Emotion recognition from posed and spontaneous dynamic expressions: Human observers vs. machine analysis

Eva G. Krumhuber

University College London

Dennis Küster

University of Bremen

Jacobs University Bremen

Shushi Namba

Hiroshima University

Datin Shah

University College London

Manuel G. Calvo

University of La Laguna

Eva G. Krumhuber and Datin Shah, Department of Experimental Psychology, University College London; Dennis Küster, Department of Mathematics and Computer Science, University of Bremen, and Department of Psychology and Methods, Jacobs University Bremen; Shushi Namba, Department of Psychology, Hiroshima University; Manuel G. Calvo, Department of Cognitive Psychology, University of La Laguna.

Correspondence concerning this article should be addressed to Eva G. Krumhuber, Department of Experimental Psychology, University College London, 26 Bedford Way, London WC1H 0AP, United Kingdom. E-mail: e.krumhuber@ucl.ac.uk

Word Count: 2493

Abstract

The majority of research on the judgment of emotion from facial expressions has focused on deliberately posed displays, often sampled from single stimulus sets. Herein, we investigate emotion recognition from posed and spontaneous expressions, comparing classification performance between humans and machine in a cross-corpora investigation. For this, dynamic facial stimuli portraying the six basic emotions were sampled from a broad range of different databases, and then presented to human observers and a machine classifier. Recognition performance by the machine was found to be superior for posed expressions containing prototypical facial patterns, and comparable to humans when classifying emotions from spontaneous displays. In both humans and machine, accuracy rates were generally higher for posed compared to spontaneous stimuli. The findings suggest that automated systems rely on expression prototypicality for emotion classification, and may perform just as well as humans when tested in a cross-corpora context.

Keywords: spontaneous, facial expression, emotion, dynamic, machine analysis

Emotion recognition from posed and spontaneous dynamic expressions: Human observers vs. machine analysis

Most past work on the perception of emotional expressions has relied on posed or acted facial behavior, often depicted in a static position at or very near the peak of an expression. Deliberately posed displays allow for good recognizability (e.g., Ekman, Friesen, & Ellsworth, 1972). However, due to their idealized and often exaggerated nature, they may be unrepresentative of spontaneous affective expressions commonplace in everyday life. Herein, we seek to assess emotion recognition from posed and spontaneous dynamic expressions, comparing classification performance between humans and machine.

Apart from their higher ecological validity, spontaneously displayed expressions often contain complex action patterns which can increase the ambiguity of their emotional content (Cohn, Ambadar, & Ekman, 2007). As a result, recognition accuracy has been argued to drop as spontaneous expressions move farther away from prototypical, stylized representations of an emotion (e.g., Motley & Camden, 1988; Naab & Russell, 2007; Nelson & Russell, 2013; Wagner, MacDonald, & Manstead, 1986, for a review see Calvo & Nummenmaa, 2016). Nonetheless, recent work points towards mixed evidence regarding the recognizability of posed and spontaneous expressions (Abramson, Marom, Petranker, & Aviezer, 2017), and suggests that the result may depend on the specific stimulus set used (Sauter & Fischer, 2018). The latter point is particularly pertinent in the context of automated facial expression analysis (AFEA).

Many machine-based systems have been trained on a few - often posed/acted - datasets (Pantic & Bartlett, 2007), raising concerns about their ability to generalize to the complexity of expressive behavior in spontaneous and real-world settings. Moreover, past efforts typically relied on in-house techniques for affect recognition. Given that AFEA is nowadays widely accessible, emotion classification using publicly/commercially available software (e.g. FaceReader, CERT, FACET) is of increasing research interest. Such software was recently found to perform similarly well (and often better) than human observers for prototypical facial expressions of standardized datasets (Del Líbano, Calvo, Fernández-Martín, & Recio, 2018; Lewinski, den Uyl, & Butler, 2014), but worse for subtle expressions that were non-stereotypical (Yitzhak et al., 2017) or produced by laypeople in the laboratory (Stöckli, Schulte-Mecklenbeck, Borer, & Samson, 2017). In none of the above studies, however, emotion recognition was tested in spontaneous affective displays.

