

Human Capabilities on Video-based Facial Expression Recognition

Matthias Wimmer¹, Ursula Zucker², and Bernd Radig²

¹ Faculty of Science and Engineering, Waseda University, Tokyo, Japan

² Institut für Informatik, Technische Universität München, Germany

Abstract. A lot of promising computer vision research has been conducted in order to automatically recognize facial expressions during the last decade. Some of them achieve high accuracy, however, it has not yet been investigated how accurately humans accomplish this task, which will introduce a comparable measure. Therefore, we conducted a survey on this issue and this paper evaluates the gathered information regarding the recognition rate and the confusion of facial expressions.

1 Introduction and Motivation

The psychologists Ekman and Friesen investigated social dependencies of facial expressions. They figure out six universal facial expressions [1] that are expressed and interpreted in the same way by humans all over the world, see Figure 1. Furthermore, they introduce the Facial Action Coding System (FACS) in order to precisely describe facial muscle activity [2]. Kanade et al. [3] gather a database (CKDB) of $I = 488$ short image sequences each showing one of the six universal facial expressions. Each sequence shows a neutral face at the beginning and then develops into one of the six universal expressions with peak activity.



Fig. 1. The six universal facial expressions as they occur in [3].

The computational task of facial expression interpretation is usually subdivided into three subordinate challenges, see Pantic et al. [5], detecting the face, extracting facial features, and inferring the facial expression. Michel et al. [4] train a Support Vector Machine (SVM) that determines the facial expression within video sequences of the CKDB by comparing the first frame with the neutral expression to the last frame with the peak expression. Schweiger et al. [6] compute the optical flow of several predefined regions within a human face in order to extract the facial features. Classification is conducted by neural networks.

However, the accuracy of these approaches cannot be stated realistically, because a comparable measure does not exist. Therefore, we conduct a comprehensive survey asking hundreds of persons to determine facial expressions. Similar to computer vision algorithms, humans are only provided visible information. The contribution of this paper is a realistic measure to state the accuracy of algorithms based on the accuracy of human beings recognizing facial expression.

2 Description of the Survey

We questioned a few hundred persons about the facial expressions visible in the CKDB. Note that this database only contains visible information and does not provide further communication channels or context information. The participants were shown randomly selected image sequences and they specified one of the six universal facial expressions or none in case they were not able to decide. Each participant annotated as many image sequences as he or she wanted.

3 Evaluation of the Survey’s Results

This section investigates which facial expressions are recognized easily and which are more likely to be confused. We gathered $q = 5413$ annotations specified by $P = 250$ persons³. On average, each participant annotated $\frac{q}{P} \approx 22$ image sequences, which results in $\frac{q}{I} \approx 11$ annotations per sequence.

For each sequence of the CKDB, licensed FACS-experts manually specified Action Units, but, unfortunately, they do not directly relate to one of the six universal facial expressions, in most cases. Therefore, we cannot decide whether or not the participants’ annotations are correct. This evaluation compares the annotations to one another, instead.

The set $\mathcal{E} = \{\text{happiness, sadness, disgust, fear, anger, surprise, none}\}$ contains the possible annotations. We subdivide the q annotations into the number of

³ The entire set of annotations is available on request (matthias.wimmer@cs.tum.edu).

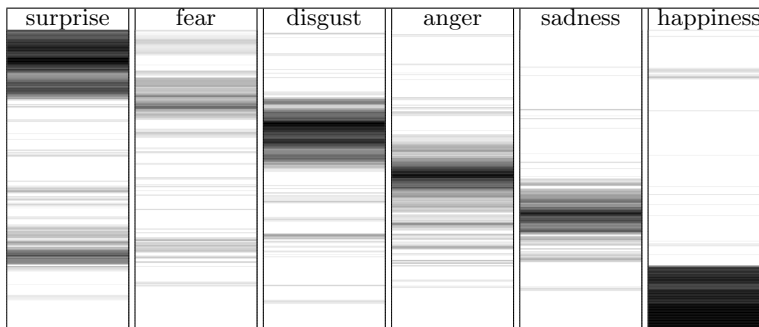


Fig. 2. The annotation rates for each facial expression of each image sequence sorted by similar annotation of the sequences. Darker regions denote a higher annotation rate.

annotations q_i for the image sequences i . Again, we subdivide each q_i into the number of annotations $q_{i,\epsilon}$ for the facial expression ϵ , see Equation 1. The annotation rate $r_{i,\epsilon}$ gives evidence about the number of annotations for the facial expression ϵ compared to the total amount of annotations for the sequence i .

$$q = \sum_{i=1}^I q_i, \quad q_i = \sum_{\epsilon \in \mathcal{E}} q_{i,\epsilon}, \quad r_{i,\epsilon} = \frac{q_{i,\epsilon}}{q_i} \quad (1)$$

Higher annotation rates indicate that the participants rather chose the same facial expression, which makes the annotation more reliable. Furthermore, $r_{\epsilon,i} \approx 0$ indicates that most participants did not specify ϵ for sequence i . Therefore, this image sequence does probably not show facial expression ϵ . Figure 2 illustrates the annotation rates for all 488 sequences. Every row denotes one image sequence i and indicates the annotation rate $r_{i,\epsilon}$ for all facial expressions $\epsilon \in \mathcal{E}$. We sort the rows of the table such that similarly specified image sequences are in adjacent rows. In this representation, the confusion of facial expressions is clearly visible. Happiness is best distinguished from other facial expressions. Sadness gets little confused with disgust or fear, but gets highly confused with anger or surprise. Fear is the hardest to tell apart.

