Evaluation of a System for Real-Time Valence Assessment of Spontaneous Facial Expressions

Kuderna-Iulian Benţa[1], Hans van Kuilenburg[2], Ulises Xolocotzin Eligio[3], Marten den Uyl[4], Marcel Cremene[5], Amalia Hoszu[6] and Octavian Creţ[7]
[1]Comm.Dept.,Technical University of Cluj-Napoca, Romania,
Iulian.Benta@com.utcluj.ro
[2]VicarVision, Amsterdam, The Netherlands
kuilenburg@vicarvision.nl
[3]Learning Sciences Research Institute, University of Nottingham, United Kingdom
lpxux@nottingham.ac.uk
[4]VicarVision, Amsterdam, The Netherlands
kuilenburg@vicarvision.nl
[5]Comm.Dept.,Technical University of Cluj-Napoca, Romania,
Marcel.Cremene,@com.utcluj.ro
[6]hoszu_amalia@yahoo.com
[7]Comm.Dept.,Technical University of Cluj-Napoca, Romania
Octavian.Cret@cs.utcluj.ro

Abstract. Over the last years many solutions have been proposed for facial expression analyses for emotion assessment, some of which claim high recognition accuracy in real time, even on natural expressions. However, in most (if not all) cases, these studies do not use a publically available and validated dataset for inducing emotions and fail to place the reported performance in the right context by comparing against human expert scoring performance. We evaluate the abilities of FaceReader, a commercially available product for fully automatic real-time emotional expression recognition, to detect spontaneously induced emotions. We presented a balanced subset of IAPS pictures to a number of test subjects and compared the automatically detected emotional valence as reported by FaceReader to the IAPS valence labels and to the evaluation of three psychologists.

Keywords: Emotional Expression Recognition, Emotional Valence, Affective Computing, IAPS, Spontaneous Elicited Emotions, Evaluation.

1 Introduction

The usefulness and wide spread of computing devices led to the appearance of a new social partner that follows us everywhere and so cannot be neglected anymore: the electronic buddy. Sometimes this buddy is not as nice or as smart as one may expect, which is partly due to its lack of affective sensitivity.

The state of the art solutions in emotion assessment are getting near to the point where they can detect the user's emotions in real-life situations and provide them as feedback to the device. But how near are they? How can we evaluate these emotion detectors' valence accuracy in real-life like situations? We evaluated the valence assesment accuracy of FaceReader, a widely used tool for facial expression assessment, using spontaneous facial expressions elicited with images from the IAPS (International Affective Picture System) [2].

The structure of this paper is as follows. In section 2 we review and compare the literature regarding other valence assessment tools and evaluation methods. Section 3 briefly describes FaceReader. Then, in section 4, we present the results of a first experiment in which we selected a set of standardized emotion inducing pictures and presented them to a group of subjects while FaceReader analyzed their facial expressions. In section 5 we present a second experiment, where we compared the data logged by FaceReader with the analysis made by human experts. We conclude our work in section 6 and indicate our future work intentions.

2 Related Work

The interest in the assessment of spontaneous displayed emotions is very important for real-time human-computer applications, as emphasized in the literature [1]. However, the number of existing real-time non-invasive emotion assessment tools described in other papers that recognize emotional facial expressions is small and lack performance evaluations for spontaneous elicited emotions. FaceReader is one of the few publically available automatic facial expression recognition systems with advanced analysis and reporting functions. Although FaceReader has been reported to be quite accurate on posed expressions [2], there has not yet been a comparitive and quantitive study on its ability to detect spontaneous emotions.

A user's emotional valence [3] is important for human machine interaction as it generally characterizes the emotional response in terms of acceptance (+), rejection (-) or neutral (0) but is rarely addressed in non-invasive emotion assessment methods like facial expression based ones. Recent publications [1, 4, 5] claim remarkable results on basic emotion assessment in real-time of spontaneous elicited emotions. In [1] video trailers were watched by 28 subjects and facial features captured by a hidden camera were classified in real time in four states (neutral, joy, surprise and disgust). In the article [4] spontaneous emotions were elicited by short films for 41 participants and measured by face features and physiological parameters (partially invasive), but only amusement and sadness as opposite emotions were considered. A similar approach [5] used emotionally-loaded films on 21 subjects and automatically classified them in six basic states based on facial features. These are promising results. But the properties of the emotion induction stimuli used to evaluate the accuracy at emotion assessment are unknown, which makes it hard to attach any meaning to the reported performances. To fill this gap, we decided to use a subset of images from, which is a set of emotion inducing images which are well characterized in terms of valence, arousal and dominance.

