Human Rating of Emotional Expressions
Scales vs. Preferences

Marco Pasch[1], Andrea Kleinsmith[2] and Monica Landoni[1]
[1]*Faculty of Informatics, University of Lugano, via Buffi 13, 6900 Lugano, Switzerland*
[2]*Department of Computer and Information Science and Engineering, University of Florida, FL 32611, Gainesville, U.S.A.*

Keywords: Perception, Emotions, Affect, Preference Rating, Scale Rating, Ground Truth Labeling.

Abstract: Human ratings of emotional expressions are the foundation for building and training automatic affect recognition systems. We compare two rating schemes for labeling emotional expressions: likert scales and pair-wise preferences. A statistical analysis shows that while there is a strong correlation between the two schemes, there are also frequent mismatches. Our findings indicate that the schemes perform differently well per affect label. We discuss reasons for this and outline planned future work based on the findings.

1 INTRODUCTION

When building automatic affect recognition systems based on bodily manifestations of affect (such as facial expressions, posture, behavior, physiology) one typically has to go through the following steps. First, collecting data that contains such bodily manifestations. Then, establishing which particular affective states can be observed in the data when it is considered that an inherently "correct" affective label does not exist, a process known as ground truth labeling. With these labels one would search for the features within the collected data that are key to the respective affective states. These then form the basis for building a system that can automatically identify affective states.

Establishing the ground truth is thus a key component in the development process. Human perception of emotions is often the benchmark that a system is tested against; it is an important step for the creation of affect recognition systems. Labels are often obtained in two ways: self-report from the person who portrayed an emotional expression or recruiting observers to rate images or sequences of affective expressions and assign a label to each sequence.

Our aim is to find better rating schemes for obtaining ground truth labels. For this, we investigate alternative methods for rating stimuli containing emotional expressions. At the very least, this allows us to assess the influences of particular schemes on the resulting affect labels. Ideally, we can make recommendations which labeling scheme works best given the particular goals of a study or system to be developed. In this initial study, we compare two rating schemes for the labeling of a corpus of affective body postures: pairwise preference rating and rating on multiple scales. The hypothesis that the rating schemes are tested against is:

H_1. *There is an inconsistency between reported preferences and reported scale ratings.*

The remainder of this paper is organized as follows: Section 2 examines typical approaches for establishing the ground truth. Section 3 describes the method implemented in our study to compare ground truth labeling approaches. The results are reported in Section 4. A discussion and conclusions are presented in Sections 5 and 6, respectively.

2 GROUND TRUTH LABELING

Emotional expressions are usually elicited in two ways: letting actors portray emotions that are not actually felt (acted) or recording the affective behavior of people in various scenarios and who were not instructed to portray particular behavior (non-acted). Acted expressions are often exaggerated portrayals of emotional states, whereas non-acted expressions are generally subtler and more complex. Not surprisingly, until recently automatic affect recognition systems were based on acted expressions, as these are easier to detect (Zeng et al., 2009). Only recently there has been a shift towards subtler naturally occurring expressions.

Pasch M., Kleinsmith A. and Landoni M..
Human Rating of Emotional Expressions - Scales vs. Preferences.
DOI: 10.5220/0004727602400245
In *Proceedings of the International Conference on Physiological Computing Systems* (PhyCS-2014), pages 240-245
ISBN: 978-989-758-006-2

However, self-report may not be feasible and is often considered unreliable (Afzal and Robinson, 2009; Kapoor et al., 2007). Yet, in particular since the shift towards non-acted data, observer ratings are becoming more common for establishing the ground truth (Kleinsmith and Bianchi-Berthouze, 2013). This method may be particularly relevant when the aim of the application into which the recognition software will be integrated is to act as a human interaction partner.

