When Human Coders (and Machines) Disagree on the Meaning of Facial Affect in Spontaneous Videos

Mohammed E. Hoque, Rana el. Kaliouby, Rosalind W. Picard

Media Laboratory, Massachusetts Intitute of Technology
20 Ames Street, Cambridge, MA 02142
{mehoque, kaliouby, picard}@media.mit.edu

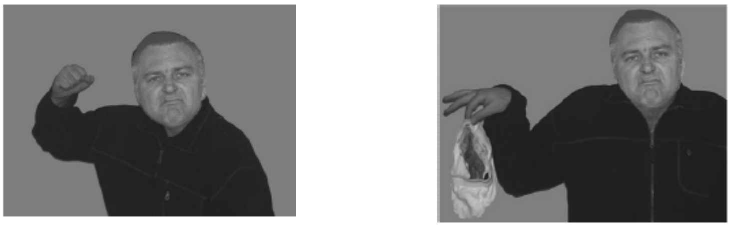
Abstract. This paper describes the challenges of getting ground truth affective labels for spontaneous video, and presents implications for systems such as virtual agents that have automated facial analysis capabilities. We first present a dataset from an intelligent tutoring application and describe the most prevalent approach to labeling such data. We then present an alternative labeling approach, which closely models how the majority of automated facial analysis systems are designed. We show that while participants, peers and trained judges report high inter-rater agreement on expressions of delight, confusion, flow, frustration, boredom, surprise, and neutral when shown the entire 30 minutes of video for each participant, inter-rater agreement drops below chance when human coders are asked to watch and label short 8 second clips for the same set of labels. We also perform discriminative analysis for facial action units for each affective state represented in the clips. The results emphasize that human coders heavily rely on factors such as familiarity of the person and context of the interaction to correctly infer a person's affective state; without this information, the reliability of humans as well as machines attributing affective labels to spontaneous facial-head movements drops significantly.

Keywords: facial expression analysis, affective computing, action units, spontaneous video

1 Introduction

One important application area for intelligent virtual agents is intelligent tutoring systems (ITS). These systems are much more effective in maximizing learning goals when the virtual agent is equipped with the ability to understand the learner's facial affect. Building an automated system that recognizes spontaneous facial expressions remains a challenging endeavor: human facial muscle activations can occur in over twenty thousand combinations and there is no codebook to describe the mapping from facial expressions to affective state. Despite the immense diversity in human facial expressions, impressive results have been reported in the literature in recognizing small sets of prototypic facial expressions. However, when examined carefully, the reported results are dependent on the dataset, environment, and simplicity of affective state categorizations. For example, a study may be based on a dataset of professional actors/random participants feigning particular affective states in a laboratory environment. While pattern recognition algorithms can be made to perform really well on such data, those algorithms do not generalize as well to spontaneous natural emotions. A more accurate validation of automated facial analysis systems uses emotional clips from movies/TV shows that have obvious emotional tags, optimized camera views and lighting conditions. These clips are carefully selected, clipped and validated by human judges and do not contain the variability of difficult and imperfect natural data. Another approach employs carefully selected movie clips which are believed to elicit natural spontaneous emotions from humans. A small group of participants are asked to view those clips while agreeing to be video taped. One limitation of such dataset is that it does not provide a task dependent environment where context becomes an inevitable part of elicited affective states. After building a framework to analyze and recognize affective states, most researchers also use a relatively limited set of individuals and a small set of possible labels to validate their framework. When individuals are given more labeling choices (e.g. both “proud” and “happy” for a smiling face) then agreement in labeling tends to go down.

For automated classification of facial expression data, given the lack of a robust theory that maps expressions to labels; one needs to have some portion of the data labeled by human judges. The judges are often instructed to equate a specific set of facial Action Units (AUs) with a particular affective state [3]. For example, the nose wrinkler (AU 9) is often considered a distinguishing feature of disgust. A brow lowerer (AU 4) is a common feature of confusion. Combination of lip AUs such as jaw drop (AU 26), low intensity lip corner puller (AU 12), lips funneler (AU 22), and lips part (AU 25) are often regarded as signatures of happiness. This direct mapping between AUs and affective states may work well in synthetic data, resulting in high agreement among the human judges, but it may not converge well in real life data. For example, in Figure 1 [1], the same face with identical facial expression is used in two different contexts, resulting in completely different meanings. While humans are more likely to tag Figure 1 (a) as angry and Figure 1 (b) as disgust, an automated algorithm would not be influenced by the context and would not differentiate between the two images. This is an example of how our perception of something can be biased by context and prior experience, even when the input pattern is constant.