The present research aims to fill this knowledge gap by investigating human and machine emotion recognition performance in posed *and* spontaneous facial expressions. It does so by providing cross-corpora results in which stimuli are sampled from a broad range of different databases. These include expressive behaviors ranging from directed or enacted portrayals (posed) to emotion-induced responses (spontaneous). Importantly, all of them contain dynamic expressions which are key to the differentiation between posed and spontaneous displays (Krumhuber, Kappas, & Manstead, 2013; Zloteanu, Krumhuber, & Richardson, 2018). Following common approaches, we focused on the classification of facial expressions portraying the six basic emotions. Instead of a single forced-choice task (which has been heavily criticized because it forces observers to choose a single emotion, Russell, 1993), participants indicated the relative extent of occurrence for multiple emotion categories of the same expression, thereby allowing maximum comparability to the machine recognition data.

Based on previous research pointing towards superior emotion classification from posed relative to spontaneous displays (Motley & Camden, 1988; Nelson & Russell, 2013), we predicted that recognition accuracy of posed expressions generally exceeds that of spontaneous ones; a finding which may be explained by the frequent occurrence of prototypical facial patterns when behavior is posed. Higher emotional prototypicality might further facilitate AFEA (e.g. Yitzhak et al., 2017), with the result that the machine performs better (or equally well) compared to humans in classifying emotions from posed expressions, while recognition accuracy should be similar (or worse) in the context of spontaneous expressions.

Method

Stimulus Material

Dynamic facial expressions in the form of video-clips were taken from 14 databases, and featured single person portrayals of at least four basic emotions (see Table S1). Nine of the databases contained posed facial expressions, emerging from instructions to perform an expression/facial action or scenario enactments. Five databases included spontaneous facial expressions that had been elicited in response to emotion-specific tasks or videos (for a review see Krumhuber, Skora, Küster, & Fou, 2017). For the purpose of the present study, we focused on the six basic emotions - anger, disgust, fear, happiness, sadness, surprise - as predefined by the dataset authors.[1] Two exemplars of each emotion category were randomly selected from every database, yielding 12 emotion portrayals per database. The two exceptions were DISFA and DynEmo, both of which contain only five and four of the basic emotions, respectively. This resulted in a total of 162 dynamic facial expressions (54 spontaneous, 108 posed) from 85 female and 77 male encoders. Stimuli lasted on average 5 s and were displayed in color (642 x 482 pixels).[2]

Human Observers

Participants. Seventy students (79% females) aged 18-24 years (M = 19.61, SD = 1.57) were recruited, ensuring 85% power to detect a small-sized effect (Cohen's f = .18, α = .05 two-tailed, r = 0.8) in a 2 (Machine vs. Human) x 2 (Posed vs. Spontaneous) x 6 (Emotion) within-between subjects repeated measures ANOVA. Participants were predominantly of White/Caucasian ethnicity (96%). Ethical approval was granted by the Department of Experimental Psychology, UCL.

Procedure. Participants were randomly presented with one of two exemplars of each emotion category from every database, netting 81 dynamic facial expressions per participant. Stimulus sequence was randomized using the Qualtrics software (Provo, UT), with each video-clip being played only once. For each facial stimulus, participants rated their confidence (from 0 to 100%) about the extent to which the expression reflected anger, disgust, fear, happiness, sadness, surprise, other emotion, and neutral (no emotion). If they felt that more than one category applied, they could respond using multiple sliders to choose the exact confidence levels for each response category. Ratings across the eight response categories had to sum up to 100%.

Machine Analysis

All video stimuli were submitted to automated analysis by means of FACET (iMotions, SDK v6.3). FACET is a commercial software for automatic facial expression recognition, originally developed by the company Emotient (based on the Computer Expression Recognition Toolbox (CERT) algorithm, Littlewort et al., 2011). FACET codes facial expressions both in terms of FACS Action Units (AU) as well as the 6 basic emotions. For details regarding the measurement of machine classification performance see the Supplementary Materials.