When establishing the ground truth using observers, what labeling model should be used? Two options are pairwise preference ratings and rating on multiple scales. Many studies in the Affective Computing field employ a forced-choice design, e.g., (Savva and Bianchi-Berthouze, 2012) and (Kleinsmith et al., 2011). In this design, observers are presented with a list of choices and are forced to choose from that list. An advantage of a forced-choice design over a free-form design is that it forces an absolute match and eliminates the possibility of observers providing non-emotion labels, a known issue with free-form designs (Russell, 1994). However, forcing an absolute match is also a disadvantage, as the list may not include all options considered applicable by the observers. Similarly, concurrence of more than one distinct emotional state can not be captured in a forced choice design. Also, the intensity of a perceived emotion is lost, which may be particularly problematic when dealing with subtle emotional expressions.

Pairwise preference rating is used in artificial intelligence (Fürnkranz and Hüllermeier, 2005; Doyle, 2004) and machine learning (Yannakakis, 2009) fields and may be considered an attempt to overcome the limitation of a forced choice design. In pairwise preference rating, observers are presented with pairs of stimuli and asked to choose which stimulus best represents a particular label. This process is repeated for all possible stimulus pairs. Because it does not require the observers to determine an absolute match, pairwise preference rating may help to reduce the variability that exists between observers when a forced choice design is used (Kleinsmith and Bianchi-Berthouze, 2013).

Obtaining ratings for the same stimulus on multiple scales representing discrete emotion labels has also been used in other emotion recognition research (Liscombe et al., 2003). Observers were asked to rate speech tokens on separate scales for 10 discrete emotions. The authors conclude that by rating stimuli on multiple emotion scales shows that their stimuli expressed several different emotions at the same time, resulting in a better, more complete representation of emotion.

3 METHOD

3.1 Participants

For this initial study, participants were recruited via mailing lists from within the university community. They were aware of the aims of the study and no compensation was given. 12 participants took part in the study.

3.2 Stimuli

As stimuli we use images from the *UCLIC Database of Affective Postures and Body Movements* (Kleinsmith et al., 2006; Kleinsmith et al., 2011); these are readily available online. The database consists of separate corpora of acted and non-acted whole body postures, of which we use the latter. The non-acted collection consists of 105 postures that were obtained from people playing physically active video games and have been rated for four affective labels: *concentrating*, *defeated*, *frustrated*, and *triumphant*. The postures are modeled on an abstract avatar seen from a frontal perspective in front of neutral grey background.

In order to prevent study fatigue and to account for the fact that the study was conducted online (as was the original labeling for the database used), the number of ratings the participants were required to make was kept to a level sufficient for carrying out the study.

For this reason we only use a subset of 10 postures from the UCLIC collection, chosen at random. Pairwise preference rating means that for n postures, $n * (n-1)/2$ pairs of postures have to be rated. For the entire corpus this would result in 5460 pairs. Bringing the number of stimuli down to 10 results in 45 pairs. Figure 1 shows one of the 45 screens participants see in the preferences condition. In a further attempt to keep the number of ratings at a manageable level, we let our participants rate a posture for all four labels in one screen in the scale condition as can be seen in Figure 2.

3.3 Procedure

Upon clicking on the link in the invitation email, participants reached the welcome screen. Here they received information about the study and that we ask them to rate postures on 55 screens. The order of rating conditions was assigned randomly.

In the preference condition, participants saw two postures next to each other. Below the posture images, they were asked: ”Which posture looks more

