Interpretable Explainability in Facial Emotion Recognition and Gamification for Data Collection

Krist Shingjergji  
*Educational Sciences*  
Open University of the Netherlands  
Heerlen, The Netherlands  
krist.shingjergji@ou.nl

Deniz Iren  
*Center for Actionable Research*  
Open University of the Netherlands  
Heerlen, The Netherlands  
deniz.iren@ou.nl

Felix Böttger  
*Center for Actionable Research*  
Open University of the Netherlands  
Heerlen, The Netherlands

Corrie Urlings  
*Educational Sciences*  
Open University of the Netherlands  
Heerlen, The Netherlands

Roland Klemke  
*Educational Sciences*  
Open University of the Netherlands  
Heerlen, The Netherlands

Abstract—Training facial emotion recognition models requires large sets of data and costly annotation processes. To alleviate this problem, we developed a gamified method of acquiring annotated facial emotion data without an explicit labeling effort by humans. The game, which we named Facegame, challenges the players to imitate a displayed image of a face that portrays a particular basic emotion. Every round played by the player creates new data that consists of a set of facial features and landmarks, already annotated with the emotion label of the target facial expression. Such an approach effectively creates a robust, sustainable, and continuous machine learning training process. We evaluated Facegame with an experiment that revealed several contributions to the field of affective computing. First, the gamified data collection approach allowed us to access a rich variation of facial expressions of each basic emotion due to the natural variations in the players’ facial expressions and their expressive abilities. We report improved accuracy when the collected data were used to enrich well-known in-the-wild facial emotion datasets and consecutively used for training facial emotion recognition models. Second, the natural language prescription method used by the Facegame constitutes a novel approach for interpretable explainability that can be applied to any facial emotion recognition model. Finally, we observed significant improvements in the facial emotion perception and expression skills of the players through repeated game play.

Index Terms—Affective computing, facial emotion recognition, gamification, explainable AI, interpretable machine learning

I. INTRODUCTION

Facial expressions are imperative to non-verbal human communication as they provide a means of conveying information regarding the emotional state [1] as well as the behavioral intentions [2] of the individual. Emotions are fundamental components of social interaction [3], and the ability to express and perceive emotions is an invaluable asset for building social connections. The holy grail of affective computing is to empower computer systems with the ability to perceive and express emotions, and be able to form social ties with human users [4]. Until very recently, this ability has been considered unique to humans. However, especially with the recent advances in Artificial Intelligence (AI), many studies have been conducted that focus on the automated recognition of emotions [5].

The common approach of training machine learning models for facial emotion recognition (FER) is supervised learning, which requires large sets of data [6]. Specifically, deep FER models are challenged by a lack of sufficient data for training [7]. Collecting and curating such large datasets is a costly and time-consuming endeavour since labeling by human annotators is necessary [8]. This poses an obstacle to achieving significant performance improvements in emotion recognition research.

Another major challenge lies in the explainability and interpretability of emotion recognition models. Studies mostly evaluate emotion recognition models using accuracy and confusion matrices; however, these metrics often fall short in reporting the utility of the models for humans. Interpretable models should provide explanations that are simple enough to be understood by their users, and are given in a language that is meaningful to them [9]. The explainability of emotion recognition models has been very rarely addressed in literature. The approaches to achieve explainability are limited to model-agnostic methods that explain the output of the model based on the inputs, and model-transparent methods (e.g., [10], [11]) that highlight the activation in different layers of artificial neural networks [12]. However, neither approach necessarily provides human-friendly explanations that are interpretable by their users.

The challenges regarding collecting and curating excessive amounts of labeled data for training FER models, and yielding interpretable explanations from such models call for heterodox methods. In this study, we propose a gamification approach towards the collection of annotated facial emotion data. The proposed game, which is named; *Facegame*, embodies a method for providing natural language prescription as feedback to the players, effectively serving as a means of achieving interpretable explainability. In summary, our contributions to the field of affective computing are as follows:

- We present a gamification approach for rapidly collecting...
annotated facial emotion data that is rich in variety of facial expressions, in a low-cost, low-effort manner.

