Interpretable Explainability for Face Expression Recognition

Krist SHINGJERGJI[a,1], Deniz IREN[a,b], Felix BÖTTGER[a, b], Corrie URLINGS[b], and Roland KLEMKE[a,c]
[a] *Open Universiteit, Heerlen, Netherlands*
[b] *Center for Actionable Research of the Open Universiteit, Heerlen, Netherlands*
[c] *Cologne Game Lab, Koln, Germany*

Abstract. Training facial emotion recognition models requires large sets of data and costly annotation processes. Additionally, it is challenging to explain the operation principles and the outcomes of such models in a way that is interpretable and understandable by humans. In this paper, we introduce a gamified method of acquiring annotated facial emotion data without an explicit labeling effort by humans. Such an approach effectively creates a robust, sustainable, and continuous machine learning training process. Moreover, we present a novel way of providing interpretable explanations for facial emotion recognition using action units as intermediary features and translating them into natural language descriptions of facial expressions of emotions.

Keywords. Face expression recognition, explainable AI, gamification, human-in-the-loop, crowdsourcing, interpretable machine learning

1. Introduction

The ability to perceive emotions has long been attributed to humans. However, especially with the recent advances in AI, many studies have been conducted that focus on the automated recognition of human emotions. Prominent types of emotion recognition include Speech Emotion Recognition (SER) [1] Facial Emotion Recognition (FER) [2] and Multimodal Emotion Recognition (MER) [3]. The common approach of training machine learning models for emotion recognition is supervised learning, which entails collecting and curating a large number of emotion labels from human annotators. Such an approach requires costly annotation campaigns, which poses an obstacle in front of achieving human-level performance in emotion recognition.

Another major challenge lies in the explainability and interpretability of emotion recognition models [4]. 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. The explainability of emotion recognition models have

[1] Corresponding Author, Open University of The Netherlands; E-mail: krist.shingjergji@ou.nl

been very rarely addressed in the literature. More importantly, the approaches to achieve explainability are limited to model-agnostic methods that explain the output of the model based on the inputs, model-transparent methods (e.g.,[5],[6]) that highlight the activation in different layers of artificial neural networks [7]. However, neither approach necessarily provides human-friendly explanations, i.e., explanations that are understandable and interpretable by humans.

The challenges regarding collecting and curating excessive amounts of labeled data for machine learning training, and yielding interpretable explanations from the machine learning models call for heterodox methods that promise transparency and resource-efficiency. In this study, we propose a FER approach using machine learning in the form of a game that makes two contributions to the fields of explainable artificial intelligence and emotion recognition. First, it discloses a method that inherently utilizes the data generated by user interaction (i.e., game play) in the training loop, effectively creating a robust, sustainable, and continuous self-training process. Secondly, by translating the intermediary facial features into natural language descriptions and instructions, it provides a means of creating interpretable explanations that can be applied to any FER system.

This paper is structured as follows. Section 2 provides a literature review on emotion recognition, explainable AI and human-in-the-loop, and gamified data collection and shares several related works. Section 3 describes the core contributions, i.e., (a) gamified data collection for continuous machine learning model training and (b) interpretable explanations for FER systems. Section 4 presents the details of our evaluation method that consists of a between-groups experimental study. Section 5 discloses the results of our experiments. Finally, Section 6 provides a discussion on the theoretical and practical implications of our contributions and concludes the paper.

2. Related Work

2.1. 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 the field of emotion recognition. Facial expressions are movements and positions of the facial muscles that can be expressed by Action Units (AUs); hierarchical components of movements of individual or group of facial muscles that describe the changes in facial expressions [8]. There are a plethora of studies focusing on the correlation between action units and the basic emotions, namely, happiness, sadness, fear, disgust, anger, and surprise [9]. For example, in [10], the authors reported coherence between amusement and smiling and Wegrzyn et al. in[11] 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 have been studies focusing on detecting more complex emotional states, such as confusion, by utilizing AUs [12]. Nevertheless, there is research suggesting that there is a variety of how people express their emotions in different contexts and social circumstances [13].