To assess the occurrence of emotion prototypes as predicted by Basic Emotion Theory (Ekman et al., 2002, p. 174), AU combinations indicative of full prototypes or major variants (comprising more lenient criteria) were scored as 1 or 0.75, respectively. We further calculated a weighted prototypicality score by summing the FACET confidence scores of AUs within a combination, and multiplying the sum scores by 1 (full prototype) or 0.75

(major variant). This resulted in a total prototype score, with higher numbers reflecting greater emotional prototypicality.

Results

Recognition confidence scores were calculated for the two exemplars of each emotion category from every database which served as the unit of analysis.[3] The mean target emotion recognition of 54.83% (*SD* = 27.84) for human observers and 61.91% (*SD* = 41.52) for machine analysis was significantly higher than chance, set conservatively at 25%, $t_{human}(161) = 13.64$, $p < .001$, $d = 1.07$, 95% CI [25.51, 34.16]; $t_{machine}(159) = 11.24$, $p < .001$, $d = 0.89$, 95% CI [30.43, 43.40] (Frank & Stennett, 2001). Overall, FACET outperformed human observers in target emotion classification, $Z = 2.70$, $p = .007$, $r = .21$, 95% CI [1.06, 13.53].[4]

When comparing recognition performance separately for posed and spontaneous expressions, results revealed a significant human vs. machine difference in the context of posed ($M_{\text{human}} = 61.95$, $SD = 25.17$ vs. $M_{\text{machine}} = 69.82$, $SD = 38.12$), $Z = 2.67$, $p = .008$, $r = .26$, 95% CI [0.01, 15.73], but not spontaneous expressions ($M_{\text{human}} = 39.40$, $SD = 27.20$ vs. $M_{\text{machine}} = 45.49$, $SD = 43.81$), $Z = 0.79$, $p = .428$, $r = .11$, 95% CI [-4.35, 16.54].

An analysis of the emotion prototype scores showed that posed portrayals were more prototypical in their facial AU patterns than spontaneous ones, $U = 1980.5$, $p = .002$, $r = .24$, 95% CI [5.14, 24.68] (see Figure 1 for mean prototype frequencies). This applied to all emotions (*U*s < 52, *p*s < .081) except for happiness whose prototypicality didn't differ as a function of elicitation condition ($U = 57$, $p = .217$, $r = .24$, 95% CI [-30.78, 8.38]). A regression analysis revealed that the prototypicality of an expression significantly predicted the machine advantage over humans in emotion classification, $\beta = .287$, $t(158) = 2.78$, $p = .006$, 95% CI [0.08, 0.49].

In both humans and machine, emotion recognition accuracy was on average higher for posed than spontaneous expressions, $U_{human} = 1654$, $p < .001$, $r = .35$, 95% CI [12.76, 29.90]; $U_{machine} = 2036$, $p = .004$, $r = .23$, 95% CI [10.23, 38.43]. As shown in Figure 2, this performance advantage applied to posed expressions of anger (human: $U = 9$, $p = .003$, $r = .61$, 95% CI [18.45, 68.00]), disgust (machine: $U = 48$, $p = .088$, $r = .33$, 95% CI [5.11, 57.22]), sadness (human: $U = 21$, $p = .005$, $r = .56$, 95% CI [10.74, 51.15]; machine: $U = 16$, $p = .001$, $r = .63$, 95% CI [33.31, 93.30]), fear (human: $U = 34$, $p = .007$, $r = .51$, 95% CI [8.23, 44.42]; machine: $U = 29$, $p = .003$, $r = .55$, 95% CI [9.73, 68.34]), and surprise (human: $U = 34$, $p = .007$, $r = .51$, 95% CI [8.78, 46.07]). Also, human observers made less use of the categories ‘other emotion’, $U = 2256.5$, $p = .019$, $r = .18$, and ‘neutral’, $U = 2001$, $p = .001$, $r = .26$, when rating posed than spontaneous expressions.

