Inferring Emotional State from Facial Micro-Expressions

Alessandro Aiuti[a], Alessio Ferrato[a], Carla Limongelli[a], Mauro Mezzini[b] and Giuseppe Sansonetti[a]

[a]Department of Engineering, Roma Tre University, Via della Vasca Navale 79, 00146 Rome, Italy
[b]Department of Education, Roma Tre University, Viale del Castro Pretorio 20, 00185 Rome, Italy

Abstract

Personalized systems are becoming more and more popular in everyday life. Their goal is to adapt the output to the characteristics (i.e., interests and preferences) of the active user. To achieve this purpose, a process of inferring these characteristics is needed. In this paper, we verify the existence of some significant correlation between the facial micro-expressions of individuals and their emotional state. If so, we could think of monitoring the user while enjoying a certain visual stimulus, to understand her emotional response. For example, we could comprehend whether a visitor of a museum or an exhibition likes or dislikes the object she is observing, thus deriving her interests and tastes, regardless of the reality from which she comes. It could foster the role of the museum/exhibition intended as a vehicle of aggregation between a broad range of users, thus favoring their cultural and social inclusion. It could also allow us to design and realize recommender systems for enhancing the experience of users with difficulty in explicitly expressing their interests, such as people belonging to vulnerable groups (e.g., elderly, children, disabled people) or different cultures. Although the sample analyzed is limited and concerns a specific context (i.e., music video clips), the experimental results have been encouraging, thus spurring us to carry on with our research activities.

Keywords

User interfaces, Computer vision, Deep Learning, Museum visitors

1. Introduction and Background

Nowadays, technology accompanies and often affects our lives as individuals [1] and members of a large community [2]. For example, Machine Learning models and methods [3] (e.g., Deep Learning [4]) allow the realization of increasingly effective customized systems [5]. Among these, there exist also systems capable of integrating user profiles with additional information about their personality [6], their emotional state [7] as well as the temporal dynamics [8] and the actual nature [9] of their interests. Moreover, information related to the browsing activities on the Web can be considered in the user modeling process [10]. This work concerns the application of Computer Vision techniques [11] for the analysis of a user's facial micro-expressions while viewing video sequences, to identify her emotional state. The action units (AUs) are the individual components of muscle movements in which it is possible to break down facial expressions and, in certain combinations, allow us to analyze a person's emotional state. The study of the action units originated thanks to the research work by Ekman and Friesen who proposed the Facial Action Coding System (FACS) in 1978 [12], then updated in 2002 [13]. In the research literature, emotions are represented in various ways [14, 15]. In our research, we have chosen the model based on the valence-arousal scale proposed by Russel [16], which defines arousal (or intensity) as the level of autonomous activation that an event creates and the valence as the level of pleasure that an event generates. The research question underlying our research work is the following: is there a significant correlation between a user's facial micro-expressions and her level of arousal and valence? If this correlation were proven, it could be exploited for various purposes including in the context of personalized systems.

2. Method

Initially, we wanted to carry out a live analysis with real users (e.g., see [17]). We intended to monitor the user while viewing certain visual stimuli both through the RGB video recording of her facial expressions and the recording of her electroencephalogram (EEG) signal. Unfortunately, the pandemic situation we are currently experiencing has not allowed us to proceed as planned. We, therefore, decided to use datasets that are publicly available online [18]. In particular, we chose the DEAP dataset proposed by Koelstra *et al.* [19]. To collect the DEAP dataset, music videos were used as visual stimuli to arouse different emotions. One minute of each music video was selected, more specifically, the one with the

Joint Proceedings of the ACM IUI Workshops 2022, March 2022, Helsinki, Finland
ale.aiuti@stud.uniroma3.it (A. Aiuti);
ale.ferrato@stud.uniroma3.it (A. Ferrato);
limongel@dia.uniroma3.it (C. Limongelli);
mauro.mezzini@uniroma3.it (M. Mezzini);
gsansone@dia.uniroma3.it (G. Sansonetti)
0000-0003-4953-1390 (G. Sansonetti)

highest level of solicitation. A frontal face video was recorded for 22 participants while viewing those videos. Participants evaluated each video in terms of valence and arousal through the use of Self-Assessment Manikin with values ranging from 1 to 9. Once the dataset was obtained, it was necessary to process the videos obtained by recording the facial expressions of the users while they observed the music videos. There exist several facial recognition software tools, some of which are free. For this purpose, we used OpenFace[1], an opensource toolkit capable of capturing and analyzing the action units. Initially, we considered the mean and standard deviation of the action units for each user. Then, we calculated the correlation between those values and their positions in the Cartesian plane (see Figure 1). Considering the analysis of the average values, as regards the action units alone, the most significant correlations were found between AU-25 (lips part) and AU-17 (chin raiser) and between AU-01 (inner brow raiser) and AU-02 (outer brow raiser). On the other hand, we did not find any noteworthy correlation between the AUs and their positions in the quadrants. So setting aside this analysis and the quadrant classification, we decided to move on to the signal analysis of the various action units. Figure 2 shows the activation and variation of AU-12 (lip corner puller). We can see one activation and increase in the intensity of the AU signal when the user smiles and, therefore, the angle of his lips varies. Specifically, we have extracted several features of a higher order than the previous ones, which we have classified into four main types, as proposed in [20]:

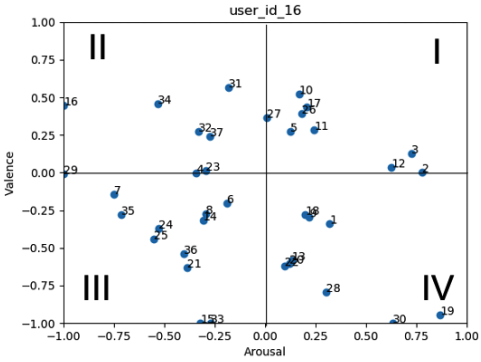

Figure 1: Distribution of valence-arousal pairs in the Cartesian plane for one of the test users.

- Statistical features
- Discrete features
- Dynamic features
- Quantitative features

The *statistical* features represent the first four statistical moments (mean, variance, skewness, and kurtosis) and

[1]https://github.com/TadasBaltrusaitis/OpenFace

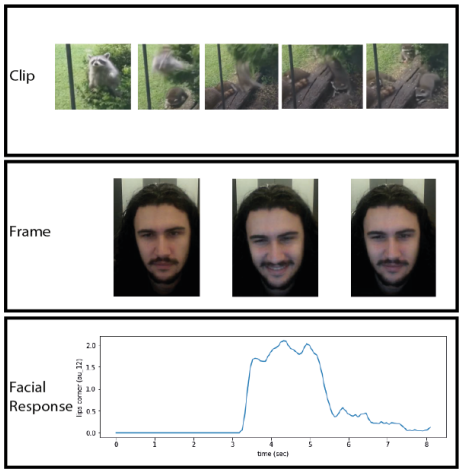

Figure 2: In the lower row, the intensity signal of the AU-12 (lip corner puller) related to the facial video (middle row) in response to the video clip shown in the upper row.

were computed for each AU intensity signal obtained from each facial recording. Then, we quantized the intensity signal over time for each action unit. The quantization was obtained through the k-means algorithm, with four clusters, a value obtained through the Elbow method. Figure 3 shows an example of a quantized signal. From each quantized signal, we calculated the following

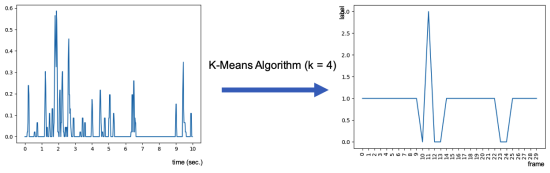

Figure 3: On the left, the intensity signal of a specific action unit, on the right, the signal quantized over time through k-means (with k=4).

three *discrete* facial activity features:

- Activation ratio: Percentage of frames with an AU activation;
- Activation length: Average number of continuous frames for which the AU is activated;
- Activation level: Average intensity of the AU activation.

For each quantized signal, we then generated a transition matrix that measures the number of transitions between the different levels. From this matrix, for each action unit, we extracted three *dynamic* features:

- Change ratio: Proportion of transitions between levels;

	mean	std	skewness	kurtosis	activation_ratio	activation_lenght	activation_level	change_ratio	slow_ratio	fast_ratio	act/frame	valence	arousal
mean	1	0.36	-0.5	-0.18	0.48	-0.35	0.56	0.54	0.27	0.38	0.63	0.18	0.37
std	0.36	1	-0.42	-0.26	0.4	-0.34	0.42	0.35	0.17	0.27	0.36	0.21	0.16
skewness	-0.5	-0.42	1	0.88	-0.3	0.4	-0.31	-0.49	-0.26	-0.34	-0.55	0.15	-0.28
kurtosis	-0.18	-0.26	0.88	1	-0.12	0.27	-0.1	-0.28	-0.14	-0.19	-0.36	0.22	-0.13
activation_ratio	0.48	0.4	-0.3	-0.12	1	0.03	0.9	0.5	0.34	0.31	0.55	0.03	0.32
activation_lenght	-0.35	-0.34	0.4	0.27	0.03	1	-0.1	-0.49	-0.18	-0.34	-0.59	-0.14	-0.26
activation_level	0.56	0.42	-0.31	-0.1	0.9	-0.1	1	0.56	0.22	0.38	0.6	0.06	0.31
change_ratio	0.54	0.35	-0.49	-0.28	0.5	-0.49	0.56	1	0.52	0.67	0.82	-0.64	0.74
slow_ratio	0.27	0.17	-0.26	-0.14	0.34	-0.18	0.22	0.52	1	0	0.43	-0.12	0.25
fast_ratio	0.38	0.27	-0.34	-0.19	0.31	-0.34	0.38	0.67	0	1	0.56	-0.45	0.53
act/frame	0.63	0.36	-0.55	-0.36	0.55	-0.59	0.6	0.82	0.43	0.56	1	0.05	0.81
valence	0.18	0.21	0.15	0.22	0.03	-0.14	0.06	-0.64	-0.12	-0.45	0.05	1	0.39
arousal	0.37	0.16	-0.28	-0.13	0.32	-0.26	0.31	0.74	0.25	0.53	0.81	0.39	1