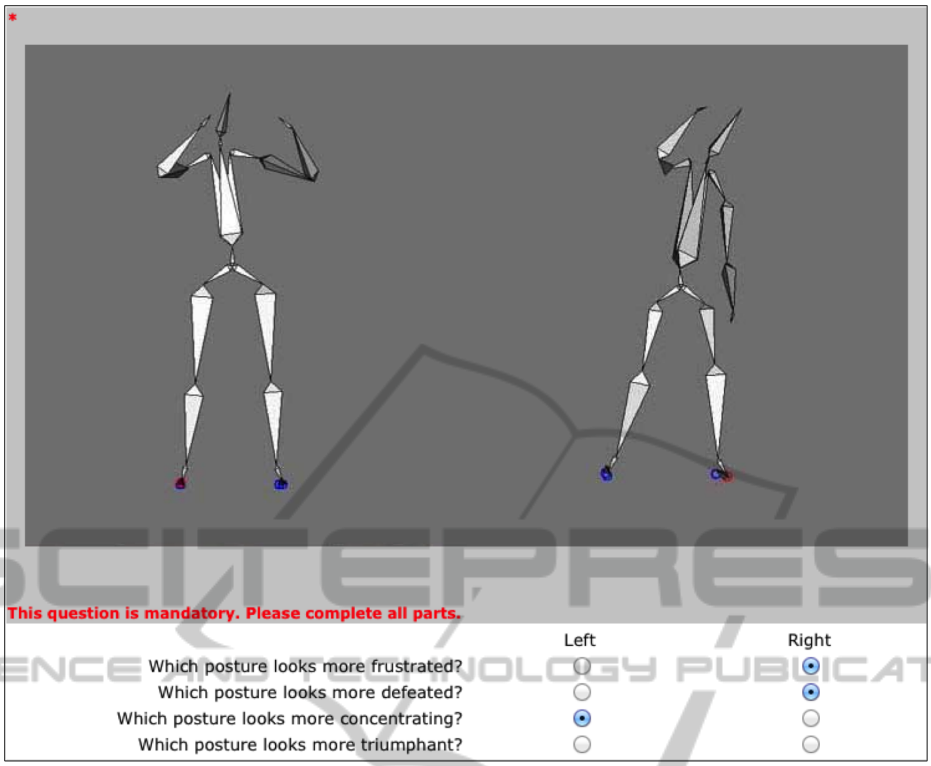

Figure 1: Screenshot Preference condition.

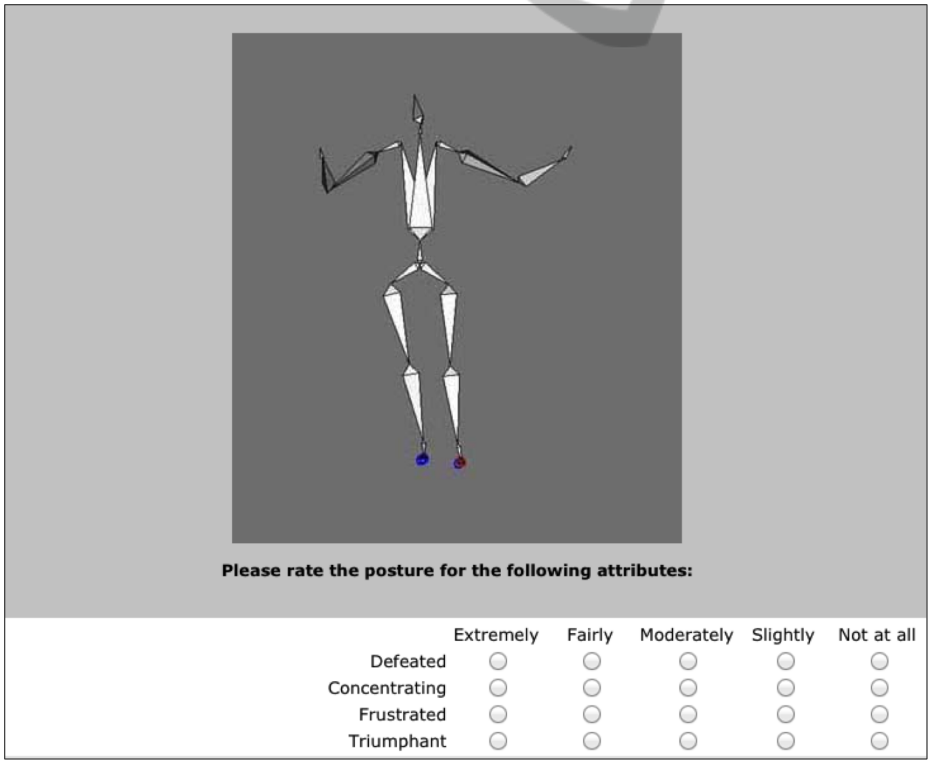

Figure 2: Screenshot Scales condition.