- We propose a novel approach for interpretable explainability by translating the intermediary facial features into natural language prescriptions, and providing them as an explanation for the emotion classifications provided by any FER model.
- The presented gamification approach leads to significant improvements in the facial emotion perception and expression skills of the players.

The remaining of this paper is structured as follows. Section II provides a literature review on emotion recognition, explainable AI, and gamified data collection. Section III describes the core contributions. Section IV presents the details of our experimental study. Section V discloses the results of our experiments. Finally, Section VI provides a discussion on the theoretical and practical implications of our contributions, and concludes the paper.

II. RELATED WORK

A. Face expressions, action units, and their automated recognition

Facial expressions are a means for humans to express their emotions; thus, their automated detection is an important goal of affective computing. Facial expressions are movements and positions of the facial muscles that can be identified by Action Units (AUs); hierarchical components of movements of individual or groups of facial muscles that describe the changes in facial expressions [13]. There are studies focusing on the correlation between AUs and the basic emotions, namely, happiness, sadness, fear, disgust, anger, and surprise [14]. For example, Reisenzein et al. [15] reported coherence between amusement and smiling, and Wegrzyn et al. [16] presented a detailed mapping between the basic emotions and different parts of the face, e.g., lid raiser is essential for fear detection and lid tightener for anger. Apart from the basic emotions, there are studies aiming at detecting more complex emotional states, such as confusion, by utilizing AUs [17].

The strong relationship between emotional facial expressions and AUs has motivated researchers to develop AU-detection algorithms as well as curating AU-labeled face expression datasets such as CK+ [18] and DISFA [19]. For instance, Baltrušaitis et al. [20] presented an AU occurrence and intensity algorithm based on Histogram of Oriented Gradients (HOG); a method to describe an image by the distribution of intensity gradients or edge direction [21], and geometrical features (e.g., shape and landmarks; detection and localization of certain characteristic points on the face [22]). This work also highlights the positive impact of using various datasets to the generalizability of the model performance. Shao et al. [23] presented a framework for detecting 10 AUs using the attention mechanism, i.e., finding the region-of-interest for each AU. Jacob and Stenger [24] outperformed their previous model by employing a correlation network based on a transformer-encoder architecture, to capture the relationship between different AUs for a wide range of expressions of emotions. Other prominent examples of architectures for AU detection are the JAA-Net [25], which uses high-level features of face alignment for AU detection, and DRML [26] which uses feed-forward functions to induce regions on the face that are important.

B. Explainable and Interpretable AI

As AI finds widespread application across many domains, the need for explainable AI (XAI) is rapidly growing as well. However, most explainability approaches do not target end-users, and their outcomes are not directly interpretable by humans. One way to address this issue is to improve the transparency of AI models. Model transparency focuses on explaining “how the system made a decision” [27]. There are models that are transparent by design, e.g., decision trees, and others that are “black box” and require additional tools for explainability [28]. In recent years, explanation tools have been designed to provide users insights on the decision-making process of a system. The study of Jeyakumar [12] showed that the users prefer explanations by example in most of the cases. Rosenfeld [29] presents a set of metrics that are suitable for evaluating the effectiveness of explainable AI, namely, i) the difference between the explanations’ logic and the agent’s actual performance, ii) the number of rules in the agent’s explanation, iii) the number of features used to construct the explanation, and iv) the stability of the agent’s explanation.

C. Crowdsourcing and Gamification for Data Collection

Crowdsourcing is defined as the act of outsourcing a task that is commonly performed by designated agents to a large number of individuals [30]. Crowdsourcing has been used both by industry and scientific communities for a variety of purposes, one of which is labeled data collection. In most cases, crowd workers complete a task with the motivation of monetary gain. Even though this approach has been proven cost-effective, it has also been criticized because it potentially leads to questionable data quality unless the necessary quality assurance mechanisms are put in place [31]. A subtype of crowdsourcing, games-with-a-purpose [32], provides a different kind of incentive [33] for the workers to complete the tasks to the best of their abilities, and it generally incurs no additional costs to the employer. The design of crowdsourcing tasks in the form of a game is considered a part of a much larger concept: gamification. Gamification can be defined as a technique of using game elements in non-game systems to improve user experience and engagement [34], increasing the motivation of the respondents by satisfying psychological and social needs [35]. Gamification of data collection finds application in different domains [36], such as education [37] [38] and health [39].

