Measuring the induction of affect using facial expression analysis technology: a pilot study

Marko Meža[1], Andrej Košir[1], Gregor Strle[1,2]

[1]User-adapted Communication and Ambient Intelligence Lab, Faculty of Electrical Engineering, University of Ljubljana
[2]Research Centre of the Slovenian Academy of Sciences and Arts
E-pošta: marko.meza@fe.uni-lj.si

Abstract. *The paper presents a pilot study investigating usefulness of facial expression analysis technology for measuring emotion induction. The study involved six male participants and a subset of 85 images from the International Affective Picture System dataset (IAPS), rated on the valence and arousal dimensions of affect. The results show low reliability and agreement, both when comparing the mean induction ratings gathered from the face reader to the ground truth (the IAPS ratings), as well as when comparing the induction ratings between the six participants. We conclude that the presented setup does not adequately measure the induction of affect and that the time interval for measuring participant responses from the stimulus onset should be carefully selected.*

1 Introduction

Non-verbal cues, particularly affect and emotions, are critical to maintaining, interpreting and understanding social signals, communication, and behavior among humans [1]. One of the challenges of the current state-of-the-art is the analysis and prediction of affective cues using sensor technology and machine learning in different setting, from researches in the wild or based on the existing datasets [2–6].

One area of research in our laboratory is a development of sensor technology framework that would support reliable measurements of emotional, behavioral, and social cues. To this end, the paper presents a pilot study investigating usefulness of facial expression analysis technology for measuring emotion induction.

2 Method

2.1 Participants

The pilot study was conducted on 6 young male participants enrolled in the curriculum at the The Faculty of Electrical Engineering, University of Ljubljana. The tasks given to the participants were a part of a laboratory assignment created to investigate how affect can be measured with sensor technology.

2.2 Materials

Materials from the International Affective Picture System dataset [5] were used in the study. The IAPS dataset is a widely used stimulus set for emotion research, based on the dimensional model of affect [7]. It contains 1194 images with short descriptions, rated along the dimensions of valence and arousal. To make this pilot study feasible, a subset of 85 images was chosen by manually selecting samples with high, medium and low arousal ratings. In order to achieve this, we ordered the database in the descending order of arousal ratings and selected top, medium and bottom part of the samples for each group from the database. Some images were further removed due to their explicit content.

2.3 Recording apparatus

The recording apparatus consisted of a laptop computer with an image playback software, a real-time face expression analysis software, and a control script.

The *image playback application* (Figure 1) was developed for the purpose of this study and used to present the participant the 85 IAPS images, in the same order, one after another, with each image presented for 5 seconds. The application stores exact timestamps of when each individual image was shown to the user.

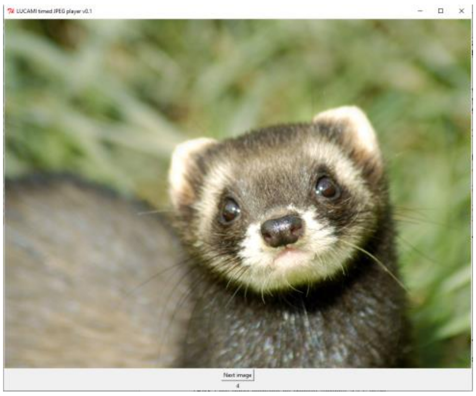

Figure 1. Image playback application showing an IAPS image from the test session.

Noldus FaceReader [8], a *facial expression analysis software*, was used for measuring the induction of affect. The Noldus face reader computes facial emotion expressions in real time, from the video recording of the participant. The software records the valence and arousal, as well as timestamps (Figure 2). For our setup, we recorded between 10 and 12 data samples per second. The Noldus API was used to control the recording sessions and analysis of the data. A *data logging application* was developed in the course of the study to obtain the face reader data and store it into a .csv file for further analysis.

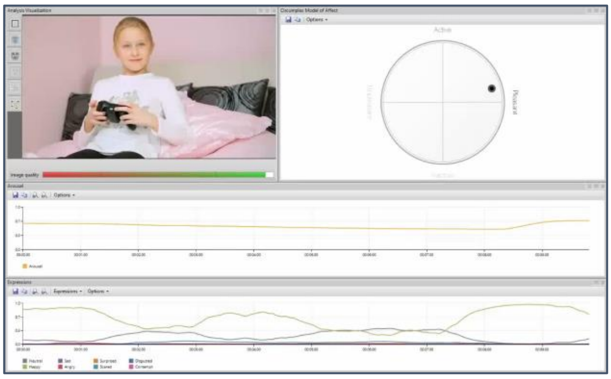

Figure 2. A screenshot of Noldus Face reader software

All components of the recording setup were controlled by a single *control script*. The script started the Noldus face reader software, initiated the analysis, started the data logging application and the image playback application, as well as finished the recording sessions by terminating all software components used during the sessions.