[blank] ?" four times; each time with the blank replaced by the four affective labels *concentrating, defeated, frustrated, triumphant*, respectively. With radio buttons they then indicated whether the posture on the left or on the right better corresponded to the label in question. Clicking a button labeled next brought them to the next screen. The order of the pairs was designed so that the same posture never appeared on two consecutive screens.

In the scales condition, participants saw a single posture and the instruction: "Please rate the posture for the following attributes:" Below, they saw four likert scales for the four affective labels, each with five options to rate the intensity of the particular attribute: *extremely, fairly, moderately, slightly*, and *not at all*.

After having rated the postures in both conditions, participants saw a debriefing screen, thanking them for their participation and giving them the opportunity to write any comments or thoughts they might have had in a text field.

4 RESULTS

The average time participants took to take part in the survey was 26.8 minutes, with a median of 23.5 minutes. We first check for order effects between the group of participants that first rated preferences and the group that first rated on scales. No significant differences can be found.

4.1 Comparison to existing Labels

The postures of the UCLIC Database of Affective Postures and Body Movements have already been labeled with the most frequent label that were assigned by raters when the data corpus was created. From the postures chosen for the study, 2 out of 10 postures are assigned different labels by our participants than labels that were assigned in the initial rating.

4.2 H_1 Test Statistic

To measure the degree of agreement between scale ratings and preference ratings we calculate the correlation coefficients between them. For this we follow the statistical analysis procedure for pairwise preference data introduced by (Yannakakis and Hallam, 2007). This procedure has been previously applied for the comparison of scale ratings and preference ratings for self-report data in (Yannakakis and Hallam, 2011). To make a comparison possible, pairwise preferences are inferred from scale ratings. These are then compared to the direct pairwise preferences.

Following (Yannakakis and Hallam, 2007), we obtain the correlation coefficients using

$$c(z) = \sum_{i=1}^{n} \frac{z_i}{N} \tag{1}$$

where N is the number of pairs i to correlate and $z_i = +1$, if scale ratings and preference ratings and $z_i = -1$ where there is no match. P-values of $c(z)$ are obtained from the binomial distribution.

We only take into account pairs where we can infer a clear preference from the scale ratings. If, e.g., a participant chose moderately for posture A and slightly for posture B for the same affective label, we can infer a preference for posture A in a pairwise comparison. This way, we do not assume a numerical basis of the scale. Also, we do not compare one participant's scale ratings to another.

As can be seen in Table 1, the direct preference ratings matched the preferences inferred from the scale ratings for 74% of the data samples (1595 out of 2160 possible matches). The number of incidents where we can infer a preference from scale ratings varies from 116 to 148 per participant, out of a possible total of 180 ratings (45 posture pairs x 4 affect labels). In total we have 1595 incidents where we can correlate between the preference rating condition and the scale rating condition. Correlation coefficients vary from .43 to .74 per participant and the total agreement correlation coefficient is .61. All correlation coefficients are highly statistically significant, ruling out the null hypothesis H_1.

4.3 Agreement Rates Across Affect Labels

Next, we take a closer look at the agreement rates between participants for each posture pair. We calculate the difference in agreement rates between direct preference ratings and inferred preference ratings for each posture pair where there is a clear preference inferred from scale ratings. We split the differences in agreement rates into three groups: scale rating has a lower agreement rate as preference rating, scale rating has the same agreement rate than preference rating, and scale rating has a higher agreement rate than preference rating.