The strong relationship between emotional facial expressions and AUs has motivated researchers to focus their attention on developing algorithms for detecting AUs as well as curating AU-labeled face expression datasets such as CK+ [14] and DISFA [15]. For instance, Baltrušaitis et al. [16], presented an AU occurrence and intensity algorithm based on appearance features (i.e., histogram of oriented gradients) and geometry features (e.g., shape and landmarks), highlighting the generalizability benefits of using data from different datasets. Shao et al. [17] presented a framework for detecting 10 AUs using the attention mechanism, i.e., finding the region of interest for each AU. Jacob and Stenger [18] outperformed their current state-of-the-art model by employing a correlation network, based on a transformer encoder architecture, to capture the relationships between different AUs for a wide range of expressions. Other prominent examples of architectures for AUs detection are the JAÂ-Net [19] which uses high-level features of face alignment for AU detection and DRML [20] that uses feed-forward functions to induce regions in the face that are important. Our proposed method can be used for collection of emotion labeled and AU labeled data for training FER models.

2.2. Explainable and Interpretable AI

As artificial intelligence finds application in an increasing number of domains, the need for explainable AI (XAI) is rapidly growing as well. However, most explainability approaches do not target end users, and are not directly interpretable by humans. One way to address this issue is to focus on the transparency of AI models. Model transparency focuses on explaining “how the system made a decision” [21]. There are models that are transparent by design, e.g., decision trees, and others that are “black box” and require additional tools for explainability [22]. In recent years, explanation tools have been designed to provide users insights on the decision-making process of a system. The study of Jeyakumar [7], showed that the users prefer the explanation-by-example in the image domain. Rosenfeld in [23] presents a set of metrics that are suitable for evaluating the effectiveness of explainable AI, namely, i) the performance difference between the agent’s model and the performance of the logic presented as an explanation, ii) the number of rules in the agent’s explanations, iii) the number of features used to construct the explanation, and the stability of the agent’s explanation. In this paper we present a novel way of providing interpretable explanations for facial emotion recognition and we evaluate their effectiveness.

2.3. Gamified Data Collection

Crowdsourcing is a technique in which crowds are incorporated into the labeling procedure. It was firstly defined by Howe in (Howe, 2006). It is a powerful tool used by industry and scientific research for a variety of purposes, including 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 [24] unless necessary quality assurance mechanisms are put in place. A subcategory of crowdsourcing; games-with-a-purpose [25]provides a different kind of incentive [26] for the workers to complete the tasks to the best of their abilities, and it generally incurs no cost. The designing of crowdsourcing tasks in the form of a game is considered a part of a much larger concept, that is gamification. Gamification can be defined as a technique of using game elements in non-game systems to improve user experience and engagement [27] increasing the motivation of the respondents by satisfying psychological and social needs [28].

Gamification of data collection finds application in different domains [29] such as education [30] [31] and health [32]. In this study we suggest a gamified method of collecting labelled facial emotion data.

3. Gamified Data Collection and Interpretable FER

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

3.1. Gamified Data Collection for FER: Face Game

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 our game; i.e., Face Game, is to provide the players with a challenging way to exercise the skills of facial expression perception and mimicking.

The goal of Face Game is to mimic the facial expression that is shown on a target image. The interface of Face Game displays two images together; (a) a target image from the database of the game, which contains a face that exhibits a certain emotion, (b) 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 [8] by human experts, as well as 20 AUs automatically using Py-Feat [33]. The player is given five seconds to mimic the target expression. Afterward, the player image is processed and automatically labeled with AUs. The Jaccard Index of the two AU sets, P for the player AUs and T for the target AUs, yields the score. Players can retry imitating the same facial image to increase their scores, or move on to another image.

Every time a player plays the game, a new data point is generated. The turns that yield a high score are considered good representations of the face expression on the target image, which is already labeled with one of six basic emotions. Thus, the player image can automatically be annotated with the same label as the target image. The small 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.