(a) (b)

Figure 1. The prototypic facial expression of disgust is placed in two different contexts, where the majority of the participants label (a) as anger, and (b) as disgust (Figure used by permission, copied from Aviezer et al. [1]).

The remaining part of the paper is divided into the following sections – Section 2 presents details regarding experimental setup, data collection and labeling. Section 3 provides the results of how the inter-rater reliability among coders drops when context information is removed. It also shows details of correlation between a set of AUs and affective states, based on manual labeling. Section 4 summarizes the lesson learned through this study which may prove to be useful towards customizing an automated AU detection framework to work with natural data.

2 Data Collection Methods

In this paper, we have used data collected from an experiment where humans interact with a computer agent – Autotutor [2]. Autotutor is designed to simulate a human tutor while having the ability to interact with the learner using natural language. Autotutor engages students in learning by asking them questions on a given topic and then providing them useful clues to get to the correct or complete answer.

2.1 Materials, Data size and Procedure

The participants consisted of 28 undergraduate at the University of Memphis who participated for extra course credit. In this paper, 10 sessions consisting of 10 participants were randomly selected. Each session was about 30 minutes long.

2.2. Tagging of Affective States

The affective states that were considered in this experiment were boredom, confusion, flow, delight, frustration, surprise and neutral (surprise and flow were not included in the manual discriminative AU analysis, described in Section 4.1, due to the infrequency of occurrences). These categories were ones that were frequently experienced in a previous experiment with Autotutor. Boredom was defined as lack of interest, whereas confusion was defined as lack of understanding. Flow symbolized a state that was a mix of drive, alertness, interest, concentration, and being self-determined that results from pleasurable engagement in an activity that is challenging but not too challenging. Delight indicated a high degree of satisfaction from accomplishment or comprehension. Frustration was described as a state of disappointment or annoyance with the Autotutor interaction. Surprise was labeled as an amazement being triggered from something unexpected. Neutral was equivalent to having no apparent affect or feeling.

There were four levels of tagging that took place in order to collect ground-truth data. In the first phase, learners (Coder 1) watched their own video of interacting with Autotutor, and then, were asked to label the affective states that they had experienced during the interaction; this was termed *self-judgment.* The second phase of the tagging included each participant returning to the lab after a week and then tagging a video of another participant interacting with Autotutor; this was termed as *peer judgment* or Coder 2. In the third level of tagging, two trained judges (Coders 3 and 4) with experience of facial expression tagging were asked to tag all the videos for a particular set of affective states. Both judges were undergraduate research assistants with extensive training on tutorial dialogue characteristics and FACS coding. Inter-rater reliability was measured using Kohen's kappa between self vs. peer (Coder 1 vs. Coder 2), self vs. judge1 (Coder 1 vs. Coder 3), self vs. judge2 (Coder 1 vs. Coder 4), peer vs. judge1 (Coder 2 vs. Coder 3), peer vs. judge2 (Coder 2 vs. Coder 4), judge1 vs. judge2 (Coder 3 vs. Coder 4). Among all these pairs, judge1 and judge2 had the highest agreement (kappa =0.71). We took the subset of videos where these two judges perfectly agreed and used these videos with the judges labels as the ground-truth data.

Next, we extracted 8-second clips around those points that the trained judges had identified. Then, those clips for all participants were presented in random order to three independent coders. The coders were expected to watch the 8 seconds of clips and assign one of the affective states label to each clip. The main rationale behind producing small segments of videos around each point that the trained judges labeled, was to produce a set of training and test examples that one would use to train a classifier to recognize affective states. However, the trained judges had the opportunity to view the entire video to tag affective states versus a machine counterpart that is typically trained on smaller video segments in random order. Therefore, we felt that it is more appropriate to analyze the agreement among humans on smaller segments of videos to get an idea of how difficult it may be for humans to label affective states without context.

3 Results

Coders 5, 6, and 7 were given the ground truth video clips where the expert judges agreed 100% of the time. Therefore, it was expected that coders 5, 6 and 7 would agree with the ground truth labels more often than "chance" to the least. Chance was calculated as 1-B.