III. GAMIFIED DATA COLLECTION AND INTERPRETABLE FER

In this section, we present our solution that addresses the challenges of labeled data collection for FER model
training and devising human-friendly explanations for emotion recognition systems. Specifically, we elaborate on the gamified data collection approach of Facegame and the underlying interpretable FER method.

A. Facegame

Emotional facial expressions emerge rapidly and mostly involuntarily on human faces. Nevertheless, humans are exceptionally good at recognizing even the subtlest cues that appear on the faces of others. Even though humans inherently possess these abilities, it is surprisingly challenging to exercise them deliberately. The motivation of Facegame is to provide the players with a challenging way to exercise the skills of facial expression perception and mimicking.

In Facegame, the goal of the players is to mimic the facial expression shown on a target image. The interface of Facegame (Fig. 1) displays two images together; (a) a target image from the database of the game, which contains a face that exhibits a certain basic emotion and (b) the player’s camera feed. Thus, the interface allows the player to compare both faces and try to imitate the target face.

All target images in the database are labeled based on six basic emotions [13] by human experts, as well as the 20 AUs (Table I) automatically using a pre-trained classifier provided by Py-Feat [40]. The classifier’s inputs are the following two vectors: the facial landmarks, a \((68 \times 2)\) vector of the landmark locations that is computed with the dlib package [41], and the HOGs; a vector of \((5408 \times 1)\) features that describe an image as a distribution of orientations [21]. The HOGs are calculated for the faces that are aligned using the position of the two eyes, and masked using the positions of the landmarks. The model’s output is a list of AUs detected on the face image. The pipeline of the AU detection is shown in Fig. 2.

The player is given five seconds to mimic the target expression. Afterward, the player’s image is processed and automatically labeled with AUs. The set of AUs on the target image are already known beforehand. Consecutively, the Jaccard Index of the two AU sets yields the score as seen in the Equation 1; \(P\) and \(T\) depicting the AU sets of the player and target respectively. Players can retry imitating the same facial image to increase their scores or move on to another image.

\[
\text{Score}(P, T) = \frac{|P \cap T|}{|P \cup T|}
\]

B. Quality of data collected from the game

Every time a player plays the game, new data are generated. The score as a game element provides feedback to the player, and the players are motivated to do better and improve their scores, thus, generating better representations of the target facial emotion. This acts as an inherent quality assurance mechanism. Nevertheless, some players may show poor performance due to various reasons. For instance, they might be just exploring and testing the game, or their lighting conditions might be sub-optimal, or they simply do not feel motivated to do good in the game. In any case, such turns would yield low scores, and the data originating from players that consistently score low can easily be eliminated, thus, diminishing noise in data.

The turns that yield a high score are considered good representations of the emotional face expression on the target image, which is already labeled with one of six basic emotions. Thus, the players’ image can automatically be annotated with the same label as the target image. The minor differences between the player and target AU sets provide a desirable variance in the distribution of AUs corresponding to a certain emotional face expression. This way, the variance in the AU distribution is created naturally by human players instead of automatically generated by means of simulation, potentially improving the in-the-wild performance of FER when used in training.

C. Interpretable FER Explanation

Even though there is no precise formula for how combinations of AUs translate into emotional expressions, some strong correlations exist. For instance, a happy face generally exhibits a smile, which is characterized by the existence of the “lip corner puller.” We propose using AUs as a means for explaining the the outcome of FER models. Specifically, we utilize AU detection parallel to FER, and translate the identified AUs into natural language descriptions, which constitute human-friendly, interpretable explanations of FER (Figure 3). The natural language descriptions are generated by a rule-based dictionary approach.