In order to quantify the similarity of confusions between machine and human, each matrix was transformed into a single vector (see Kuhn et al., 2017). Correlational analyses indicated a significant overlap between both matrices for posed expressions (*rho* = .637, $S = 2818$, $p < .001$) and spontaneous expressions (*rho* = .598, $S = 3123.8$, $p < .001$), suggesting that recognition patterns of target and non-target emotions were positively related in humans and machine.

Discussion

In this paper, we sought to compare emotion recognition rates from posed and spontaneous dynamic expressions. Rather than relying on single stimulus sets as done previously, numerous dynamic databases were employed that feature a variety of expression elicitation techniques. Whilst the small number of chosen stimuli per dataset may not be representative of the full database, we think that this approach importantly allows for a cross-corpora evaluation of posed and spontaneous expressions.

In accordance with prior findings (e.g. Motley & Camden, 1988), posed expressions were better recognized than spontaneous ones. Also, facial patterns were more prototypical in posed displays, which made classification by the machine highly successful. Similar to Yitzhak et al. (2017), FACET outperformed humans in the context of posed datasets; a recognition advantage which was driven by the prototypicality of expression. Hence, AFEA based on specific configurations of prototypical facial activity appears to be sufficiently robust (Zeng, Pantic, Roisman, & Huang, 2009). Although performance dropped when the stimuli were spontaneous, accuracy rates and confusion patterns were similar for humans and machine. This is an important finding as it suggests that AFEA can be equally sensitive to spontaneously occurring behavior (Bartlett et al., 2005).

To allow for a variety of potential interpretations of facial expressions, in the current study human observers could select multiple emotion labels as well as no/other emotion. Besides avoiding potential artifacts observed with forced-choice tasks (Frank & Stennett, 2001), the chosen approach was shown to reveal results that were similar to traditional paradigms without additional response options (see Supplementary Materials). Nonetheless, in the future it would be important to equate the number of response categories by presenting only six emotion terms. Also, a larger amount of portrayals could be included, potentially aiming for a full validation of the 14 dynamic sets. This may also provide a benchmark for comparison between different automated methods of measuring facial expressions. The present study provides the first evidence suggesting that computer-based systems perform as well as (and often better than) human judges in affect recognition from facial expressions sampled from a broad range of databases.

References

Abramson, L., Marom, I., Petranker, R., & Aviezer, H. (2017). Is fear in your head? A comparison of instructed and real-life expressions of emotion in the face and body. *Emotion, 17*(3), 557-565. doi:10.1037/emo0000252

Bartlett, M. S., Littlewort, G., Frank, M., Lainscsek, C. Fasel, I., & Movellan, J. (2005). Recognizing facial expression: Machine learning and application to spontaneous behavior. *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition* (pp. 568-573). San Diego, CA: IEEE. doi:10.1109/CVPR.2005.297

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality data? *Perspectives on Psychological Science, 6,* 3-5. doi:10.1177/1745691610393980

Calvo, M. G., & Nummenmaa, L. (2016). Perceptual and affective mechanisms in facial expression recognition: An integrative review. *Cognition and Emotion, 30*(6), 1081-1106. doi:10.1080/02699931.2015.1049124

Cohn, J. F., Ambadar, Z., & Ekman, P. (2007). Observer-based measurement of facial expression with the Facial Action Coding System. In J. A. Coan & J. J. B. Allen (Eds.), *Series in affective science. Handbook of emotion elicitation and assessment* (pp. 203-221). New York, NY: Oxford University Press

Del Líbano, M., Calvo, M. G., Fernández-Martín, & M., Recio, G. (2018). Discrimination between smiling faces: Human observers vs. automated face analysis. *Acta Psychologica, 187,* 19-29. doi:10.1016/j.actpsy.2018.04.019

Dente, P., Küster, D., Skora, L., & Krumhuber, E. G. (2017). Measures and metrics for automatic emotion classification via FACET. In J. Bryson, M. De Vos, & J. Padget

(Eds.), *Proceedings of the Conference on the Study of Artificial Intelligence and Simulation of Behaviour* (AISB) (pp. 164-167), Bath, UK (April).

