A Comparison of Human and Machine Learning-based Accuracy for Valence Classification of Subjects in Video Fragments

Y.H. Holkamp[1], J.G.M. Schavemaker[2]

[1]Faculty of EEMCS, Delft University of Technology, Delft, The Netherlands.
[2]TNO, Delft, The Netherlands. John.Schavemaker@tno.nl

Introduction

Facial expressions are the primary way to show one's emotional state. Automatic recognition of these cues from video using software allows for various improvements in human-computer interaction, ranging from improved feedback for recommender systems to automatic labeling of movies according to the emotions they induce. A number of affective display databases have been created to aid development in this field. These datasets are frequently available for academic use [1, 2, 3], use picture or video stimuli and range from highly controlled [1, 2] to more natural settings [3]. We observe that methods using these datasets report accuracy figures that leave room for improvement [5].

The contribution of our research is a comparison of the accuracy of valence estimation using facial expressions between human annotators and machine-learning classification when using an open affective computing dataset that uses video to induce emotion. This comparison helps to determine to what extent existing methods can be improved or whether optimal accuracy has been reached for facial expression-based methods. To the best of our knowledge, few comparisons have been made on datasets where video stimuli were used.

In order to make this comparison, we have reproduced the method described by Koelstra and Patras [5] for the case of using facial expressions to estimate the valence experienced by the subject. In contrast to the original approach, we used the Noldus FaceReader software [8] to detect and quantify facial activity using FACS [6] Action Units (AU). Additionally, we have conducted an experiment with 15 users and found that human observers exhibit high inter-rater agreement. Yet, we found that certain video fragments obtained from [1] are difficult to classify, even for human observers. The structure of this paper is as follows; first we describe the methods used. Finally we present our findings and discuss these results and conclude this work.

Method

In our experiments we used the MAHNOB-HCI data set [1], which was created for affect recognition and implicit tagging applications and contains face video recordings, EEG data and more. From the affect recognition part of this dataset, we selected the 24 subjects for whom the full set of data was collected without error. We cut the recordings to match the stimulation part of the experiments. To allow for a comparison between our work and [5], we mapped the subjects' self-assessment ratings to two categories; negative valence (rating 1-5) and positive (rating 6-9). Using this data, we conducted two experiments.

Annotation experiment. In the first experiment, we selected two fragments for random stimuli for each subject in the dataset, resulting in a total of 48 videos. We developed a web-based video annotation system where users were asked to view these videos after providing us with their age, gender and nationality. We presented the users with an explanation of their task, after which each user watched the same four videos. After these items were rated, the remaining videos were presented in random order. The former allowed us to use these sessions as 'gold standard' data to possibly explain outliers whereas the latter ensures that the order of the sessions does not influence the ratings.

Once the user watched a video, she was asked to estimate the emotion of the person they saw in the video using a 5-point Likert-scale, ranging 'negative' to 'positive'. We explicitly instructed the subjects to choose 'neutral' if

Proceedings of Measuring Behavior 2014, (Wageningen, The Netherlands, August 27-29, 2014).
Editors: A.J. Spink, L.W.S. Loijens, M. Woloszynowska-Fraser & L.P.J.J. Noldus. www.measuringbehavior.org

they could not distinguish an emotion. Furthermore the user was asked to rate how visible the emotion of the subject was in their facial expression using a 5-point Likert scale, ranging from 'not visible' to 'very explicit'.

For our experiments we gathered a group of 4 female and 11 male participants in the age groups 20-29 (n=13) and 30-39 (n=2). They have a Swedish (n=1), French (n=1) or Dutch (n=13) nationality and were asked to anonymously complete our annotation task from their home or office computer as courtesy or for research purposes, no actual rewards were provided. To speed up the annotation process, videos were played at four times the normal speed, which was reported in [10, 11] to have limited impact on detection accuracy. This resulted in an average length of 19.4 seconds (σ=4.9). To keep our subjects focused, we presented the raters with a page that encouraged taking a break at three intervals during the experiment but this was not enforced.

To allow for comparison between the valence estimation performance of humans and software-based estimators, we reduced the responses of our raters to two categories, positive and negative valence, where neutral ratings were included in the negative valence category, as was done in [5]. We then compared the ratings provided by our subjects against the self-assessment ratings provided in the MAHNOB-HCI datasets. In addition to the average human accuracy we also take the majority vote into consideration.

Machine-learning classification experiment. In addition to human estimators, we partially reproduced the method described in [5], which uses facial expressions to estimate the valence. In contrast to their system, we used a support vector machine rather than a Gaussian Naive Bayes classifier and we used the off-the-shelf Noldus FaceReader software to provide us with information regarding Action Unit (AU) activation levels.