We created the dictionary of AU descriptions based on the definitions on Facial Action Coding System [13]. The
TABLE I
THE 20 AUs PREDICTED BY THE AU DETECTION MODEL.

| AU   | Description            |
|------|------------------------|
| AU1  | Inner Brow Raiser      |
| AU2  | Outer Brow Raiser      |
| AU3  | Brow Lowerer           |
| AU4  | Upper Lid Raiser       |
| AU5  | Lip Suck               |
| AU6  | Cheek Raiser           |
| AU7  | Lid Tightener          |
| AU8  | Nose Wrinkler          |
| AU9  | Upper Lip Raiser       |
| AU10 | Deepner                |
| AU11 | Chin Raiser            |
| AU12 | Lip Corner Puller      |
| AU13 | Dimpler                |
| AU14 | Lip Corner Depressor   |
| AU15 | Jaw Drop               |
| AU16 | Lip Corner Raiser      |
| AU17 | Lip Stretcher          |
| AU18 | Lip Tightener          |
| AU19 | Lip Pressor            |
| AU20 | Lips Part              |
| AU21 | Jaw Raiser             |
| AU22 | Lip Corner Depressor   |
| AU23 | Lip Corner Raiser      |
| AU24 | Jaw Drop               |
| AU25 | Lips Part              |
| AU26 | Jaw Drop               |
| AU27 | Lip Suck               |
| AU28 | Eyes Closed            |

The dictionary contains AU combinations categorized based on the facial muscle types and areas of appearance: cheeks, eyebrows, eyelids, lips, chin and nose, mouth, horizontal, oblique, and orbital. Every combination of the AUs that fall into these categories are represented in the dictionary. One example entry from the dictionary is as follows:

Eyebrows, AU4: “brow lowerer”
- description: “eyebrows are lowered.”
- prescription (+): “lower your eyebrows,”
- prescription (-): “do not lower your eyebrows.”

The prescriptions of what players must do in order to improve their scores are given based on the outcome of a comparison between the target AU set and the player AU set. The intersection of both sets (P for the player AUs and T; target AUs) are the correctly mimicked AUs, while the difference between them show two kinds of mistakes; The set P-T includes the AUs that should be removed from the player’s expression, and the set T-P covers the AUs that are missing on the player’s expression to mimic the target successfully. The AUs in both sets of mistakes are expressed as natural language prescriptions in different polarities; for example; “raise your eyebrows” and “do not raise your eyebrows”.

IV. METHODOLOGY

In this section, we present the overall methodology of this study, and describe the experimental setup and procedure. In order to evaluate the efficacy and utility of the proposed gamified data collection method, i.e., Facegame, we pose the first research question as follows.

**RQ1: Do the data generated by the players represent emotional facial expressions accurately?**

We also hypothesize that through playing the game repeatedly, players exercise deliberate practice of their face expression and perception skills, and thus, be able to improve them. Hence, our second research question is as follows.

**RQ2: Do the players improve their facial expression and perception skills through repeated play?**

Finally, to evaluate how the proposed explanation method is perceived, and whether this yield an improved level of understanding regarding the outcome of the FER model. Hence, we formulate the third research question as follows.

**RQ3: Do the natural language explanations help players understand the outcome of the FER model?**

A. Experiment setup and procedure

In the experiment design, we adjusted the Facegame slightly. Participants were asked to play six rounds of the game, each round corresponding to one of the six target images, and each image representing one basic emotion. Participants were shown each target image five times in a row to be able observe the score change in each try. They were given five seconds, indicated by a countdown on the screen, to mimic the face. Following, the score of their attempt was displayed (Eq. 1). To examine the potential effect of the natural language prescriptions, we divided the participants into two groups. One group received the natural language prescriptions as well as the score (treatment group), and the other received the natural language prescriptions as well as the score (control group).

B. Analysis of the collected emotional facial image data

To examine the first research question, we analyzed the collected emotional facial image data, and evaluate to what extent the collected data can be used to model facial expressions of emotions. To this end, we used two commonly used in-the-wild data sets; Facial Expression Recognition 2013 (FER2013) [42], and The Real-world Affective Face Database (RAF-DB) [43] [44], and we trained a simple neural network classifier.