Ekman, P., Friesen, W. V., & Ellsworth, P. (1972). *Emotion in the human face: Guidelines for research and an integration of findings.* Oxford, UK: Pergamon Press. doi:10.1192/bjp.122.1.108

Ekman, P., Friesen, W. V., & Hager, J. C. (2002). *The Facial Action Coding System* (2nd ed.). Salt Lake City, UT: Research Nexus.

Frank, M. G., & Stennett, J. (2001). The forced-choice paradigm and the perception of facial expressions of emotion. *Journal of Personality and Social Psychology*, *80*(1), 75-85. doi:10.1037/0022-3514.80.1.75

iMotions (2016). Facial Expression Analysis: The complete pocket guide. Retrieved from https://imotions.com/blog/what-is-facial-expression-analysis/

Krumhuber, E. G., Kappas, A., & Manstead, A. S. R. (2013). Effects of dynamic aspects of facial expressions: A review. *Emotion Review, 5*(1), 41-46. doi:10.1177/1754073912451349

Krumhuber, E., Skora, P., Küster, D., & Fou, L. (2017). A review of dynamic datasets for facial expression research. *Emotion Review, 9*(3), 280-292. doi:10.1177/1754073916670022

Kuhn, L. K., Wydell, T., Lavan, N., McGettigan, C., & Garriod, L. (2017). Similar representations of emotions across faces and voices. *Emotion, 17*(6), 912-937. doi:10.1037/emo0000282

Lewinski, P., den Uyl, T. M., & Butler, C. (2014). Automated facial coding: Validation of basic emotions and FACS AUs in FaceReader. *Journal of Neuroscience, Psychology, and Economics*, *7*(4), 227-236. doi:10.1037/npe0000028

Littlewort, G., Whitehill, J., Wu, T., Fasel, I., Frank, M., Movellan, J., & Bartlett, M. (2011). The computer expression recognition toolbox (CERT). *Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition* (pp. 298-305). Santa Barbara, CA: IEEE. doi:10.1109/FG.2011.5771414

Motley, M. T., & Camden, C. T. (1988). Facial expression of emotion: A comparison of posed expressions versus spontaneous expressions in an interpersonal communication setting. *Western Journal of Speech Communication*, *52*(1), 1-22. doi:10.1080/10570318809389622

Naab, P. J., & Russell, J. A. (2007). Judgments of emotion from spontaneous facial expressions of New Guineans. *Emotion, 7*(4), 736-744. doi:10.1037/1528-3542.7.4.736

Nelson, N. L., & Russell, J. A. (2013). Universality revisited. *Emotion Review, 5(1),* 8-15. doi:10.1177/1754073912457227

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, *45*, 867–872. doi:10.1016/j.jesp.2009.03.009

Palermo, R., & Coltheart, M. (2004). Photographs of facial expression: Accuracy, response times, and ratings of intensity. *Behavior Research Methods, Instruments, and Computers, 36*(4), 634-638. doi:10.3758/BF03206544

Pantic, M., & Bartlett, M. S. (2007). Machine analysis of facial expressions. In K. Delac & M. Grgic (Eds.), *Face recognition* (pp. 377–416). Vienna, Austria: I-Tech Education and Publishing.

Russell, J. A. (1993). Forced-choice response format in the study of facial expression. *Motivation and Emotion, 17,* 41-51. doi:10.1007/BF00995206

Sauter, D. A., & Fischer, A. H. (2018). Can perceivers recognise emotions from spontaneous expressions? *Cognition and Emotion, 32*(3), 504-515. doi:10.1080/02699931.2017.1320978

Shaver, P., Schwartz, J., Kirson, D., & O'Connor, C. (1987). Emotion knowledge: Further exploration of a prototype approach. *Journal of Personality and Social Psychology, 52*(6), 1061-1086. doi:10.1037//0022-3514.52.6.1061

Stöckli, S., Schulte-Mecklenbeck, M., Borer, S., & Samson, A. C. (2018). Facial expression analysis with AFFDEX and FACET: A validation study. *Behavior Research Methods*, *50*(4), 1446-1460. doi:10.3758/s13428-017-0996-1