The method works as follows; the video fragments were automatically annotated using the FaceReader software, resulting in activation levels at different time intervals. For each entry in the list of EMFACS AU combinations associated with emotion [5], we determined the number of onsets, offsets and the difference between onset and offset strength. Using this setup, we detected 29 combinations, resulting in a total of 29 x 3 = 87 features.

As in [5], we applied recursive feature elimination (RFE) as implemented in [9] to select the features used for classification by training a linear SVM and removing the 10% lowest-weighted features until we reach a predefined number. This number was found using a 10-fold cross-validation on our training set of 6 subjects. In order to determine the classification performance, we then performed a leave-one-fragment-out cross-validation on a per-subject basis, resulting in a SVM trained using 19 fragments from one subject and tested using the 20th.

Results

Overall we obtained the statistics shown in Table 1. Our implementation of the method by Koelstra and Patras obtained an accuracy within 4% of the figures reported in the original work. We observe that the human annotators obtained an accuracy of 71.6%, higher than the 66.2% for the best performing machine-learning approach.

Method	Accuracy
Koelstra & Patras [5][1]	64.0%
Our implementation[1]	66.2%
Majority class[1]	62.6%
Human, average[2]	71.6%
Human, majority vote[2]	75.0%
Majority class[2]	66.6%

Table 1. Valence classification accuracy for two different methods. Majority class performance is included to provide a baseline. Differences are not significant according to pairwise t-tests. [1]Using full dataset (N=480). [2]Using subset (N=48).

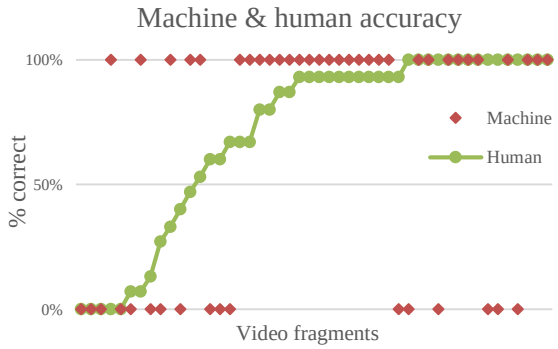

Figure 1. Machine versus human annotator accuracy, the video fragments are presented in ascending order of human accuracy. Note that the machine classifier only rates once and can thus only obtain an accuracy of 0 or 100%.

The differences in ratings between human annotators and the machine learning approach can be seen in Figure 1. We can see that, on the left side of the graph, there is a series of video fragments for which both human and computer fail to provide accurate classification. Conversely, there are nine additional video fragments for which the machine approach fails to provide accurate ratings.

We applied Fleiss' kappa to determine the inter-rater reliability of our human subjects, resulting in an agreement degree of 0.66. According to the index presented by Landis and Koch [7], this corresponds to 'substantial agreement'. We found no significant differences in performance between the different age groups, genders or nationalities among raters.

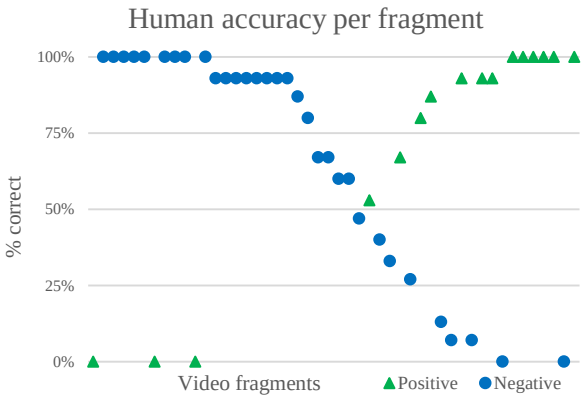

Figure 2. Percentage of correct votes per video fragment. The video fragments were ordered ascending by the number of positive valence ratings. The color and shape show the self-assessment rating provided by the subject in the video.

In Figure 2, we can see the accuracy of the human annotators when comparing their ratings against the 'ground truth', in the form of the self-assessment by the subjects of the dataset. Three fragments were incorrectly classified as negative valence, whereas nine video fragments were misclassified as positive valence.

Discussion

We will now briefly discuss our results and highlight some areas for future work. First we have seen that our implementation of the method by [5] obtained a slightly higher performance than the original. Furthermore, Table 1 showed that the performance obtained by our human annotators was higher than that obtained by the automatic methods. We might even consider our automatic classifier advantaged as it was trained using other behavior shown by the same subject, as was also the case in [5]. Conversely it can be argued that this small amount of knowledge about the subject may not level the playing field when compared with human annotators with years of experience. Preliminary results indicate that the accuracy does not degrade when the automatic classifier is only trained using information from other subjects, suggesting that the knowledge of one specific subject might not be as advantageous as might be expected.