FER2013 is a large-scale dataset that includes images collected automatically by the Google image search API labelled with the six basic emotions and neutral. RAF-DB is a real-world dataset that includes images of faces collected from the Internet and manually labelled with crowdsourcing. The neural network architecture we trained is similar to the baseline model used in the study of Bishay et al. [45], a shallow convolutional neural network (CNN) including four convolutional layers with 32, 32, 64, and 64 filters respectively and the ReLU activation function. The first three layers were followed by a max-pooling layer with a $2 \times 2$ filter, and the last
one by a flatten layer. The final two layers are fully-connected. The first fully-connected layer has 96 neurons, while the second has six sigmoid units representing the predictions of the six emotions.

Considering that FER2013 and RAF-DB are large-scale, unbalanced datasets, samples were taken in order to respect the size and class distribution of the Facegame data. The samples were created by random sampling without replacement 200 and 50 instances from each of the six emotion from the training and testing set, respectively. With this technique we obtained a balanced sample of each set containing 1,200 instances for training and 300 instances for testing. To achieve generalizable observations, we formed five different samples from each of both FER2013 and RAF-DB datasets. Finally, we compared the performance of the models trained on each sample set twice; first, without the inclusion of the Facegame Data, second, with the inclusion of Facegame Data.

C. Participant Survey

To receive quantitative and qualitative feedback regarding the design of Facegame, and to partially answer the third research question, we prepared a questionnaire. The participants were asked to fill in the questionnaire after completing the game experiment. The questions covered demographics questions, i.e., age and gender, technical information, i.e., type of device and browser used for the game, and their quantitative and qualitative feedback on the game. The quantitative feedback was given with a 5-Likert scale score (Very Satisfied, Somewhat Satisfied, Neutral, Somewhat Unsatisfied, Very Unsatisfied) on different aspects of the game; the ease of use, time to load and browser compatibility, and design. The survey for the participants in the treatment group included additional information regarding the natural language prescriptions. Specifically, they were asked to give a score on the usefulness and the design of the prescriptions. The qualitative feedback was sought with two open-ended questions about comments and suggestions regarding the functionality and the design of the game. Similar to the quantitative feedback, the participants of the treatment group were asked two additional open-ended questions regarding the clarity and other comments on the prescriptions that were displayed. The analysis of the survey data consisted of descriptive statistics on the age of the participants, quantitative analysis of the satisfaction scores regarding the design aspects of the game, and qualitative analysis on the open-ended questions.

D. Analysis of Skill Improvement

To answer the second and the third research questions, we analyzed the consecutive scores of the players in each round. Specifically, we compared the first score of each round against the mean of the remaining four scores of the same round. We tested the significance of the difference between the mean of the scores for both groups, and also between the groups using a t-test.

V. Results

A. Facegame Scores and Skill Improvement

A total number of 36 individuals participated in the experiment of which 18 received the natural language prescriptions while the remaining 18 received only a score. In total, 216 games were played, each game yielding five consecutive scores: \( S_1, S_2, S_3, S_4, S_5 \). We examined the score change by comparing the distributions of \( S_1 \) and the mean of the rest; \( M_{\text{rest}} = (S_2, S_3, S_4, S_5) \).

Our results show that, for all games \((N = 216)\) the score increased significantly between \( S_1 \) and \( M_{\text{rest}} \) with \( t(215) = 2.61, p < 0.01 \) which is a clear indication of the learning effect of Facegame.

For the control group \((N = 108)\), the same comparison also yielded a significant increase in the distributions \( S_1(M = 0.41, SD = 0.23) \) and \( M_{\text{rest}}(M = 0.46, SD = 0.21), t(107) = 2.26, p = 0.02 \). For the treatment group \((N = 108)\) that received natural language prescriptions, the comparison showed an insignificant increase in the score distributions \( S_1(M = 0.40, SD = 0.24) \) and \( M_{\text{rest}}(M = 0.44, SD = 0.22), t(107) = 1.44, p = 0.15 \).

