https://doi.org/10.1145/3656650.3656711

Towards the dataset for analysis and recognition of facial expressions intensity

Marina Tiuleneva
marina.tiuleneva@studenti.unitn.it
University of Trento
Rovereto, Italy

Emanuele Castano
emanuele.castano@unitn.it
University of Trento
Rovereto, Italy
ICST, CNR
Rome, Italy

Radoslaw Niewiadomski
radoslaw.niewiadomski@unige.it
University of Genoa
Genoa, Italy

ABSTRACT

We propose a novel dataset for studying and modeling facial expression intensity. Facial expression intensity recognition is a rarely discussed challenge, likely stemming from a lack of suitable datasets. Our dataset has been created by extracting facial expressions from actors across twelve fiction films, followed by crowd-sourced online annotation of the expression intensity and variability levels. It consists of over 400 automatically extracted video segments ranging from 3 to 5 seconds, as well as annotations and facial landmarks. We also present preliminary statistics derived from this dataset.

CCS CONCEPTS

• **Computing methodologies** → *Computer vision.*

KEYWORDS

affective computing, facial expressions, video dataset, intensity

ACM Reference Format:
Marina Tiuleneva, Emanuele Castano, and Radoslaw Niewiadomski. 2024. Towards the dataset for analysis and recognition of facial expressions intensity. In *International Conference on Advanced Visual Interfaces 2024 (AVI 2024), June 03–07, 2024, Arenzano, Genoa, Italy.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3656650.3656711

1 INTRODUCTION

Numerous models have recently been proposed for emotion recognition in videos (refer to [10, 18] for recent surveys). However, the topic of facial expression intensity recognition is seldom explored, with notable exceptions [4, 8, 17, 19, 20, 23]. One of the primary reasons for this is the lack of suitable datasets for studying facial expression intensity. In related works, intensity is typically considered locally and within a specific context, such as the intensity of particular action units (e.g., local changes in the face involving one or a few muscles) or of facial expressions associated with specific emotions (see, e.g., [1, 11, 13, 14]). Humans, however, process emotion cues in a more holistic manner. Here we thus focus on overall perception of the intensity in a facial expression, i.e., not on comparing single frames of one specific expression in terms of intensity (e.g., to find their onset, offset). Datasets and/or computational models that focus on the intensity in this broader sense are scarce. Moreover, facial expression intensity models should not be restricted solely to the expression of emotions, as many other internal states (e.g., engagement [21]) and attitudes (e.g., politeness, dominance [12]) are often conveyed through facial expressions that can be evaluated in terms of intensity.

In this paper, we propose a novel dataset to study facial expression intensity. We extracted short video segments that contain just one actor's face from twelve fiction movies. Next, we asked people to rate the facial expression of the actor in terms of intensity and variability. The resulting dataset comprises of 409 video segments ranging from 3 to 5 seconds, rated on a scale of 1 to 7. Movies are easily accessible resource of a wide variety of expressions related to both positive and negative emotions, as well as other internal states, expressed by various characters. On the one hand, using movies to create affective datasets has been successful in the past, e.g., [5]. On the other hand, several theorists postulate that there may be visible differences between spontaneous expressions and fake (or "acted") ones [6]. However, for our purposes, it is not relevant whether the actors express genuine emotions or they just "act", as our focus is solely on the perceived intensity of facial changes. Another reason for using movies is that the editing techniques employed in movie production often lead to meaningful expression segmentation. Typically, actors' entire facial expressions are shown in a single shot, and, in fiction movies, there are not many unintentional or meaningless expressions. Contrary to what is typically the case for data collected in laboratory conditions, e.g. [1], that are limited to specific emotions [9], the use of movies allows us to capture a variety of facial expressions that are likely representative of real life. Because of these characteristics, this dataset can be used to 1) study human perception of facial expressions intensity, and 2) develop computational approaches to estimate facial intensity in a broader, holistic sense.