3 FaceReader System Description

FaceReader is a system for fully automatic real time facial expression analysis developed by VicarVision and commercially available since 2007. It is currently used worldwide for numerous (consumer) behaviour studies. The software tool can process still images, video and live camera feeds and produces approximately 15 analysis results per second on a modern PC, allowing it to be used in real-time. FaceReader can classify expressions corresponding to one of the 6 basic emotions as defined by Ekman in 1970 [7] plus neutral and also classifies the emotional valence of the expression and some personal characteristics like gender, age and ethnicity.

A detailed description of the technology used in the FaceReader is beyond the scope of this article, but can be found in [6]. In a nutshell, a deformable template matching method is used to find faces after which a photorealistic model of the depicted face is created using a proprietary implementation of the Active Appearance Model [8]. Finally, classification is performed by feeding the parameters of these models to a battery of neural network classifiers.

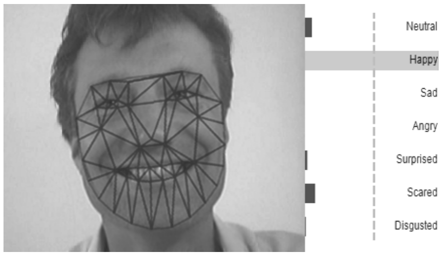

Fig. 1. Example of shape and triangulation derived by the Active appearance model

A trained appearance model has limits on the amount of variation that it is able to model, which manifests itself for example as a lower modelling accuracy (or failure) for people of certain ethnicities and age and difficulties modelling faces under certain lighting and orientation. FaceReader therefore includes several user-selectable models which each provide an optimal performance for a different subset of individuals and conditions. Live (camera) analysis results containing continuous values for all basic emotions and valence can be written to a log with a 200 ms interval and are accessible on-the-fly, allowing integration with other software for direct user interaction applications.

4 Experiment 1: Basic emotions induction

4.1 Emotion induction pictures

Our objective is to evaluate the accuracy of FaceReader at detecting induced emotions as occurring in real-time. Because there is no standard way of inducing emotions, the first experiment was aimed to determine a methodology to do so in a spontaneous and yet systematic way. As emotion-induction stimuli, we used pictures from the International Affective Picture System (IAPS). The picture set of the IAPS is well distributed over the valence and arousal dimensions [2]; the more frequently studied dimensions in emotion research. But FaceReader assesses the probability values for each of the basic emotions (happy, angry, sad, disgusted, neutral, surprised, and scared). In comparison to valence and arousal properties, the relationship between the IAPS pictures and specific basic emotions is less clear. Therefore, in order to obtain statistically and conceptually correct data, our first step was to make a picture selection.

Our objective is to have a wide set of emotion induction pictures but also needed to avoid fatigue of the participants. Therefore we reduced the 1182 pictures of the IAPS to a reasonable number to be reviewed by a person in one session. We decided that 31 pictures per emotional state should be adequate to balance between statistically enough data and fatigue avoidance. To start with, a set of 217 pictures (31 pictures per state * 7 emotions) was manually selected on potential to induce the basic emotions based on the personal criteria of one co-author.

Experiment 1a. Participants were 31 students (19-24 years old, 17 male and 14 female) with the same computer science background. To evaluate, we developed a web based application that displayed: a picture, a scale with the seven basic emotions to classify the emotion induced by the picture and the valence SAM (Self-Assessment Manikin) [9].