Additionally, we investigated the number of times a game ends with an increased score for both groups. Our observations showed that 62.9\% of the games resulted with an increased score when the participants received natural language descriptions, while the score increased 57.4\% when the participants received only a score.

B. Survey Results and Player Feedback

In the online survey, 36 participants (22 male; 14 female) provided feedback. The age of the participants ranged between 25 and 55 years \((M = 33.77, SD = 7.66)\). Fig. 4 presents the satisfaction scores provided by all participants regarding various aspects of the design of the game. Fig. 5 shows the feedback of the participants who received natural language prescriptions \((N = 18)\) in the experiment regarding their content, design, and usefulness. The results revealed that most of the participants were satisfied or neutral to the content and the design of the prescriptions while attention should be given to the usefulness prescriptions.

We manually semantically grouped the answers from the two open-ended questions. The open-ended question that inquires participants’ suggestions on the prescriptions shows that participants find the use of visualization, prioritization, and personalization useful. Specifically, four participants mentioned that on-screen visualization of the part of the face that they needed to change would help them follow the natural language prescriptions better. Three of the participants found the text too long to read in a short time span and suggested the display of a few of the most important instead. Two participants indicated that more personalized prescriptions would be helpful. Lastly, the 12 participants that commented on the natural language of the prescriptions stated that they found them understandable.
C. Facial Emotion Recognition with Facegame Data

The Facegame Data that were collected in the experiment consist of 636 images depicting the six basic emotions (86 angry, 89 disgusted, 127 fearful, 132 happy, 99 sad, and 103 surprised). The accuracy for each of the five samples for FER2013 and RAF-DB is shown in Fig. 6 and Fig. 7, respectively. The average accuracy of the model trained on instances from FER2013 is 32.80%, while for the model trained on the instances from the combination of FER2013 and the Facegame Data is 33.20%. Similarly, the average accuracy of the model trained on instances from RAF-DB is 51.27%, while for the model trained on the instances from the combination of FER2013 and the Facegame Data is 51.87%. The observed low accuracies of the models were expected considering the nature of the sets and the simplicity of our neural network architecture. The images in FER2013 and RAF-DB represent in-the-wild conditions with subtle facial expressions. Therefore, the emotion recognition task is more challenging, and it requires complex classifiers to achieve better accuracy. Moreover, our goal was to examine the quality of the data collected via Facegame in comparison with the existing in-the-wild sets, and for this purpose, a simple neural network sufficed. The results show that we were able to collect labelled data that can potentially improve FER in-the-wild.

D. The Mapping of AUs to Emotion Classes and the Variability

The results of the correlation between the six basic emotions and the AUs are shown in Fig. 8. For this analysis, the data from the trials that scored below $\frac{1}{2}$ were excluded. The results suggest that we were able to capture some strong correlations between certain emotional facial expressions and their signature AUs. For instance, lips part (AU25) and jaw drop (AU26) highly correlate with both surprise and fear, while lip corner puller (AU12) highly correlates with happiness. The high occurrence number of nasolabial deepener (AU11) is an indication of the underlying Py-Feat model creating a high false-positive rate for AU11, which needs further examination. The results show that we were able to define the emotion classes as a distribution of multiple AUs. Such naturally occurring variety in facial expression data can potentially be used to improve FER in-the-wild.

VI. DISCUSSION AND CONCLUSION

In this paper, we introduced a novel approach for explainable FER that promises three related contributions. First, by means of gamification, we have developed a method for collecting annotated face expression data continuously, which allows us to describe the facial expressions of six basic emotions as a distribution of AUs. Secondly, we proposed and evaluated an interpretable FER explainability method that uses AUs as features to describe the outcomes of FER models, i.e., facial emotion classes. The experimental observations indicate that the natural language explanations of face expressions are interpretable by humans. Our quantitative and qualitative results highlight improvement opportunities regarding the design of Facegame and how we communicate the face expression explanations. Finally, we observed that the players were able to...
to collect more data by crowdsourcing a large number of players who will result in a more in-depth analysis of

the collected data set. In this setting, the player was asked to mimic the facial expressions of a face displaying one emotion and not triggered to experience that emotion. This approach has limitations considering that there are differences between the posed and spontaneous facial expressions in morphological and dynamic aspects of certain emotions [48] as well as demographical mismatches between the target image and the player. In future studies, we aim to increase the variance of the dataset by including in the set of the target images faces with spontaneous facial expressions as well as increasing the demographic diversity of the target images.