Using Bayes Error, $B= \sum_{i} P_i (1 - P_i)$, where

P_i = i-th class prior probability based on the frequency of seven different labels in the training set. Based on the frequency of labels in the given dataset, chance was 51%. If the distribution of the labels in the training data set was uniform, then chance would have been 1/7 = 14.28% for the 7 classes.

Table 1 demonstrates the kappa and percentage agreement between ground truth labels and Coder 5, 6 and 7. Both kappa and percentage agreement between ground truth values and independent coders were lower than chance which was 51%. Results indicated high agreement (above 80%) on delight, while disagreeing significantly on other categories. The lowest percent agreement was for frustration and confusion while the highest was for delight and surprise.

Table 1. Kappa and percentage agreement among the ground truth labels and Coders 5, 6, 7. Ground truth corresponds to the labels agreed upon by Coders 3 and 4.

Combinations	Kappa	% agreement
Ground truth vs. Coder 5	0.25	0.38
Ground truth vs. Coder 6	0.38	0.50
Ground truth vs. Coder 7	0.28	0.39
Coder 5 vs. Coder 6	0.35	0.45
Coder 5 vs. Coder 7	0.31	0.41
Coder 6 vs. Coder 7	0.46	0.54

4.1. Analysis of Discriminative Power of AUs

Most automated facial expression analysis systems assume a one-to-one mapping between facial expressions and affect. However, it has been shown that when affective states beyond the six basic emotions [4] are considered, or when non-prototypic expressions of an affective state are included, the discriminative power of action units drops. In other words, the relationship between a single AU or facial expression and an affective state is not direct, and the same AU can appear in more than one affective expression. In this study, we trained judges manually coded a randomly chosen 20% of the original data for AUs. The main goal was to distinguish a smaller subset of AUs which may be unique to a particular affective state.

After the manual recognition of AUs, an analysis was done to predict how good of a discriminator a particular AU is, given a mental state. We define a heuristic variable, $H = P(Yj\ |Xi) - P(Yj\ |\sim Xi)$, where Y = AUs, and X = mental states. The magnitude of H quantifies the discriminative power of a display for a mental state; the sign depicts whether an action unit increases or decreases the probability of a mental state. To explain how the heuristic works, consider the following hypothetical cases of the discriminative ability of a lip corner pull in identifying delight:

Assume that a lip corner pull (AU 12) is always present in delight $P(Yj{=}12|Xi{=}\text{delight}) = 1$, but never appears in any of the other mental states $P(Yj{=}12\ |Xi \neq \text{delight}) = 0$. The heuristic is at its maximum value of one, and its sign is positive. The presence of a lip corner pull is a perfect discriminator of delight. Similarly, suppose that a lip corner pull (AU 12) never shows up in agreeing, $P(Yj{=}12\ |Xi{=}\text{agreeing}) = 0$, but always shows up in all other mental states $P(Yj{=}12\ |Xi \neq \text{agreeing}) = 1$. The magnitude of H would still be one, but its sign would be negative. In other words, the lip corner pull would be a perfect discriminator of agreeing, even if it never occurred in that state. (Again, this example is not actually true). Finally, if a lip corner pull is always observed in delight, and is also always observed in all other mental states, then $P(Yj{=}12|Xi{=}\text{delight}) = P(Yj{=}12\ |Xi \neq \text{delight}) = 1$. In this case, H has a value of 0, and the lip corner pull is an irrelevant feature in the classification of delight.

The result of computing H for each mental state is shown in Table 2. Table 2 helps identify a particular set of AUs which are either significant discriminators or non-discriminators of an affective state. Table 2 provides such a list, where outer brow-raiser (AU 2), mouth stretch (AU 27), eyes closed (AU 43), head turn left (AU 51) etc were positive discriminators of boredom. Lip corner pull (AU 12), lid lightener (AU

7), brow lowerer (AU 4), Jaw drop (AU 26), inner brow raiser (AU 1) were negative discriminators of boredom. Lip corner pull (AU 12) turned out to be the best discriminator for both delight and frustration and least for boredom, confusion and neutral. The highest discriminatory value of all the AU's went to AU 12 for delight (note that coders were able to identify delight more often than other categories). It was also evident that several AUs were positively correlated with more than one affective state.

Table 2. Discriminatory and non-discriminatory AUs per mental state. The AUs are listed in order of their contribution (most significant to least).