Figure 3 shows the distribution of ratings for each affect label across the three groups. We can see that the labels frustrated and concentrated are rated more often with lower agreement rates in the scales condition than the preference condition. They are rated least frequently when there are higher agreement rates

Table 1: Number of matches between preference ratings and inferred scale ratings, number of incidents where there was a clear preference (out of 180 possible preferences per participant), and correlation coefficients between preferences and inferred scale ratings (all statistically significant at p<0.01).

	Matches	Incidents	$c(z)$
P1	113	145	0.56 *
P2	110	146	0.51 *
P3	116	133	0.74 *
P4	94	116	0.62 *
P5	103	128	0.61 *
P6	110	135	0.63 *
P7	108	123	0.76 *
P8	125	144	0.74 *
P9	97	123	0.58 *
P10	106	148	0.43 *
P11	90	119	0.51 *
P12	111	135	0.64 *
Sum	1283	1595	0.61 *

in the scales condition. The label defeated is to a lesser extent also rated more often with a lower agreement for scales than for preferences. The label triumphant on the other hand is only rated in a few instances with a higher agreement rate for scales than for preferences. In most instances, there is a higher agreement rate for preferences than for scales. Where the scale condition receives the same agreement rates as the preference condition, all affect label appear with similar frequency.

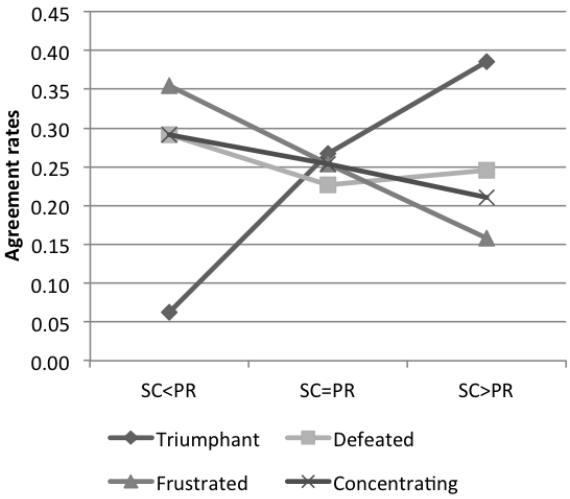

Figure 3: Distribution of ratings for an affect label across participants that happen with a lower agreement rate for scales than for preferences (SC<PR), an equally high agreement rate for scales and preferences (SC=PR), and higher agreement rate for scales and preferences (SC>PR).

5 DISCUSSION

Most significant is the finding that while direct preferences and preferences inferred form scale ratings are well correlated, there are frequent mismatches. The question remains, why is there a mismatch between scale ratings and preferences ratings from the same raters on the same set of stimuli? One possible answer is indicated by our analysis of differences between agreement rates for each affect label across the two rating schemes. This revealed differences in agreement rates. These are distributed unevenly, in that some affect labels are more often rated with a higher agreement rate in one scheme, and others are rated more often with higher agreement rates in the other scheme. It thus appears that the rating schemes may be differently well suited for various affect labels. With the limitation of our findings being based on only one corpus of affective data we must be careful to generalize this finding to other rating scenarios.

There are several other points worth noting. A study taking all of them into account is beyond the scope of the present paper, but our findings indicate that such a study may be worthwhile.

Number of ratings. As pointed out above, there were many more ratings in the preference scheme than in the scale scheme. Where n stimuli are to be rated, there have to be $n*(n-1)/2$ ratings. Our posture set is comparably small (n=10), still we already had 45 ratings in the preference scheme as opposed to 10 in the scale scheme. Rating a large number of abstract postures quickly becomes a monotonous task and in fact several participants reported that felt bored towards the end of the study. Another comment was that a participant felt increasingly insecure about previous ratings over the course of rating all stimuli.

Possible bias in experimental protocol. It is possible that our experimental protocol did favor preference rating, as there had to be more ratings made than for the scale ratings. One can argue that this leads to more familiarity. It can be also argued that pairwise preference rating is easier, because there are only two options from which to choose. The scale ratings consisted of 20 options.