2 DATASET

Segments from twelve movies with average resolution of 736×363 and a mean duration $M = 116$ minutes were used to extract the stimuli presented to human raters. The movies were initially processed using OpenFace 2.0 [3], that detects facial landmarks [2], [22]. Segments where a human face was tracked by the software for at least 3 seconds were extracted. This constraint was imposed to prevent the occurrence of spuriously detected faces in the output set.

OpenFace 2.0 detects several faces in the same frame. If there are three faces in one frame, this frame will appear three times in the resulting OpenFace .csv file, where each line describes a

AVI 2024, June 03–07, 2024, Arenzano, Genoa, Italy

ACM ISBN 979-8-4007-1764-2/24/06
https://doi.org/10.1145/3656650.3656711

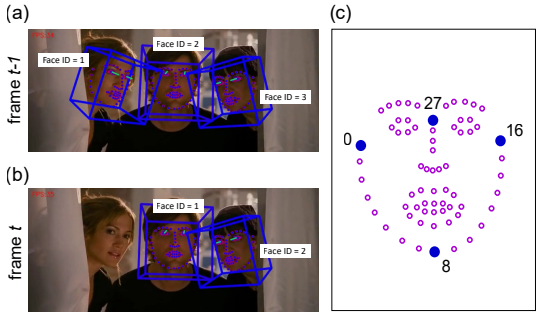

Figure 1: a) Frame $t-1$ with 3 faces recognized; b) Frame t with one face not recognized, and the IDs of the other faces changed; c) Landmarks for computing distances.

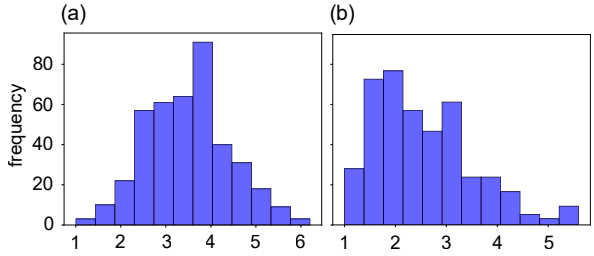

Figure 2: a) Ratings distribution for a) intensity; b) variability

separate face. We only track faces OpenFace recognized with high confidence ($success = 1$). However, the identifiers assigned to the faces in the frame t do not necessarily correspond to the same faces in the previous frame $t-1$, see Figure 1a, 1b. We measure the distance between 2D coordinates (x,y) of specific points (landmarks) in each face (the nose, ears, chin), and the corresponding coordinates in the previous frame. We match faces from different frames if $|x_i(t) - x_i(t-1)| < width$ and $|y_i(t) - y_i| < height$, where: $height = \frac{|y_{27}(t-1) - y_8(t-1)|}{2}$, $width = \frac{|x_{16}(t-1) - y_8(t-1)|}{2}$, and x_i and y_i are landmark i coordinates, $i \in \{0, 8, 16, 27\}$, as on Figure 1c. The pool of 3537 segments that was initially extracted was reduced by applying the following selection criteria: 1. **Maximum duration of 5 seconds:** Longer segments often contain more than one expression. These expressions can exhibit strong variability in terms of intensity, making the rating task challenging. 2. **Segments with only one face:** Although our procedure can track multiple faces simultaneously, to avoid ambiguity and keep the task simple, we excluded segments where more than one face appears in the frame. 3. **Minimum face dimensions:** Although our procedure can detect smaller faces, we excluded all segments where a human face covers less than 20% of the frame to ensure they are clearly visible to raters. 434 segments met these criteria. Next, we removed segments containing unrealistic deformations created with special effect techniques (e.g., incomplete faces after being shot) and segments that were erroneously extracted, by which we mean that two different faces appear in exactly the same position in two consecutive frames (as described in the tracking procedure above). The final dataset consists of 409 video segments.