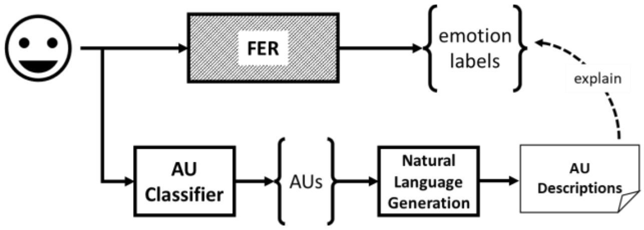

Figure 1. A graphical representation of the overall explainability model.

3.2. Interpretable FER Explanation

Emotions manifest in the form of facial expressions on human faces. Such facial expressions of emotions comprise combinations of AUs. Even though there is no clear formula of 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 operation and the outcome of FER models. Specifically, we utilize AU detection in parallel to FER, and translate the identified AUs into natural language descriptions, which constitutes human-friendly, interpretable explanations of FER (Figure 1). The natural language descriptions are generated by a rule-based dictionary approach that uses the outcome of a comparison between the target AU set (T) and the player AU set (P). The intersection of both sets 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 a prescription in different polarities; for example; "raise your eyebrows" and "try not to raise your eyebrows".

4. Methodology

In this section, we present the overall methodology of this study, reporting on the experimental setup for the data collection and the data analysis.

4.1. Experiment setup and procedure

For the experiment, we adjusted the Face Game slightly. Participants were asked to play six rounds of the game, each round corresponding to one of the six target images. Participants received each target image five times in a row. They were given three seconds, indicated by a countdown on the screen, to mimic the face. Following, the score of their attempt was displayed. To examine the potential effect of natural language instructions, we divided the participants into two groups. One group received only the score, and the other received the natural language instructions as well as the score. Our particular interest was to test if the natural language instructions lead to improved results. We considered an increase in the score a signal for learning, which implies the interpretability of the instructions.

4.2. The survey

After the completion of the game, the participants were asked to fill in a questionnaire. The questions included demographics, 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 as well as how likely it is for them to play the game again. The survey for the participants in the treatment group included additional information regarding the instructions. Specifically, they were asked to give a score on the usefulness and the design of the instructions. The qualitative feedback was asked 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 instructions that were displayed.

5. Results

5.1. Survey results

In the online survey, N = 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) . Figure 2 presents the satisfaction scores provided by all participants regarding various aspects of the design of the game. Figure 2 shows the feedback of the participants who received natural language instructions (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 instructions while attention should be given to the usefulness instructions.

We manually coded the answers of the two open-ended questions. The open-ended question that inquires participants' suggestions on the instructions show 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 instructions 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 instructions instead. Two participants indicated that more personalized instructions would be helpful. Lastly, all the participants that commented on the natural language of the instructions stated that they found them understandable.

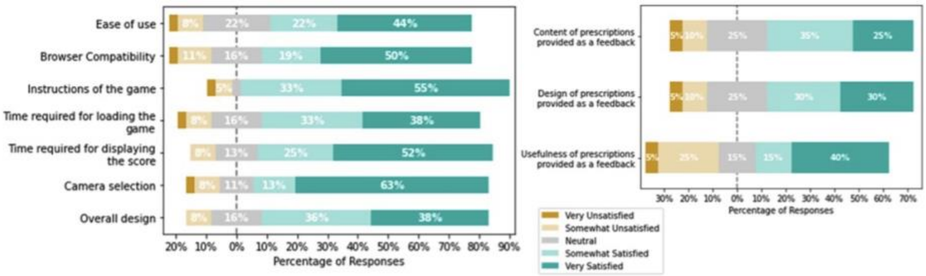

Figure 2. Feedback on different aspects of the game from players from both groups (left). Feedback on the instructions from the treatment group (right).

5.2. Face Game scores and the learning effect

A total number of 36 individuals participated in the experiment of which 18 received the natural language instructions while the remaining 18 received only a score. In total, 216 games were played, each game yielding five consecutive scores: S1, S2, S3, S4, and S5. We examined the score change by comparing the distributions of S1 and the mean of the rest; M_{rest}=M(S2,S3,S4,S5).