**ETHICAL IMPACT STATEMENT**

In this study, the personal data collection was limited to facial images of the players. The collected data were securely stored in the server of the research institute which can only be accessed by the researchers of this study. The participants were informed regarding the experiment and data collection, and they provided consent prior to taking part in the experiment.

**REFERENCES**

[1] K. R. Scherer, and P. H. Tannenbaum, “Emotional experiences in everyday life: A survey approach,” Motivation and Emotion, vol. 10, pp. 295–314, 1986.

[2] A.J. Fridlund, “Human facial expression: An evolutionary view,” Academic press, March 2014.

[3] D. Matsumoto, K. Dacher, N. M. Shiota, M. O’Sullivan, and M. Frank, “Facial expressions of emotion,” The Guilford Press, 2008.

[4] R. W. Picard, “Affective computing,” MIT press, July 2000.

[5] M. Maihri, U. Raghavendra, A. Gudivar, J. Samanth, D. B. Prabal, M. Murugappan, et al. “Automated Emotion Recognition: Current Trends and Future Perspectives,” Computer Methods and Programs in Biomedicine, pp. 106646, 2022.

[6] B. C. Ko, “A brief review of facial emotion recognition based on visual information,” Sensors, vol. 18, pp.401, February 2018.

[7] S. Li, and D. Wei Hong, “Deep facial expression recognition: A survey,” IEEE transactions on affective computing, 2020.

[8] S. Li, and D. Wei Hong, “Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition,” IEEE Transactions on Image Processing, vol. 28, pp.356–370, 2018.

[9] I.H. Gipin, D. Bau, B.Z. Yuan BZ, A. Bajwa, M. Specter, and L. Kagal. “Explaining explanations: An overview of interpretability of machine learning,” 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), pp. 80-89, October 2018.

[10] P. Kumar P. Kauhlik V, and B. Raman, “Towards the Explainability of Multimodal Speech Emotion Recognition,” Interspeech, pp 1748–1752, 2021

[11] E. Ghaleb, A. Mertens, S. Asteriadis, and G. Weiss, “Skeletion-Based Explainable Bodily Expressed Emotion Recognition Through Graph Convolutional Networks,” 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021), pp. 1–8, 2021.

[12] I.V. Jeyakumar, J. Noor, Y.H Cheng, L. Garcia, and M. Srivastava, “How can i explain this to you? an empirical study of deep neural network explanation methods,” Advances in Neural Information Processing Systems, vol. 33, pp. 4211–4222, 2020.

[13] P. Ekman, and W. Friesen, “Facial action coding system,” Environmental Psychology & Nonverbal Behavior, 1978.

[14] P. Ekman, “Facial expression and emotion,” American psychologist vol. 48, pp. 384, April 1993.

[15] R. Reisenzein, M. Studtmann, and G. Horstmann, “Coherence between emotion and facial expression: Evidence from laboratory experiments,” Emotion Review, vol. 5, pp. 16–23, January 2013.

[16] M. Wegryn, M. VOGT, B. KireclioGLU, J. Schneider, and J. KISSLER, “Mapping the emotional face. How individual face parts contribute to successful emotion recognition,” PloS one, vol. 12, pp. 11–12, May 2017.
[17] N. Borges, L. Lindblom, B. Clarke, A. Gander, and R. Lowe, “Classifying confusion: autodetection of communicative misunderstandings using facial action units,” 2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (AChIWi), pp. 401–406, September 2019.

[18] 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,” 2010 IEEE computer society conference on computer vision and pattern recognition-workshops, pp. 94–101, June 2010.