Mental states	**Discriminatory Aus**	**Negatively discriminatory Aus**
Boredom	2, 27, 43, 51	12, 7, 4, 26, 1
Confusion	4, 7, 17, 52	12, 53
Delight	12, 25, 26, 7	43
Frustration	43, 12, 7	57, 25, 54,
Neutral	None	7, 12, 4, 25, 43

4 Discussions and future work

In this paper, we provide a methodical approach to facial expression analysis when dealing with challenging natural data obtained from interaction with an automated agent in a learning context. Over 300 minutes of video data were collected from experiments where a human interacted with an animated agent, where the human played the role of learner and the agent played the role of tutor. The data were manually coded for seven different mental states of boredom, confusion, flow, delight, frustration, neutral and surprise by two human judges with inter-rater reliability being 0.7. These ground truth videos were then segmented into 8-second clips and given to 3 independent coders for tagging. The percent agreement among the independent coders was less than chance. This finding is very important because in pattern recognition, classifiers are typically trained on similarly short video clips and in most cases, the classifiers do not perform well with natural data.

Developing a system that reliably recognizes over 31 different AUs is a difficult problem. In this study, we have manually coded a random 20% of our data to detect the most discriminative and least discriminative AUs for the relevant affective states. Due to the experimental set up of our study, participants had to sit very close to the camera. Therefore, even a slightest movement/tilt of the head as part of natural movement would trigger most of the Action Descriptors (AD) related to head movement (AD 51 to AD 58 and AD 71 to AD 76). Therefore, it is probably not useful trying to incorporate those AUs in our analysis. Based on observation and manual coding of the data, AUs related to lip movement (AU 12, AU 15, AU 18, AU 25), eye/lid/brow movement (AU 1, AU 2, AU 7) were more relevant. From this experience, given the task, camera position, and context, it may be possible to group a bunch of AUs based on relevance and importance.

In the past, there has been a trend to associate a set of AUs with particular affective states regardless of the task and context. However, even faces made of AU's that correspond to basic emotions can take on a label of a different basic emotion if the context is modified [1]. In this paper, we argue that a blind association between a set of AUs and a particular affective state could potentially confuse the automated classifier. Instead of looking for one-to-one or many-to-one association between AUs and affective states, it is important to investigate the interplay among AUs in sequence for a given affective state. Even though a video clip may contain AU signatures not unique to one affective state, the sequence in which the AUs appear and interact with each other may reveal unique patterns.

In the 8-second video clips, it was often the case that participants moved away from the viewable range of the camera, looked to the side with a tilted face, and occluded their face with their hands. While it is not possible to fully address these concerns with the current state of computer vision algorithms, it may be possible to add meaning to these phenomena, given a context.

Our immediate future work involves incorporating all these lessons learned towards modifying our existing AU detection framework for the given dataset. While it is desirable to develop a one-size-fits-all affect recognition system by inferring meaning from the detected AUs, it is very likely that we may have to perform a fair bit of customization of the framework and manual work depending on the dataset. Future work will also involve conducting similar exploratory studies of different real-world datasets to statistically model the way affect is communicated through natural data. It is also a possibility to incorporate and fuse other features such as shoulder movement, hand gestures, task-related behaviors, and prosodic aspects of speech towards affect recognition for animated agents.

5. Acknowledgment

This research was partially supported by NSF (IIS 0325428). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding institution. The authors would like to acknowledge their colleagues at the University of Memphis for sharing the Autotutor data.

References

1. H. Aviezer, R. Hassin, J. Ryan, G. Grady, J. Susskind, A. Anderson, M. Moscovitch and S. Bentin, Angry, Disgusted or Afraid? Studies on the Malleability of Emotion Perception. *Psychological Science*, vol. 19, 2008, pp. 724-732.
2. A. C. Graesser, B. McDaniel, P. Chipman, A. Witherspoon, S. D’Mello, and B. Gholson, Detection of Emotions During Learning with AutoTutor. *Proceedings of the 28th Annual Meeting of the Cognitive Science Society*, Vancouver, Canada, 2006, pp. 285-290.
3. P. Ekman and W. Friesen, Facial Action Coding System: A Technique for the Measurement of Facial Movement. Consulting Psychologists Press, Palo Alto, 1978.
4. P. Ekman, W. V. Friesen, and P. Ellsworth, *Emotion in the Human Face: Guidelines for Research and an Integration of Findings*. New York: Pergamon Press, 1972.