3 ANNOTATION STUDY

To collect human ratings we created an online survey on Qualtrics [16]. 409 videos were split randomly into 8 groups to balance the workload of each rater. This decision was made after a pilot study in which we asked 4 people to report when they noticed a significant decline in concentration while performing the task. This time ranged from 18 to 23 minutes, during which they were able to annotate 42-55 videos. Their responses are not included in the final dataset. In the main study, each rater was randomly assigned to one of the eight video sets. The videos were presented in a random order. Raters were allowed to replay a segment multiple times before answering the questions. Once the answers are submitted, they cannot be changed. Each video starts and finishes with a white frame to make sure that the raters are not exposed to a face before and after playing the video. All videos were scaled to the same resolution 854 × 464 and were evaluated in terms of intensity and variability using two Likert scales (1-7).

47 participants took part in the study: 32 volunteers and 15 recruited via Prolific [15], and were payed £5. They were given instructions, definitions of variability and intensity, and asked to complete training to ensure they knew how to work with the platform. At the beginning of the survey, a definition of intensity and variability was provided. Intensity: *"We refer to the strength or clarity with which signals are conveyed through facial movements. "Intense" in this context refers to the strength, vividness, or prominence of the expressions displayed on individuals' faces."* Variability: *"We refer to the diversity or range of facial expressions observed in the given video. In this context "variable" implies the degree of differences or variations in facial expressions displayed by individuals in the video."* All participants gave their written consent and remained anonymous.

To exclude raters who did not carry out the task as instructed, we included four attention questions per group. In these questions, raters would see a white number against a black background and had to set both sliders to match this number. Four participants were excluded due to failing more than one of these attention questions.

The data collection resulted in 5 ratings per segment. Figure 2 shows the distribution of the ratings for intensity and variability. Regarding intensity, it can be observed that the rankings are well distributed, and the dataset contains some stimuli of very low and very high intensity, with the majority falling somewhere in the middle. Variability, in contrast to intensity, is somewhat skewed to the left, suggesting that in the present dataset there are just a few videos that displays high variability in facial expression. Overall, we believe that the dataset is suitable for modeling intensity.

4 CONCLUSION AND FUTURE WORKS

We presented a novel dataset for studying the perception intensity in facial expressions, and the development of computational models of facial intensity. The code developed for this work is freely available [7]. In future, we plan to extend the ranking study with more stimuli. In parallel, we will develop a computational approach with machine learning methods for automatic estimation of the intensity. Our work may have several implications, mainly in Human-Computer Interaction (e.g., social robots able to interpret human nonverbal behavior), entertainment (e.g., video games with affective feedback), medicine (e.g., pain estimation), well-being, and research in psychology and cognitive science, among others.

REFERENCES

[1] Niki Aifanti, Christos Papachristou, and Anastasios Delopoulos. 2010. The MUG facial expression database. In *11th International Workshop on Image Analysis for Multimedia Interactive Services WIAMIS 10*. 1–4.

[2] Tadas Baltrusaitis, Peter Robinson, and Louis-Philippe Morency. 2013. Constrained local neural fields for robust facial landmark detection in the wild. In *Proceedings of the IEEE international conference on computer vision workshops*. 354–361.

[3] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. IEEE, 59–66.

[4] Abhinav Dhall, Roland Goecke, and Tom Gedeon. 2015. Automatic Group Happiness Intensity Analysis. *IEEE Transactions on Affective Computing* 6, 1 (2015), 13–26. https://doi.org/10.1109/TAFFC.2015.2397456

[5] Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. 2012. Collecting Large, Richly Annotated Facial-Expression Databases from Movies. *IEEE MultiMedia* 19, 3 (2012), 34–41. https://doi.org/10.1109/MMUL.2012.26

[6] P. Ekman and W.V. Friesen. 1982. Felt, false, and miserable smiles. *Journal Nonverbal Behavior* 6 (1982), 238–252.

[7] GitHub. [n. d.]. https://github.com/estiei/facial_expression_intensity. Accessed: 2024-03-02.