The output for each recording session includes two .csv files, one containing the timestamps of the presented images and the other containing the face reader timestamps and valence and arousal ratings sampled approximately 10 times per second.

2.4 Recording procedure

The participants were first informed about the aim and details of the pilot study and the experimental procedure. They were instructed to act normally and to not exaggerate and play their expressions. Each participant was subjected to a test session, during which they became accustomed to the recording apparatus and the experiment procedure. After the test trial, each participant was set in a quiet environment behind a laptop computer used for the experiment. The participant was given time to relax and told to start the recording session when ready. After each session, the recorded data were first archived, then the experimental setting was prepared for the next participant.

2.5 Statistical analysis

The IAPS ratings were first normalized to correspond to the valence-arousal measuring scale (-1,1) used by the Noldus face reader. For the Noldus face reader data, the mean valence and arousal ratings were taken for each IAPS image the participant was exposed to, in order to assess the reliability among the participants. The overall mean valence and arousal Noldus ratings were also calculated for each IAPS image, in order to evaluate the reliability of the face reader ratings with the ground truth (IAPS ratings).

Independent Samples *t*-test (Welch's *t*-test for unequal variances) was used to test the means of both groups of ratings (IAPS vs. Noldus), and a one-way Anova was used to test the difference between the face reader participants for the valence and arousal dimensions. The intraclass correlation (ICC) was used to assess the reliability of the ratings, both between the IAPS (the ground truth) and the Noldus ratings, as well as between the participants using the Noldus face reader.

3 Results

A comparison of the mean ratings from IAPS and the mean ratings from the Noldus face reader shows significant differences, both for valence and arousal. The IAPS ratings for the valence (M = 0.21, SD = 0.43) are significantly different than the Noldus ratings (M = 0.06, SD = 0.03), with t = 3.19, p < .002. Similarly, significant difference was found for the arousal, between the IAPS ratings (M = 0.01, SD = 0.34) and the Noldus ratings (M = 0.32, SD = 0.01), with t = 8.3, p < .001.

The significant difference in the distribution of the ratings between both groups is shown also in Figures 4 and 5. The low value of the ICC shows there is no agreement in the ratings between the IAPS and the Noldus face reader. The *ICC*(1,1)= -0.004 for valence is extremely poor, with the 95% confidence interval for the ICC (-0.2, 0.2). Similarly, the agreement for the arousal is also extremely poor, with the *ICC*(1,1)= -0.27 and the 95% confidence interval for the ICC (-0.45, 0.06).

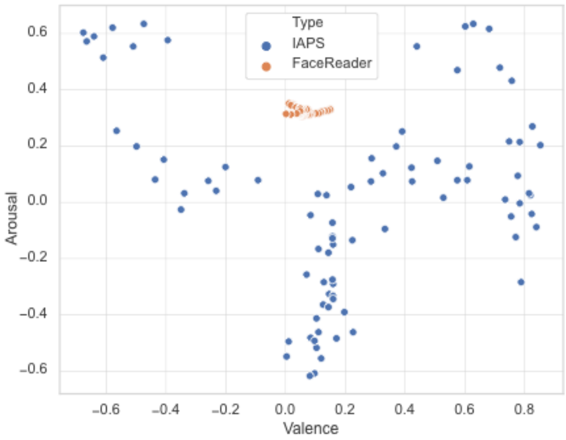

Figure 4. Valence-arousal space of the mean IAPS (blue) and face reader ratings (orange).

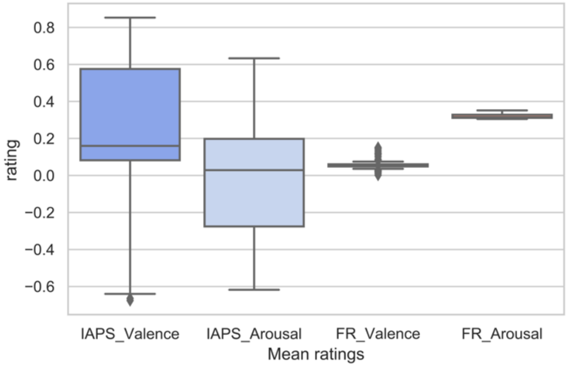

Figure 5. Distribution of mean valence/arousal ratings for IAPS and Noldus (FR).