Reliability of participants. In our present study there is no redundancy in the stimuli. This makes it impossible to assess the consistency and reliability with which participants rate the stimuli and whether there are differences for the rating conditions. A possible solution is given in (Kleinsmith et al., 2011), where participants had to rate the same set of postures multiple times. In addition, this would lead to increasing familiarity over time, thus addressing the possible bias in the experimental protocol.

6 CONCLUSIONS

We investigated differences between ratings of affective stimuli expressed through preferences and through scales. Our aim is to find appropriate schemes for rating affective stimuli in order to better define ground truth labels for the training and testing of automatic affect recognition systems.

Based on a corpus of abstract body postures, we find that while there is a strong correlation between preference and scale ratings, there are also frequent mismatches. We discuss reasons as indicated by our findings, as well as other potential causes. We plan to address these in future work. As we find that different rating conditions work better for different affect labels, we believe it is worthwhile to investigate rating schemes that make use of scales as well as preference ratings. Preference rating quickly requires large numbers of ratings to be made. We believe that data mining techniques can be helpful here. When a stimulus-label combination is rated with a high-consistency from a low number of raters it can be retired, leaving more capacity of further raters to more disputed cases. Ultimately, we envisage a protocol to assist assessors to achieve higher levels of consistency and agreement rates when rating affective stimuli for ground truth labeling.

REFERENCES

Afzal, S. and Robinson, P. (2009). Natural affect data: Collection and annotation in a learning context. In *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, pages 1–7.

Doyle, J. (2004). Prospects for preferences. *Computational Intelligence*, 20(2):111–136.

Fürnkranz, J. and Hüllermeier, E. (2005). Preference learning. *Künstliche Intelligenz*, 19(1):60–61.

Kapoor, A., Burleson, W., and Picard, R. W. (2007). Automatic prediction of frustration. *Int. J. Hum.-Comput. Stud.*, 65(8):724–736.

Kleinsmith, A. and Bianchi-Berthouze, N. (2013). Affective body expression perception and recognition: A survey. *Affective Computing, IEEE Transactions on*, 4(1):15–33.

Kleinsmith, A., Bianchi-Berthouze, N., and Steed, A. (2011). Automatic recognition of non-acted affective postures. *Trans. Sys. Man Cyber. Part B*, 41(4):1027–1038.

Kleinsmith, A., De Silva, P. R., and Bianchi-Berthouze, N. (2006). Cross-cultural differences in recognizing affect from body posture. *Interact. Comput.*, 18(6):1371–1389.

Liscombe, J., Venditti, J., and Hirschberg, J. (2003). Classifying subject ratings of emotional speech using acoustic features. In *Proceedings of Interspeech'2003 - Eurospeech*.

Russell, J. A. (1994). Is there universal recognition of emotion from facial expression? a review of the cross-cultural studies. *Psychological Bulletin*, 115:102–141.

Savva, N. and Bianchi-Berthouze, N. (2012). Automatic recognition of affective body movement in a video game scenario. In Camurri, A. and Costa, C., editors, *Intelligent Technologies for Interactive Entertainment*, volume 78 of *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, pages 149–159. Springer Berlin Heidelberg.

Yannakakis, G. (2009). Preference learning for affective modeling. In *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, pages 1–6.

Yannakakis, G. and Hallam, J. (2011). Ranking vs. preference: A comparative study of self-reporting. In D'Mello, S., Graesser, A., Schuller, B., and Martin, J.-C., editors, *Affective Computing and Intelligent Interaction*, volume 6974 of *Lecture Notes in Computer Science*, pages 437–446. Springer Berlin Heidelberg.

Yannakakis, G. N. and Hallam, J. (2007). Towards optimizing entertainment in computer games. *Appl. Artif. Intell.*, 21(10):933–971.

Zeng, Z., Pantic, M., Roisman, G., and Huang, T. (2009). A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(1):39–58.