Pictures were presented in random order. We asked the subject to classify each picture as inducing one of the seven basic emotions and rate its inducting valence. The session duration was about 30-40 minutes/test. When analysing the results we concluded:

1. Subjects were able to make a decision on the emotion induced by most of the pictures (197). In comparison, participants could not make a decision about the emotion induced for only 20 pictures. This indicates that overall, the selected picture set allows for discrimination of discrete basic emotions, therefore valid to our testing purposes.
2. There were pictures rated by the majority of the subjects as being inductors for just one emotional state (the average score for that emotion was bigger then for all other six basic emotions)
3. Surprise, anger and sadness had a small number of pictures to induce them, less than 31. This was more extreme for surprise, since only 2 pictures were rated by the majority of the responders as inducing surprise.

Picture set refinement. Further data treatment was meant to create an even distribution of 31 pictures for each of the 7 basic emotions.

For the more represented emotional states (happy, neutral and scared), we took off the less rated pictures to finally get to 31 pictures for each state. For the less represented emotional states, we added twice the number of needed pictures. A total of 137 pictures were added to the initial picture set. These new pictures were rated by 13 persons. Then we selected the best rated pictures (with the highest score) for each state, e.g. the best 6 from 13 in case of disgust.

As a result we should have had, after the second selection, 31 pictures/emotion * 7 basic emotions = 217 a new pictures set of potentially inducing basic emotions.

Experiment 1b. Participants were 14 students from the same category (8 males, 6 females), age 23-24. They used the same web page for rating the pictures as in experiment 1. All the seven basic emotions could be rated, not just the 4 (disgust, sad, surprised, angry) we were looking for and the instructions where to rate one out of seven on a seven scale category.

The results of this stage where consistent with experiment 1b: we got one more picture rated as 'anger' out of 47, 9 pictures labeled as 'surprise' from 63 and the 9 needed pictures for 'sad'. The results for 'anger' and 'sad' are probably explained by people's tendency to experience or report these emotions with low frequency [10, 11]. The remark of one person seems to explain the results for 'surprise': "how can a picture be surprising if each of the presented pictures already is new and appears suddenly?"

In the end we selected a set of 175 IAPS pictures. We tried to have an even number of pictures to represent each basic emotional state. But the experiment results and reviewed literature suggested that this may be quite difficult to do. Moreover, it may not reflect how these emotions occur in real life scenarios, e.g., in everyday life episodes of surprise tend to be less frequent than episodes of neutral or happy emotional states. We measured a 0.916 correlation to IAPS in terms of valence as described in next section.

Experiment 1c. We investigated the effect of inducing the basic states in a set-up (see Fig. 2) with 16 students and noticed the following errors:

a. The moment of image presentation and the interval during which the subject was observing the presented image could not be determined, due to internet connection speed limitations and the fact that the self-reporting task was done on the same monitor.

b. The needed processing power for FaceReader was less than necessary (running on an older laptop) and that led to a lot of missing values.

c. The on-top-of-the-monitor camera position (see Fig. 2) biased the results towards 'anger'

d. The self-evaluation task that followed each of the emotion induction pictures confined the subject to a more 'neutral' or 'anger'-like face.

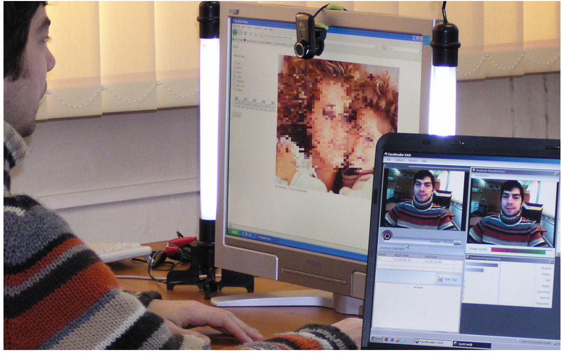

Fig. 2. The student experimental setup

In the experiment 1c, the user's self-reported their basic emotion (one force choice out of seven) and the valence (one force choice out of 9 SAM levels) was recorded.

The correlation between the IAPS valence scores and the valence self-report was 0.916.

Table 1 shows the distribution of the self-reported valence for the self-reported emotion. You may see that the negative basic emotions (disgust, scared, anger and sad have negative values) while the only positive basic emotion has a positive mean valence value (2.125).