The agreement is also low between the Noldus participants. The *ICC*(2,1)= 0.04 for valence is extremely poor, with the 95% confidence interval for the ICC (0.002, 0.08). Similarly, for the arousal, with the *ICC*(2,1)= 0.02 and the 95% confidence interval for the ICC (-0.002, 0.05). The one-way Anova on the Noldus participants' ratings shows significant differences among the participants for both valence and arousal, with $p < .001$. The differences in valence and arousal ratings between the participants are shown in Figure 6. We can observe that the valence ratings are consistently lower and with several outliers, as compared to the arousal ratings.

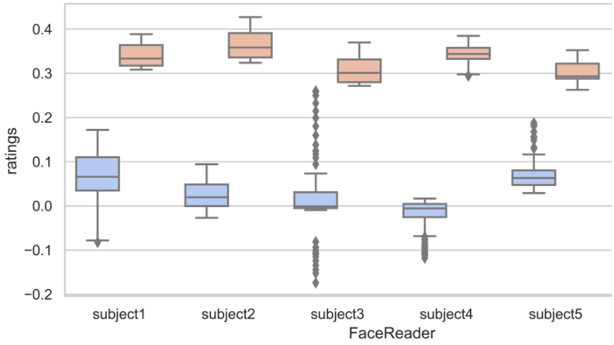

Figure 6. Distribution of the participants' mean valence/arousal ratings (Noldus face reader).

4 Discussion

The presented pilot study investigated the induction of affect as measured by the Noldus face reader. The results show low reliability and agreement, both compared to the ground truth (the IAPS dataset), as well as among the participants of the study.

For the presented setup, the results generated by the Noldus face reader do not adequately measure the emotional induction from images (e.g., see Figures 4 and 5). We speculate the main reason for this are the mean ratings taken of the entire time interval each image was presented to the participant, as the distributions of the valence and arousal ratings are small for all the participants. It might be more appropriate to sample the ratings in a narrower time interval starting from the onset an image is shown, as the face reader recordings show higher values for the initial responses to the stimuli, for both dimensions, which then decrease over the five second timespan.

It is also important to note that, in part, the ratings are influenced by the individual's personality and their facial expressions which vary among the participants, with some participants being more expressive than others. However, as shown by the results (see also Figures 4 and 5), the low reliability and agreement with the ground truth (IAPS) cannot be ascribed to differences in emotional expressivity among the six participants, as the distribution of the ratings is extremely low.

In our future work, the presented study will be extended with a larger sample and, particularly, with in-depth analysis of the time interval that could potentially better capture the induction of affect.

Literature

[1] Vinciarelli A, Pantic M, Bourlard H. Social signal processing: Survey of an emerging domain. Image and vision computing. 2009 Nov 1;27(12):1743-59.

[2] Sawyer R, Smith A, Rowe J, Azevedo R, Lester J. Enhancing student models in game-based learning with facial expression recognition. InProceedings of the 25th conference on user modeling, adaptation and personalization 2017 Jul 9 (pp. 192-201).

[3] Mollahosseini A, Hasani B, Mahoor MH. Affectnet: A database for facial expression, valence, and arousal computing in the wild. IEEE Transactions on Affective Computing. 2017 Aug 21;10(1):18-31.

[4] Kollias, Dimitrios, et al. "Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond." International Journal of Computer Vision 127.6-7 (2019): 907-929.

[5] Lang, PJ.; Bradley, MM.; Cuthbert, BN. International affective picture system (IAPS): Technical manual and affective ratings. University of Florida, Center for Research in Psychophysiology; Gainesville: 1999.

[6] Stöckli, S.; Schulte-Mecklenbeck, M.; Borer, S. & Samson, A.C. Facial expression analysis with AFFDEX and FACET: A validation study. Behavior Research Methods, 50 (4), (2018): 1446-1460.

[7] Russell JA. A circumplex model of affect. Journal of personality and social psychology. 1980 Dec;39(6):1161.

[8] Noldus FaceReader: Tool for automatic analysis of facial expression: Version 6.0. Wageningen, the Netherlands: Noldus Information Technology B.V. (2014).