[19] S.M. Mavadati, M.H Mahoor, K. Bartlett, P. Trinh, and J.F. Cohn, “Disfa: A spontaneous facial action intensity database,” IEEE Transactions on Affective Computing, pp. 151–160 2013 Mar 7;4(2):151–60.

[20] T. Baltrušaitis, M. Mahmoud, and P. Robinson, “Cross-dataset learning and person-specific normalisation for automatic action unit detection,” 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), vol. 6, pp. 1–6, May 2015.

[21] N. Dalal, and B. Triggs, “Histograms of oriented gradients for human detection.” In 2005 IEEE computer society conference on computer vision and pattern recognition (CVP’05), vol. 1, pp. 886–893, June 2005.

[22] O. Çelikutan, S. Ulukaya, and B. Sankur, “A comparative study of face landmarking techniques,” EURASIP Journal on Image and Video Processing, pp. 1–27, December 2013.

[23] Z. Shao, Z. Liu, J. Cai, Y. Wu, and L. Ma, “Facial action unit detection using attention and relation learning,” IEEE transactions on affective computing, October 2019.

[24] G.M. Jacob, and B. Stenger, “Facial action unit detection with transformers,” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7680–7689, 2021.

[25] Z. Shao, Z. Liu, J. Cai, Y. Wu, and L. Ma, “JAA-Net: joint facial action unit detection and face alignment via adaptive attention,” International Journal of Computer Vision, vol. 129, pp. 321–340, February 2021.

[26] K. Zhao, W.S Chu, and Zhang, “Deep region and multi-label learning for facial action unit detection,” Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3391–3399, 2016.

[27] A. Rosenfeld, and A. Richardson, “Explainability in human–agent systems,” Autonomous Agents and Multi-Agent Systems, vol. 33, pp. 673–705 November 2019.

[28] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, “A survey of methods for explaining black box models,” ACM computing surveys (CSUR). 2018 Aug 22;51(5):1-42.

[29] A. Rosenfeld. “Better metrics for evaluating explainable artificial intelligence,” Proceedings of the 20th international conference on autonomous agents and multiagent systems, pp. 45–55, May 2021.

[30] J. Howe J, “The rise of crowdsourcing,” Wired magazine, vol. 6, pp. 1–4, June 2006.

[31] D. Iren, S. Bilgen, “Cost of quality in crowdsourcing,” Human Computation, vol. 1. December 2014.

[32] A.J Quinn, and B.B Bederson, “Human computation: a survey and taxonomy of a growing field,” Proceedings of the SIGCHI conference on human factors in computing systems, pp. 1403–1412, May 2011.

[33] M. Hosseini, A. Shahri, K. Phalp, J. Taylor, and R. Ali, “Crowdsourcing: A taxonomy and systematic mapping study,” Computer Science Review, vol. 17, pp. 43–69, August 2015.

[34] S. Deterding, D. Dixon, R. Khaled, and L. Nacke, “From game design elements to gamefulness: defining “gamification,” Proceedings of the 15th international academic MindTrek conference: Envisioning future media environments, pp. 9–15, September 2011

[35] A.F. Aparicio, F.L. Vela, J.L. Sánchez, and J.L. Montes, “Analysis and application of gamification,” Proceedings of the 13th International Conference on Interacción Persona-Ordenador, pp. 1–2, October 2012.

[36] V. Gurav, M. Parkar, and P. Kharwar, “Accessible and Ethical Data Annotation with the Application of Gamification,” International Conference on Recent Developments in Science, Engineering and Technology, pp. 68–78, November 2019.

[37] J.J López-Jiménez, J.L. Fernández-Alemán, L.L. González, O.G Sequeros, B.M Valle, J.A García-Berná, et al. “Taking the pulse of a classroom with a gamified audience response system,” Computer Methods and Programs in Biomedicine, vol. 213, pp. 106459, January 2022.

[38] A.I. Wang, and R. Tahir, “The effect of using Kahoot! for learning–A literature review.” Computers & Education, vol. 149, pp. 103818, May 2020.