Table 1. The distribution of the valence for the self-reported emotion

	Disgust	Happy	Scared	Angry	Neutral	Surprise	Sad
Mean	-2.117	2.125	-1.912	-1.439	0.195	0.45062	-1.9
SD	1.5405	1.39	1.645	1.707	1.152	2.00016	1.59

5 Experiment 2: FaceReader and human observer evaluation and comparison

5.1 Generation of facial expression data

We used the methodology for emotion induction used in experiment 1 with 16 subjects ranging from 23 to 55 years (13 males, 3 females), most of whom have a computer science background. Each participant was asked to watch the pictures and to react naturally Participants were left alone in the room for the next 13 minutes. Each person sat on a chair in front of a widescreen in such way that the camera could clearly ‘see’ his/her face but still being able to move leaning forwards and backwards, right-left. The lightning conditions were controlled by two neon tubes that distributed the light evenly on the subjects face. The subjects were video recorded while watching the 175 emotion induction picture slide show displayed in a random order.

5.2 Individual differences

We noticed very large individual differences in the facial expressions and behaviours of the subjects while taking part in the experiment. Both the emotional sensitivity as well as the ability of people to express emotions throught facial expressions is known to vary [12], which was reflected clearly in the experimental results. While some subjects showed very clear variations in expressions throughout the experiment and reported feeling a bit ‘shaken’ by the emotions they had felt, others showed nearly no variations in facial expression at all during the experiment. Some subjects even showed emotional expressions clearly opposite to the emotions that were meant to be evoked by the shown IAPS image. These subjects reported that they had experienced images meant to evoke disgust as funny or entertaining.

In addition, we observed some head gestures reflecting strong negative emotions. One subject in particular kept turning her head for all disgusting images. This also forced us to eliminate this subject for further analysis, as FaceReader requires near-frontal face images only.

5.3 FaceReader scoring

The recorded videos were analyzed by FaceReader, producing a continuous emotional output signal for the entire experiment for each subject. Although FaceReader is able to analyze single still-images, the dynamic information contained in a video will increase the accuracy of the analysis. These continuous output values (for valence and the basic emotions) provided us with some interesting additional insights and complications to further deal with.

First, we noticed a significant emotional overflow effect into the next displayed image. Fig. 3 shows the mean correlation between the IAPS value of the presented image and the FaceReader valence output from the moment of image presentation. At 4.5 seconds, a new image is shown and at 9 seconds again a new image is shown. The emotion/expression onset is clearly visible, indicating that FaceReader valence indeed correlates to the emotional content of the presented image, but it takes rather long to reach a maximum after about 4 seconds. Also, we notice that the expression is shown far into the next image presentation and even after that, the correlation remains significant.

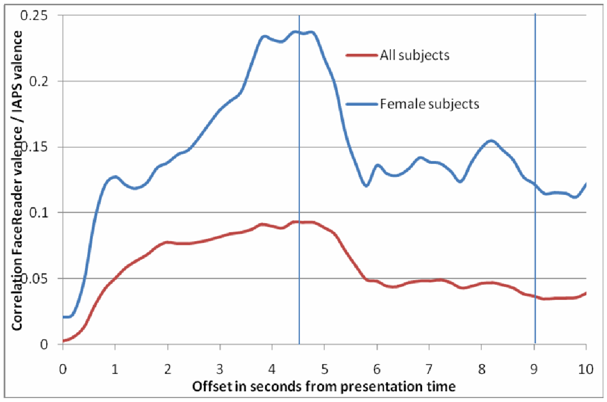

Fig. 3. The emotion in time evolution after each elicitation picture display

Just how significant this overflow effect is, became clear by observing the mean FaceReader valence output in the beginning and in the end of the experiment. Where the subjects mostly showed a slightly negative mean valence at the beginning of the experiment (caused by having more images with a negative content than images with a positive content), many subjects showed a strongly negative mean valence in the middle and last part of the experiment. This strongly suggests that evoked emotions to a large extend add up to previous emotions and we must thus conclude that at any point in the experiment, an emotion is shown that reflects not just the currently displayed IAPS image, but a complex sum of all images shown up until that point. Fig. 3 also helps us to identify the moment after image presentation that can best be used to sample the valence and use for further analysis, which is about 4 seconds after image presentation.