Figure 4: Pearson correlation matrix between the extracted features and the arousal and valence values.

- Slow change ratio: Proportion of slow transitions (difference of one level);
- Fast change ratio: Proportion of fast transitions (difference of two or more levels).

As a *quantitative* feature, we considered the ratio between the number of activations of each action unit and the number of frames. The criterion with which we calculated the number of activations is that peaks occur when they have a value higher than the variance and a duration higher than three frames. Once all features were calculated, we determined the Pearson correlation matrix between the extracted features and the arousal and valence values. The resulting matrix is shown in Figure 4. Analyzing the correlation values between the extracted features, the three strongest correlations are between kurtosis and skewness, activation ratio and activation level, and change ratio and the number of activations per the number of frames (act/frame). As for the correlation values between the extracted features and the valence values, it can be noted that the strongest (negative) correlations occur between valence and change ratio, and valence and fast change ratio. As for the correlation values between the extracted features and the arousal values, the highest correlation values are with the number of activations per the number of frames (act/frame), change ratio, and fast change ratio.

The obtained findings are interesting because they show the possibility of inferring information relating to a user's emotional state by monitoring her facial expressions. From a practical point of view, this can make it possible to automatically derive the degree of appreciation that the user has towards the displayed object, without having to resort to long and annoying questionnaires. This information can be usefully exploited to design and develop recommender systems [21] capable of suggesting, for example, points of interest [22] and itineraries between them [23]. A possible application of the proposed approach could be in museums [24] and cities [25], intended as open-air exhibitions, because it would allow the automatic detection of visitors with similar tastes and interests [26], regardless of their social, demographic, and cultural peculiarities. In this way, museums and cities could represent possible inclusive places through the sharing of common interests and preferences [27].

3. Conclusions and Future Works

In this article, we have described an approach that consists in analyzing facial recordings of users subjected to visual stimuli to extract the action units relating to their facial micro-expressions. From these, we obtained some features and verified their correlation with the valence and arousal values.

Although the results obtained are encouraging and suggest that there is indeed a significant correlation between the features extracted from the action units and the user's emotional response, this study has some limitations. First of all, it was carried out on a small sample of users and video sequences. User reactions were collected using the Self-Assessment Manikin that, while having the advantage of simplifying the evaluation by the user, cannot be considered completely reliable. Furthermore, the users' reactions were collected by showing them music video clips and not, for example, artworks. Among future possible developments, there is, therefore, certainly a live analysis on real users, showing them different visual stimuli and collecting their emotional responses more accurately, for example, also using the EEG signal [28]. Moreover, in our study, we only considered valence and arousal. Further development could include, for example, the use of other emotional dimensions, such as likability and rewatch.

References

[1] M. Gordon, Solitude and privacy: How technology is destroying our aloneness and why it matters, Technology in Society 68 (2022) 101858.

[2] G. D'Aniello, M. Gaeta, F. Orciuoli, G. Sansonetti, F. Sorgente, Knowledge-based smart city service system, Electronics (Switzerland) 9 (2020) 1–22.

[3] L. Vaccaro, G. Sansonetti, A. Micarelli, An empirical review of automated machine learning, Computers 10 (2021).

[4] G. Sansonetti, F. Gasparetti, G. D'Aniello, A. Micarelli, Unreliable users detection in social media: Deep learning techniques for automatic detection, IEEE Access 8 (2020) 213154–213167.

[5] H. A. M. Hassan, G. Sansonetti, F. Gasparetti, A. Micarelli, J. Beel, Bert, elmo, use and infersent sentence encoders: The panacea for research-paper recommendation?, in: M. Tkalcic, S. Pera (Eds.), Proceedings of ACM RecSys 2019 Late-Breaking Results, volume 2431, CEUR-WS.org, 2019, pp. 6–10.