Wagner, H., MacDonald, C., & Manstead, A. (1986). Communication of individual emotions by spontaneous facial expressions. *Journal of Personality and Social Psychology*, *50*(4), 737-743. doi:10.1037/0022-3514.50.4.737

Yitzhak, N., Giladi, N., Gurevich, T., Messinger, D. S., Prince, E. B., Martin, K., & Aviezer, H. (2017). Gently does it: Humans outperform a software classifier in recognizing subtle, nonstereotypical facial expressions. *Emotion, 17*(8), 1187-1198. doi:10.1037/emo0000287

Zeng, Z., Pantic, M., Roisman, G. I., & Huang, T. S. (2009). A survey of facial affect recognition methods: Audio, visual and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*, 39–58. doi:10.1109/tpami.2008.52

Zloteanu, M., Krumhuber, E. G., & Richardson, D. C. (2018). Detecting genuine and deliberate displays of surprise in static and dynamic faces. *Frontiers in Psychology*, *9,* 1184. doi:10.3389/fpsyg.2018.01184

Footnotes

[1] Due to lack of uniformity in emotion labelling across databases, amusement (BINED, DynEmo) and joy (ADFES, DISFA, GEMEP) were included under the umbrella of happiness. In one database (MPI), missing portrayals of surprise were substituted with those of disbelief, belonging to the same emotion family (Shaver, Schwartz, Kirson, & O'Connor, 1987). Action Unit configurations characteristic of the six basic emotions as proposed in the Facial Action Coding System manual (FACS, Ekman, Friesen, & Hager, 2002) were selected in the context of the D3DFACS database which itself does not include emotion labels.

[2] Portrayals that lasted longer than 15 s (BINED, DynEmo) were edited to display the emotional peak of the expression from onset (neutral face), through apex, to offset (if applicable), resembling portrayals from the majority of databases. None of the final facial stimuli exceeded 10 s in duration.

[3] Two portrayals (one happy, one disgust) from the BINED database could not be processed by FACET. For six cases in which the evidence values for both target and non-target emotions were below the set threshold (< 0), equal weightings were assigned to the six response options of a portrayal.

[4] A 2 (Machine vs. Human) x 2 (Posed vs. Spontaneous) x 6 (Emotion) ANOVA revealed a significant three-way interaction, $F(5, 148) = 3.29$, $p = .008$, $\eta_p^2 = .10$. This effect remained significant when portrayer gender and video duration were entered as covariates, $F(5, 146) = 3.21$, $p = .009$, $\eta_p^2 = .10$. Non-parametric tests were used to analyze human vs. machine differences in recognition performance (due to violations of the assumption of homogeneity of variance) and when comparing posed vs. spontaneous expressions (due to unequal cell sizes).