As discussed in section 3, there are several appearance models with corresponding classifiers available in the FaceReader and each may work better or worse under specific circumstances and for specific individuals. We have therefore analyzed the videos with two different models, both included in the default FaceReader product and both designed for a wide range of use and for people of all ethnicities.

We calculated per subject the correlation of FaceReader's valence output value to the IAPS valence of the image that was used to induce the emotion. The results are shown in the Table 2.

Table 2. The FaceReader's – IAPS valence correlation per subject

Subject nr	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Model1 corr	0.46	-0.08	0.17	0.03	0.15	0.00	-0.12	0.02	0.38	0.18	0.25	0.12	0.25	0.23
Model2 corr	0.40	-0.03	0.02	0.09	0.07	-0.10	-0.03	-0.10	0.45	0.19	0.26	0.17	0.21	0.20

As can be seen, the correlation ranges from slightly negative to medium positive. In section 5.2 we discussed the individual differences we noticed. These differences are reflected in the results above, as the negative correlations correspond to the subjects who tended to laugh/smile when looking at images that were meant to evoke feelings of disgust, the near-zero correlations correspond to subjects who showed little to no difference in expression at all during the experiment.

5.4 Human observer scoring

We sampled still images from the recorded videos to be analyzed by human observers at the exact same points in time where FaceReader valence values were sampled, building on the assumption that human observer valence correlation would peak at the same moment.

We used a subset of 35 face images for each subject from the total of 175 corresponding to the emotion induction pictures. The same web application was adapted for the human observers to rate the face images by selecting one of the 9 valence levels using SAM. Again we notice the individual differences between subjects by analysing the data presented in Table 3.

Table 3. Human observer – IAPS valence correlation per subject

Subject nr	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Observer1 corr	0.30	-0.16	0.13	0.03	0.17	0.20	0.18	0.09	0.47	0.19	0.17	0.34	0.22	0.25
Observer2 corr	0.24	0.18	0.17	0.08	0.23	0.20	0.03	0.25	0.74	0.45	0.18	0.43	0.01	0.24
Observer3 corr	0.17	0.24	0.15	-0.19	0.07	0.02	-0.03	0.20	0.60	0.42	0.20	0.36	0.23	0.23

5.5 Comparison between FaceReader and human analyses of emotional facial expression

Both the scoring by FaceReader as well as the human expert scoring produced correlations which are far lower than we initially expected and are lower than those reported in other studies [1, 4, 5]. We therefore have to conclude that the experimental setup or the material used only managed to evoke a mild emotional response which was often not reflected in the subject's facial expression. Additionally, the large overflow effect of emotions evoked by previous images will have introduced a great amount of noise. Calculating the correlation between the predicted valence values (of either human observers or FaceReader) and the current image IAPS value plus the IAPS valence value last 3 presented images (weighted using least squares analysis), gives correlation score of up to 80% higher.

A direct comparison between the correlation scores of the two FaceReader models and the 3 human observers shows that for some subjects, FaceReader outperforms all or some of the human observers, but when considering the mean correlation for all 14 subjects, the human observers each have a higher mean correlation, see Table 4.

Table 4. Comparison between the FaceReader and three human observers

14 subjects	FaceReader model 1	FaceReader model 2	Human observer 1	Human observer 2	Human observer 3
Mean correlation score	0.14	0.13	0.19	0.24	0.19

However, this score includes the correlation values for subjects who clearly showed little to no expressions at all, plus those who were smiling at many negative images. In real-life applications, these subjects would most likely not have been selected in a pre-screening as their data introduces noise which will be highly biased by the image selection.

Removing the 4 individuals who showed little or inverse emotional expressions (by observing the videos and paying special attention to the episodes when the most negatively valanced images were presented), the mean scores will look as you may see in Table 5.

Table 1. Mean correlation scores to IAPS valence for 10 selected subjects

10 subjects	FaceReader model 1	FaceReader model 2	Observer 1	Observer 2	Observer 3
Mean correlation score	0.22	0.21	0.23	0.28	0.22

This puts FaceReader in the same league as two of the observers, where only observer 2 scores significantly better.