Our results show that, for all games (N=216) the score increased significantly between S1 (M=0.4, SD=0.23) and Mrest (M=0.45, SD=0.21) with t(215)=2.61, p<0.01 which is a clear indication of the learning effect of Face Game. However, the same comparison for the group that received the natural language instructions (N=108) yielded a difference in the distributions S1 (M=0.4, SD=0.24) and Mrest (M=0.44, SD=0.22) p(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.

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

The results of the correlation between the six emotions and the action units are shown in Figure 3. For this analysis, the data from the trials that scored below 0.33 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. Additionally, 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.

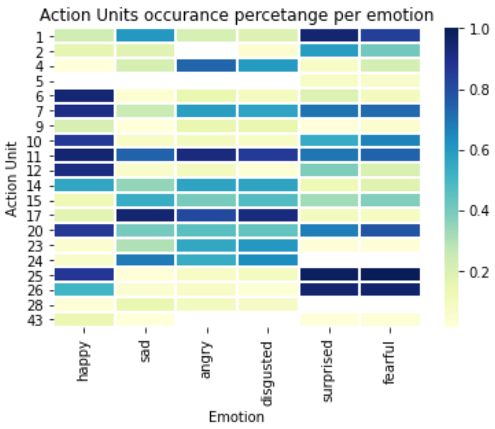

Figure 3. Heatmap of the occurrences of AUs detected on the participants and the emotions of the targeted image (threshold = 0.33).

6. Discussion and Conclusion

In this paper, we introduced a novel approach for explainable FER that promises two 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 Face Game and how we communicate the face expression explanations.

Our results have potential theoretical and practical implications. Our method of acquiring nuanced face expressions (i.e., distributions of AUs) that correlate with facial emotion classes provides a means to improve the performance of in-the-wild FER models. Moreover, the gamification approach offers a sustainable, continuous self-training process of FER models. Finally, our explainability method that uses AUs as intermediary features to describe facial emotions provides a novel approach towards achieving interpretable, human-friendly explanations of FER models. In the future, we will continue our studies and develop improved ways of delivery for the explanations by combining the natural language explanations with graphical methods.

References

[1] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, “Speech Emotion Recognition Using Deep Learning Techniques: A Review,” *IEEE Access*, vol. 7, pp. 117327–117345, 2019.

[2] B. Ko, “A Brief Review of Facial Emotion Recognition Based on Visual Information,” *Sensors (Basel)*, vol. 18, 2018.

[3] T. Bänziger, D. Grandjean, and K. R. Scherer, “Emotion recognition from expressions in face, voice, and body: the Multimodal Emotion Recognition Test (MERT).,” *Emotion*, vol. 9 5, pp. 691–704, 2009.

[4] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. A. 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, 2018.

[5] P. Kumar, V. Kaushik, and B. Raman, “Towards the Explainability of Multimodal Speech Emotion Recognition,” 2021.

[6] E. Ghaleb, A. Mertens, S. Asteriadis, and G. Weiss, “Skeleton-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.

[7] J. Jeyakumar, J. Noor, Y.-H. Cheng, L. Garcia, and M. B. Srivastava, “How Can I Explain This to You? An Empirical Study of Deep Neural Network Explanation Methods,” 2020.

[8] P. Ekman and W. v Friesen, “Facial action coding system: a technique for the measurement of facial movement,” 1978.

[9] P. Ekman, “Facial expression and emotion.,” *Am Psychol,* vol. 48 4, pp. 384–92, 1993.

[10] R. Reisenzein, M. Studtmann, and G. Horstmann, “Coherence between Emotion and Facial Expression: Evidence from Laboratory Experiments,” *Emotion Review,* vol. 5, pp. 16–23, 2013.

[11] M. Wegrzyn, 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, 2017.