[8] Siti Khairuni Amalina Kamarol, Mohamed Hisham Jaward, Heikki Kälviäinen, Jussi Parkkinen, and Rajendran Parthiban. 2017. Joint facial expression recognition and intensity estimation based on weighted votes of image sequences. *Pattern Recognition Letters* 92 (2017), 25–32. https://doi.org/10.1016/j.patrec.2017.04.003

[9] Fanny Larradet, Radoslaw Niewiadomski, Giacinto Barresi, Darwin G. Caldwell, and Leonardo S. Mattos. 2020. Toward Emotion Recognition From Physiological Signals in the Wild: Approaching the Methodological Issues in Real-Life Data Collection. *Frontiers in Psychology* 11 (2020). https://doi.org/10.3389/fpsyg.2020.01111

[10] Shan Li and Weihong Deng. 2022. Deep Facial Expression Recognition: A Survey. *IEEE Transactions on Affective Computing* 13, 3 (2022), 1195–1215. https://doi.org/10.1109/TAFFC.2020.2981446

[11] S. Mohammad Mavadati, Mohammad H. Mahoor, Kevin Bartlett, Philip Trinh, and Jeffrey F. Cohn. 2013. DISFA: A Spontaneous Facial Action Intensity Database. *IEEE Transactions on Affective Computing* 4, 2 (2013), 151–160. https://doi.org/10.1109/T-AFFC.2013.4

[12] Radoslaw Niewiadomski and Catherine Pelachaud. 2010. Affect expression in ECAs: Application to politeness displays. *International Journal of Human-Computer Studies* 68, 11 (2010), 851–871. https://doi.org/10.1016/j.ijhcs.2010.07.004

[13] Radoslaw Niewiadomski and Catherine Pelachaud. 2012. Towards Multimodal Expression of Laughter. In *Intelligent Virtual Agents (Lecture Notes in Computer Science, Vol. 7502)*, Yukiko Nakano, Michael Neff, Ana Paiva, and Marilyn Walker (Eds.). Springer Berlin Heidelberg, 231–244. https://doi.org/10.1007/978-3-642-33197-8_24

[14] M. Pantic, M. Valstar, R. Rademaker, and L. Maat. 2005. Web-based database for facial expression analysis. In *2005 IEEE International Conference on Multimedia and Expo*. 5 pp.–. https://doi.org/10.1109/ICME.2005.1521424

[15] Prolific. [n. d.]. https://www.prolific.com. Accessed: 2024-03-02.

[16] Qualtrics. [n. d.]. https://www.qualtrics.com. Accessed: 2024-03-02.

[17] Ognjen Rudovic, Vladimir Pavlovic, and Maja Pantic. 2012. Multi-output Laplacian dynamic ordinal regression for facial expression recognition and intensity estimation. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. 2634–2641. https://doi.org/10.1109/CVPR.2012.6247983

[18] Evangelos Sariyanidi, Hatice Gunes, and Andrea Cavallaro. 2015. Automatic Analysis of Facial Affect: A Survey of Registration, Representation, and Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37, 6 (2015), 1113–1133. https://doi.org/10.1109/TPAMI.2014.2366127

[19] Md Taufeeq Uddin and Shaun J. Canavan. 2021. Quantified Facial Expressiveness for Affective Behavior Analytics. *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* (2021), 131–140.

[20] Robert Walecki, Ognjen Rudovic, Vladimir Pavlovic, Björn Schuller, and Maja Pantic. 2017. Deep Structured Learning for Facial Action Unit Intensity Estimation. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), 5709–5718.

[21] Jacob Whitehill, Zewelanji Serpell, Yi-Ching Lin, Aysha Foster, and Javier R. Movellan. 2014. The Faces of Engagement: Automatic Recognition of Student Engagementfrom Facial Expressions. *IEEE Transactions on Affective Computing* 5, 1 (2014), 86–98. https://doi.org/10.1109/TAFFC.2014.2316163

[22] Amir Zadeh, Yao Chong Lim, Tadas Baltrusaitis, and Louis-Philippe Morency. 2017. Convolutional experts constrained local model for 3d facial landmark detection. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 2519–2528.

[23] Rui Zhao, Quan Gan, Shangfei Wang, and Qiang Ji. 2016. Facial Expression Intensity Estimation Using Ordinal Information. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3466–3474. https://doi.org/10.1109/CVPR.2016.377