One more interesting observation is that if FaceReader would be able to automatically asses which model works best for each individual, the mean correlation score would go up to 0.24 for 10 subjects. However, whether this selection can be made automatically based on modeling errors or emotional expression variation remains to be investigated and is not a current FaceReader feature.

Acknowledgments. This work benefits by the support of the UEFISCSU PN2 "Idei" code 1062.

6 Conclusions and Future Work

We defined by experiments a subset of well balanced IAPS pictures and we used them to induce spontaneous emotions. We measured the elicited emotions' valence for each subject using FaceReader, a face feature based emotion assessment tool. Analyzing the data obtained we concluded that human observers score valence slightly more accurate than FaceReader, but the correlation difference is rather small. For emotional expression recognition applications where traditionally human observers would be used, like in usability studies, this means that though a few more observations or test subjects might be needed in order to produce statistically significant and reliable results, this will be more than compensated by the fact that automatic analysis can produce results at far lower costs in time and resources, cutting analysis time by a factor 10 and making the use of trained experts obsolete.

In order to use an automatic recognition tool directly in a human-computer interfacing, the found accuracies will have to be further improved. We intend to investigate the possibility of personalizing FaceReader's emotion valence assessment based on the large individual differences we observed in the ability to show emotions and form in which the emotions are shown. Just as human observers are best at judging the facial expressions of people they know or have observed for a longer time, an automatic tool like FaceReader could benefit from some person specific calibration or training.

References

1. Sebe, N., Lew, M.S., Cohen, I., Yafei S., Gevers, T., Huang, T.S.: Authentic facial expression analysis, Automatic Face and Gesture Recognition, 2004. Proceedings. Sixth IEEE International Conference on, pp. 517--522 (2004)
2. Lang, P.J., Bradley, M.M., Cuthbert, B.N.: International affective picture system (IAPS): Affective ratings of pictures and instruction manual. Technical Report A-8. University of Florida, Gainesville, FL. (2008)
3. Colombetti, G.: Appraising valence, J. Consciousness Studies, no. 8/10, pp. 103--126 (2005)
4. Bailenson, J.N., Pontikakis, E.D., Mauss,I.B., Gross, J.J., Jabon, M.E., Hutcherson, C.A.C., Nass, C., John, O.: Real-time classification of evoked emotions using facial feature tracking and physiological responses, Int. J. Human-Computer Studies, vol. 66, pp. 303–-317 (2008)
5. Yeasin, M., Bullot, B., Sharma, R.: Recognition of Facial Expressions and Measurement of Levels of Interest From Video, IEEE Transactions On Multimedia, vol. 8, no. 3 (2006)
6. van Kuilenburg, H., Wiering, M., den Uyl, M.: A model-based method for automatic facial expression recognition. Springer LNAI, Vol. 3720, pp. 194—205 (2005)
7. Ekman, P.: Universal facial expressions of emotion, California Mental Health Research Digest, 8, pp. 151–-158 (1970)
8. Cootes, T., Taylor, C.: Statistical models of appearance for computer vision, Technical report, University of Manchester, Wolfson Image Analysis Unit, Imaging Science and Biomedical Engineering (2000)
9. Lang, P. J.: Behavioral treatment and bio-behavioral assessment: Computer applications, In J. B. Sidowski, J. H. Johnson, & T. A. Williams (Eds.), Technology in mental health care delivery systems, pp. 119--137, Norwood, NJ: Ablex (1980)
10. Tong, E. M. W., Bishop, G. D., Enkelmann, H. C., Why, P. Y., Diong, M. S.: Emotion and appraisal: A study using ecological momentary assessment, Cognition and Emotion, vol. 21, pp. 1361--1381. (2007)

11. Carstensten, L., Pasupathi, M., Mayr, U., Nesselroade, J. R.: Emotional experience in everyday life across the adult life span, J. Pers. and Soc. Psychology, pp. 644--655. (2000)
12. Elfenbein, H.A., Der Foo, M., Boldry, J., Tan, H.H.: Dyadic effects in nonverbal communication: A variance partitioning analysis, Cognition and Emotion, vol. 20, pp. 149—159. (2006)