In our human annotation experiment, we found that the agreement among our raters is high with only 10 out of 48 video fragments receiving less than 75% of the ratings in one specific category. In Figure 2, we saw that three video fragments showing a positive valence were misclassified as negative valence. This effect might be caused by the subjects in these sessions showing little emotion, which makes it difficult for the observer to determine the experienced valence. We also found a significant fraction of negative valence video fragments which were misclassified by human observers. Similarly, in Figure 1, we see that many of the fragments misclassified by human observers were also mislabeled by the machine learning classifier.

Investigation of the misclassified sessions showed that several subjects show either little emotion or show expressions that might be difficult to relate to their self-assessment ratings. An example of the latter is a smiling response when confronted with a sad video. This difference between the observed emotions and the emotions that were reported by the subjects might be one important factor in explaining our results. The cause of the discrepancy could be because they actually felt the reported emotion (but did not clearly show this) or for example because they have given socially desirable answers.

In other comparisons between human and automatic annotations, the used datasets also differ. For instance, in [4], the subjects were asked to recall and describe a situation in which they experienced a specific emotion. There might be a difference in the expressiveness of the facial expressions depending on the type of emotion elicitation. Finally, where many similar studies provided the subjects with monetary rewards, we were not able to do so, which might have influenced the performance of our subjects.

Conclusion

In this paper we present the results of a comparison between classification accuracy of humans and machine-learning classifiers. For this we used the MAHNOB-HCI affective computing dataset and we have reproduced and extended the facial expression-based method by Koelstra and Patras. Our results show that both humans and machine classifiers agree to a large portion on the appropriate class for video fragments. In our experiments, we found that human annotators obtained higher accuracy than the automatic classification methods.

Future work could include evaluating how well the automatic classification method performs when training using the human labels, as described in [12], allowing for an estimation as to how close the performance of the automatic classifier is to the human annotators, rather than to the self-assessment ratings. Evaluating multiple datasets using the same human annotation method will also allow for comparisons across datasets, which could help compare automatic classification methods developed and tested using different datasets and allows datasets to be compared based on an human difficulty level of classification.

References

1. Soleymani, M., Lichtenauer, J., Pun, T., & Pantic, M. (2012). A multimodal database for affect recognition and implicit tagging. *Affective Computing, IEEE Trans. on*, ***3***(1), 42-55.

2. Koelstra, S., Muhl, C., Soleymani, M., Lee, J. S., Yazdani, A., Ebrahimi, T. ... & Patras, I. (2012). DEAP: A Database for Emotion Analysis; Using Physiological Signals. *Affective Comp., IEEE Trans. on*, ***3***(1), 18-31.

3. Sneddon, I., McRorie, M., McKeown, G., & Hanratty, J. (2012). The Belfast Induced Natural Emotion Database. *Affective Computing, IEEE Trans. on*, ***3***(1), 32-41.

4. Janssen, J.H., et al. Machines Outperform Laypersons in Recognizing Emotions Elicited by Autobiographical Recollection. *Human–Computer Interaction* **28.6** (2013): 479-517.

5. Koelstra, S., & Patras, I. (2013). Fusion of facial expressions and EEG for implicit affective tagging. *Image and Vision Computing*, ***31***(2), 164-174.

6. Ekman, P., & Friesen, W. V. (1977). Facial Action Coding System.

7. Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, ***33***, 159–174.

8. Den Uyl, M. J., & Van Kuilenburg, H. (2005, August). The FaceReader: Online facial expression recognition. In *Proceedings of Measuring Behavior*, ***30.***

9. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O. ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *The Jrnl of Mch. Learning Resrch*, ***12***, 2825-2830.

10. Bould, E., Morris, N., & Wink, B. (2008). Recognising subtle emotional expressions: The role of facial movements. *Cognition and Emotion*, ***22***(8), 1569-1587.

11. Kamachi, M., Bruce, V., Mukaida, S., Gyoba, J., Yoshikawa, S., & Akamatsu, S. (2001). Dynamic properties influence the perception of facial expressions. *Perception*, ***30***, 875-887.

12. Steidl, S., M. Levit, A. Batliner, E. Nöth, and H. Niemann (2005). Of All Things the Measure Is Man: Automatic Classification of Emotions and Inter-Labeler Consistency. In ICASSP (1), pp. 317-320.