[6] M. Onori, A. Micarelli, G. Sansonetti, A comparative analysis of personality-based music recommender systems, in: CEUR Workshop Proceedings, volume 1680, CEUR-WS.org, Aachen, Germany, 2016.

[7] M. Tkalcic, B. De Carolis, M. de Gemmis, A. Odic, A. Kosir, Introduction to emotions and personality in personalized systems, in: M. Tkalcic, B. D. Carolis, M. de Gemmis, A. Odic, A. Kosir (Eds.), Emotions and Personality in Personalized Services - Models, Evaluation and Applications, Springer, 2017.

[8] S. Caldarelli, D. F. Gurini, A. Micarelli, G. Sansonetti, A signal-based approach to news recommendation, in: CEUR Workshop Proceedings, volume 1618, CEUR-WS.org, Aachen, Germany, 2016, pp. 1–4.

[9] D. Feltoni Gurini, F. Gasparetti, A. Micarelli, G. Sansonetti, Temporal people-to-people recommendation on social networks with sentiment-based matrix factorization, Future Generation Computer Systems 78 (2018) 430–439.

[10] F. Gasparetti, A. Micarelli, G. Sansonetti, Exploiting web browsing activities for user needs identification, in: Proc. of the 2014 CSCI, volume 2, 2014.

[11] A. Micarelli, A. Neri, G. Sansonetti, A case-based approach to image recognition, in: Proceedings of the 5th European Workshop on Advances in Case-Based Reasoning, EWCBR '00, Springer-Verlag, Berlin, Heidelberg, 2000, pp. 443–454.

[12] P. Ekman, W. V. Friesen, Facial action coding system, 1978.

[13] J. F. Cohn, K. Schmidt, R. Gross, P. Ekman, Individual differences in facial expression: stability over time, relation to self-reported emotion, and ability to inform person identification, in: Proc. of the 4th IEEE ICMI, IEEE Computer Society, USA, 2002.

[14] D. Kollias, S. Zafeiriou, Affect analysis in-the-wild: Valence-arousal, expressions, action units and a unified framework, CoRR abs/2103.15792 (2021).

[15] S. Zhao, G. Jia, J. Yang, G. Ding, K. Keutzer, Emotion recognition from multiple modalities: Fundamentals and methodologies, IEEE Signal Processing Magazine 38 (2021) 59–73.

[16] J. A. Russell, A circumplex model of affect, Journal of personality and social psychology 39 (1980).

[17] S. Park, S. W. Lee, M. Whang, The analysis of emotion authenticity based on facial micromovements, Sensors 21 (2021).

[18] Y.-H. Oh, J. See, A. C. Le Ngo, R. C. W. Phan, V. M. Baskaran, A survey of automatic facial micro-expression analysis: Databases, methods, and challenges, Frontiers in Psychology 9 (2018).

[19] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, I. Patras, Deap: A database for emotion analysis using physiological signals, IEEE TAC 3 (2012).

[20] D. Hadar, T. Tron, D. Weinshall, Implicit media tagging and affect prediction from rgb-d video of spontaneous facial expressions, in: Proc. of the 12th IEEE Int. Conf. on Automatic Face & Gesture Recognition, IEEE Computer Society, USA, 2017.

[21] G. Sansonetti, F. Gasparetti, A. Micarelli, F. Cena, C. Gena, Enhancing cultural recommendations through social and linked open data, User Modeling and User-Adapted Interaction 29 (2019) 121–159.

[22] G. Sansonetti, Point of interest recommendation based on social and linked open data, Personal and Ubiquitous Computing 23 (2019) 199–214.

[23] A. Fogli, G. Sansonetti, Exploiting semantics for context-aware itinerary recommendation, Personal and Ubiquitous Computing 23 (2019) 215–231.

[24] A. Ferrato, C. Limongelli, M. Mezzini, G. Sansonetti, Using deep learning for collecting data about museum visitor behavior, Applied Sciences 12 (2022).

[25] D. D'Agostino, F. Gasparetti, A. Micarelli, G. Sansonetti, A social context-aware recommender of itineraries between relevant points of interest, in: HCI International 2016, volume 618, Springer International Publishing, Cham, 2016, pp. 354–359.

[26] F. Gasparetti, G. Sansonetti, A. Micarelli, Community detection in social recommender systems: a survey, Applied Intelligence 51 (2021) 3975–3995.

[27] K. Coffee, Cultural inclusion, exclusion and the formative roles of museums, Museum Management and Curatorship 23 (2008) 261–279.

[28] F. Galvão, S. M. Alarcão, M. J. Fonseca, Predicting exact valence and arousal values from eeg, Sensors 21 (2021).