Figure 3 shows the histograms of the annotation rates for all facial expressions. As mentioned above, well-recognized facial expressions have a lot of occurrences of $r_{i,\epsilon} \approx 0$ and of $r_{i,\epsilon} \approx 1$. Happiness is the most distinctive facial expression, because its histogram shows the most distinctive peaks for $r_{i,\epsilon} = 0$ and for $r_{i,\epsilon} = 1$.

3.1 Confusion Between Facial Expressions

Furthermore, we determine the level of confusion between different facial expressions. We consider two facial expressions to be confused if two different participants annotate the same image sequence with these expressions. $H(\epsilon_1 \wedge \epsilon_2)$ represents the number of sequences that are both annotated as ϵ_1 and as ϵ_2 .

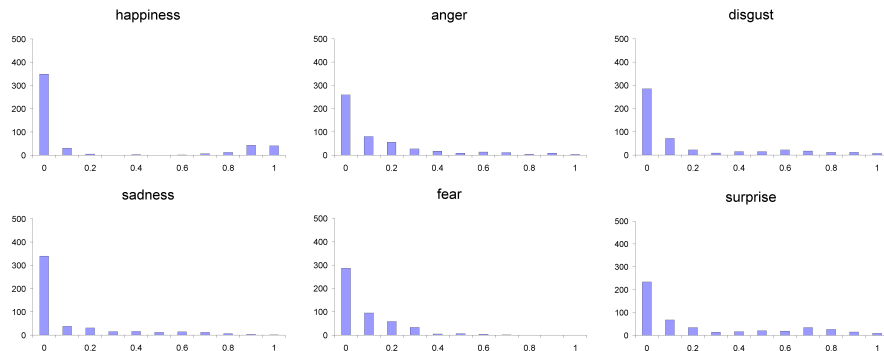


Fig. 3. Distribution of the annotation rate $r_{i,\epsilon}$ for each facial expression ϵ .

$\tau(\epsilon_1, \epsilon_2)$	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Anger	100.0%	42.4%	23.8%	7.3%	43.2%	28.9%
Disgust	42.4%	100.0%	32.6%	6.2%	19.3%	24.6%
Fear	23.8%	32.6%	100.0%	11.0%	15.8%	43.8%
Happiness	7.3%	6.2%	11.0%	100.0%	6.7%	14.5%
Sadness	43.2%	19.3%	15.8%	6.7%	100.0%	28.9%
Surprise	28.9%	24.6%	43.8%	14.5%	28.9%	100.0%

Table 1. The amount of confusion $\tau(\epsilon_1, \epsilon_2)$ between the six universal facial expressions.

$H(\epsilon_1 \vee \epsilon_2)$ represents the number of sequences that are either annotated as ϵ_1 or as ϵ_2 . The quotient $\tau(\epsilon_1, \epsilon_2) = \frac{H(\epsilon_1 \wedge \epsilon_2)}{H(\epsilon_1 \vee \epsilon_2)}$ determines the amount of confusion between two facial expressions ϵ_1, ϵ_2 .

Table 1 illustrates the confusion of either pair of facial expressions. The participants confused fear and surprise most often. According to FACS, some Action Units are similar in these two facial expressions. Further expressions, which get easily confused because of some coinciding Action Units are anger and sadness, anger and disgust, fear and disgust, and anger and surprise. People confused least between happiness and disgust, and happiness and sadness.

4 Conclusion

The interpretation of the data gathered by our survey shows that humans are not as good in determining facial expressions as computer vision researchers would expect them to be. These poor recognition rates partly originate from the fact that the CKDB does not contain natural expressions but acted ones. Furthermore, this recording was conducted in a laboratory environment rather than in a real-world scene. In our opinion, the most decisive reason for the poor results is the consideration of video information only. We expect humans to be more accurate being provided further information as well, such as audio information and long-term context information. Therefore, we recommend integrating this information into facial expression interpretation algorithms as well.

References

1. P. Ekman. Universals and cultural differences in facial expressions of emotion. In *Nebraska Symposium on Motivation*, vol 19, pp 207–283, 1972. U of Nebraska Press.
2. P. Ekman. Facial expressions. In T. Dalglish and M. Power, editors, *Handbook of Cognition and Emotion*, New York, 1999. John Wiley & Sons Ltd.
3. T. Kanade, J. F. Cohn, and Y. Tian. Comprehensive database for facial expression analysis. In *Int. Conf. on Automatic Face and Gesture Recognition*, pp 46–53, 2000.
4. P. Michel and R. El Kaliouby. Real time facial expression recognition in video using support vector machines. In *Int. Conf. on Multimodal Interfaces*, 2003.
5. M. Pantic and L. J. M. Rothkrantz. Automatic analysis of facial expressions: The state of the art. *IEEE PAMI*, 22(12):pp 1424–1445, 2000.
6. R. Schweiger, P. Bayerl, and H. Neumann. Neural architecture for temporal emotion classification. In *Affective Dialogue Systems, LNAI 3068*, pp 49–52, 2004.