[12] N. Borges, L. Lindblom, B. Clarke, A. Gander, and R. Lowe, “Classifying Confusion: Autodetection of Communicative Misunderstandings using Facial Action Units,” in *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos, ACIIW 2019*, Sep. 2019, pp. 401–406. doi: 10.1109/ACIIW.2019.8925037.

[13] L. F. Barrett, R. Adolphs, S. Marsella, A. M. Martinez, and S. D. Pollak, “Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements,” *Psychological Science in the Public Interest*, vol. 20, no. 1, pp. 1–68, Jul. 2019, doi: 10.1177/1529100619832930.

[14] P. Lucey, J. F. Cohn, T. Kanade, J. M. 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, 2010.

[15] S. M. Mavadati, P. Sanger, and M. H. Mahoor, “Extended DISFA Dataset: Investigating Posed and Spontaneous Facial Expressions,” *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW),* pp. 1452–1459, 2016.

[16] T. Baltrušaitis, M. Mahmoud, and P. Robinson, “Cross-dataset learning and person-specific normalisation for automatic Action Unit detection.” [Online]. Available: https://github.com/TadasBaltrusaitis/FERA-2015

[17] Z. Shao, Z. Liu, J. Cai, Y. Wu, and L. Ma, “Facial Action Unit Detection Using Attention and Relation Learning,” *ArXiv,* vol. abs/1808.03457, 2019.

[18] G. M. Jacob and B. Stenger, “Facial Action Unit Detection With Transformers,” *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 7676–7685, 2021.

[19] Z. Shao, Z. Liu, J. Cai, and L. Ma, “JAA-Net: Joint Facial Action Unit Detection and Face Alignment via Adaptive Attention,” *Int. J. Comput. Vis.,* vol. 129, pp. 321–340, 2021.

[20] K. Zhao, W.-S. Chu, and H. Zhang, “Deep Region and Multi-label Learning for Facial Action Unit Detection,” *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 3391–3399, 2016.

[21] A. Rosenfeld and A. Richardson, “Explainability in humanagent systems,” *Autonomous Agents and Multi-Agent Systems*, pp. 1–33, 2019.

[22] R. Guidotti, A. Monreale, F. Turini, D. Pedreschi, and F. Giannotti, “A Survey of Methods for Explaining Black Box Models,” *ACM Computing Surveys (CSUR),* vol. 51, pp. 1–42, 2019.

[23] A. Rosenfeld, “Better Metrics for Evaluating Explainable Artificial Intelligence,” 2021.

[24] D. Iren and S. Bilgen, “Cost of Quality in Crowdsourcing,” *Hum. Comput.,* vol. 1, pp. 283–314, 2014.

[25] 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*, 2011.

[26] M. Hosseini, A. Shahri, K. Phalp, J. Taylor, and R. Ali, “Crowdsourcing: A taxonomy and systematic mapping study,” *Comput. Sci. Rev.*, vol. 17, pp. 43–69, 2015.

[27] S. Deterding, D. Dixon, R. Khaled, and L. E. Nacke, “From game design elements to gamefulness: defining ‘gamification,’” 2011.

[28] A. F. Aparicio, F. L. G. Vela, J. L. G. Sánchez, and J. L. Montes, “Analysis and application of gamification,” 2012.

[29] V. Gurav, M. Parkar, and P. Kharwar, “Accessible and Ethical Data Annotation with the Application of Gamification,” 2019.

[30] J. J. López-Jiménez *et al.*, “Taking the pulse of a classroom with a gamified audience response system,” *Comput Methods Programs Biomed*, vol. 213, p. 106459, 2022.

[31] A. I. Wang and R. Tahir, “The effect of using Kahoot! for learning - A literature review,” *Comput. Educ.*, vol. 149, p. 103818, 2020.

[32] P. Tobien, L. Lischke, M. Hirsch, R. Krüger, P. Lukowicz, and A. Schmidt, “Engaging people to participate in data collection,” *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*, 2016.

[33] J. H. Cheong, T. Xie, S. Byrne, and L. J. Chang, “Py-Feat: Python Facial Expression Analysis Toolbox,” *ArXiv*, vol. abs/2104.03509, 2021.