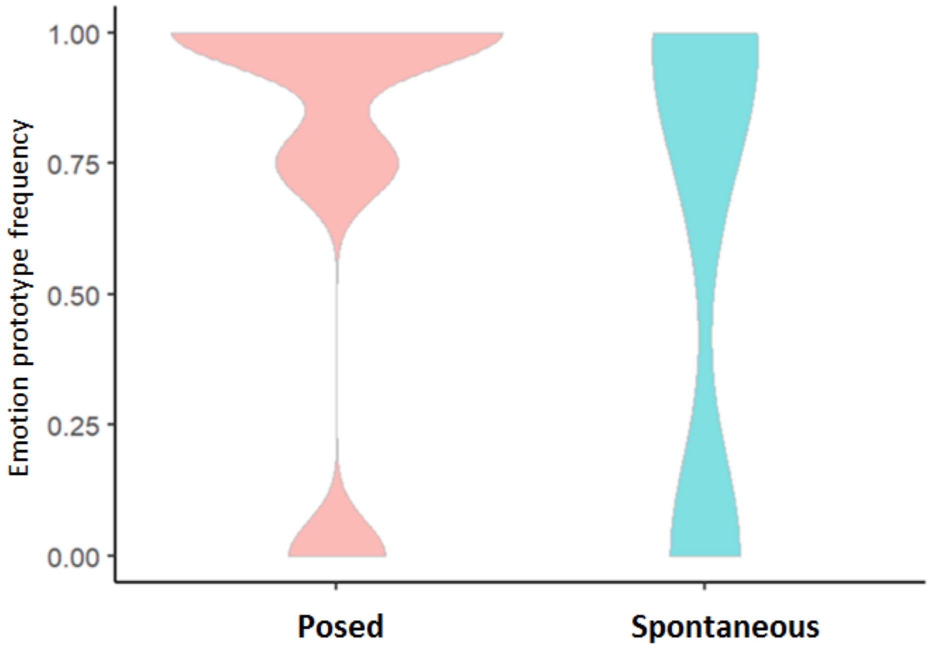

Figure 1. Mean frequency (as indicated by the density plot width) of facial emotion prototypes in posed and spontaneous expressions.

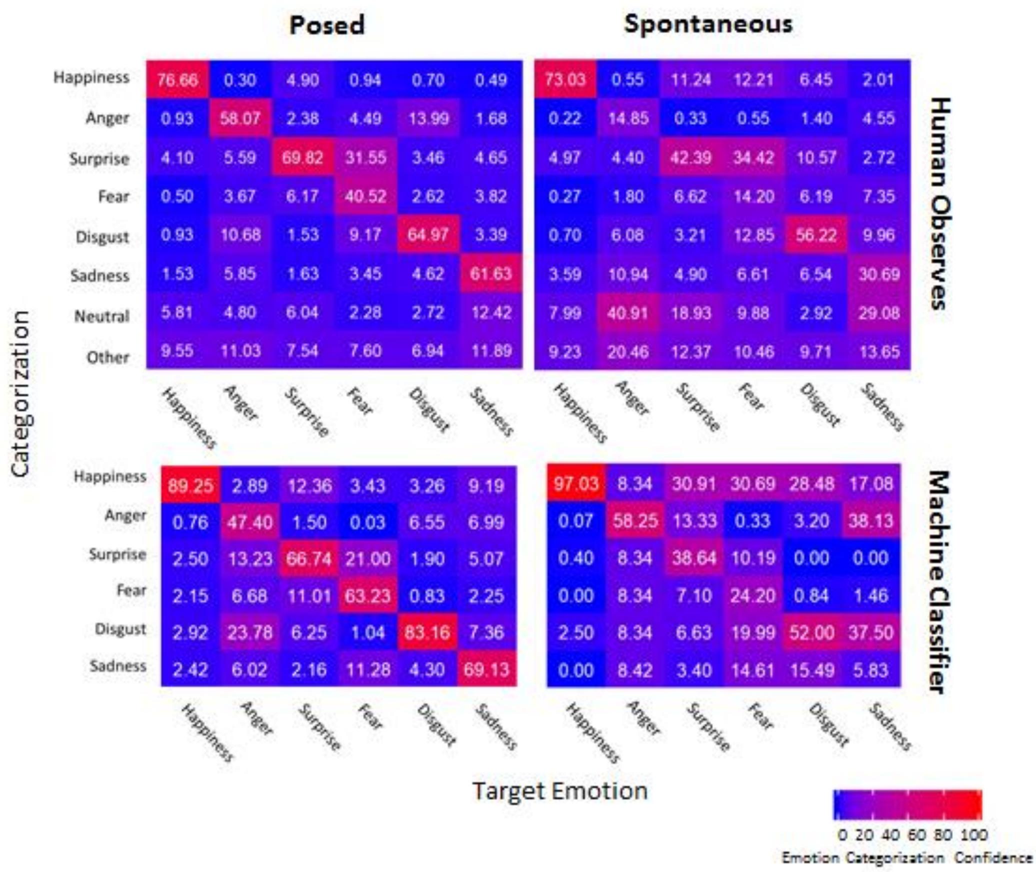

Figure 2. Confusion matrices of emotion categorization for human observers and FACET machine classifier averaged across dataset exemplars for posed and spontaneous expressions of each basic emotion.