Machine and human emotion classification diverge for naturalistic images

Version: 18 June 2024

Amy Dawel[1]*, Paige Mewton[1], Tayla Williams[1], Eva G. Krumhuber[2]

[1]School of Medicine and Psychology, The Australian National University, Australia.

[2]Department of Experimental Psychology, University College London, UK.

*Correspondence concerning this article should be addressed to Amy Dawel, School of Medicine and Psychology, The Australian National University, Canberra, ACT 2600, Australia. E-mail: amy.dawel@anu.edu.au

Acknowledgements and Funding Information

We thank Patrice Ford, Julia Gillett, Alison Schofield, and other student members of the ANU Emotions and Faces Lab for their contributions to stimulus development.

This research is supported by the Australian Government through the Australian Research Council's Discovery Projects funding scheme (Project No. DP220101026).

Abstract

The pursuit of affective computing aims to endow machines with the ability to read and respond to human emotions. Unfortunately, several high-profile publicly available machine classifiers are based on psychological theory that over-emphasises the prototypical, morphological features of six or seven basic emotional expressions. This approach is poorly suited to the complex and nuanced repertoire of naturalistic facial expressions; it also omits other physical information that humans use in emotion perception such as tears, eye-gaze, and facial colouring. Prior studies have focused on lab-generated expression stimuli, which are mostly posed, and evaluated machine and human classification against database labels. Here, we use a large, novel set of naturalistic expression stimuli (2,453 still images from 824 YouTube clips) to directly compare machine (Affdex) and human emotion classification. Overall, we found that human-machine agreement was strikingly low. At best, humans and Affdex agreed on two-thirds of happy stimuli, but this was more than double the rate for the next agreed-upon emotion (surprise = 33%). Agreement was under twelve percent for the remaining five emotion categories, with Affdex detecting only prototypical expressions. The findings underscore the importance of designing machine systems that can adapt to the subtleties of human emotional expressions for greater human-machine consistency.

Keywords: Facial expression; emotion; machine analysis; artificial intelligence; naturalistic

Machine and human emotion classification diverge for naturalistic images

There have been numerous attempts to create software capable of interpreting human emotional expressions, hoping to foster human-like rapport in customer service, gaming, psychological therapy, and other human-oriented domains (Lewinski et al., 2014; Littlewort et al., 2011; McDuff et al., 2016; Zeng et al., 2009). Many of these machine classifiers are rooted in basic emotions theory (BET; Ekman, 1992), which posits clear prototypes for seven basic emotions: anger, contempt, disgust, fear, happiness, sadness, and surprise. However, naturalistic expressions are often complex, mixed, and subtle (Cowen & Keltner, 2020; Schmidt et al., 2006), thereby challenging the compatibility of the BET approach with real-world emotional expressions.

Operationalized through morphological features known as "action units" (AUs; Ekman et al., 2002), BET prototypes represent emotional expressions as formulas. For instance, the "happy" formula combines AU12 "lip corner puller" with AU6 "cheek raiser". While BET-based machine classifiers excel at classifying intense posed expressions, they struggle with spontaneous expressions elicited in the lab (e.g., people viewing emotional movies or images) that may not exhibit the expected AU combinations (Krumhuber, Küster, Namba, Shah, et al., 2021; Krumhuber, Küster, Namba, & Skora, 2021; Yitzhak et al., 2018). BET-based approaches may further diverge from human perception as naturalistic expressions often incorporate additional non-AU information such as tears, eye and head gaze cues, and facial coloring (Kret, 2015; Küster et al., 2021; Thorstenson et al., 2021).

Previous studies evaluating BET-based emotion classification often used controlled, lab-generated expression stimuli (Dawel et al., 2022), where expressions are artificially posed or

induced by emotionally evocative stimuli rather than real-world interactions. Furthermore, database labels served as the "ground truth" for performance accuracy (e.g., Krumhuber, Küster, Namba, Shah, et al., 2021; Krumhuber, Küster, Namba, & Skora, 2021; Yitzhak et al., 2018). However, for developing software able to detect real-world social cues, naturalistic expressions should be used with human perception serving as the ground truth.[1] The present research aims to compare human versus machine emotion classification using a widely used commercial software (Affdex) to examine classification rates for a large set of naturalistic facial expressions.

Method

Stimuli

The full stimulus set comprised 2,453 still image frames extracted from 824 YouTube clips from various sources such as news, reality TV, dramas, adverts, talk/game shows, and vlogs. Each stimulus showed a facial expression displayed by young White adults (N = 1,402 expressors). The expressions were selected via two routes. First, three human coders identified the clearest version of each expression in each clip. Second, Affdex likelihood scores were used to identify the image frames with the highest evidence that each emotion was present, above a minimum threshold. Supplement S1 explains the selection process and criteria in detail. As with real-world expressions, these faces appear with varied lighting and viewpoints, and likely include both posed and genuinely-felt expressions. Stimuli were retained in the study if Affdex produced a full set of

[1] Alternatively, if software aims to classify people's emotional experiences, the ground truth should capture the expressor's felt emotion, such as physiological responses and/or self-reported emotions.

emotion classification output for the target image frame. Affdex classification was performed on the entire video-clips as the software is optimized for dynamic stimulus classification. However, our analyses were based on the output for individual image frames, aligning with how we measured human classification for these frames. Given that Affdex uses only the face region whilst humans typically integrate contextual information in their emotion judgements (Aviezer et al., 2008), we cropped the images for the human observer study so that only the head/face remained visible.

Machine Classification

Affdex outputs a score between 0 and 100 that "indicates the likelihood each <basic> emotion or facial expression to occur, as recognized by a human observer" (iMotions, 2018). Our main analyses used the Affdex default cut-off score of 50 to classify an emotion as present (score >= 50) or absent (< 50) for each image. If none of the seven emotion category scores for an image met the cut-off threshold, we labelled it as showing "no emotion" and excluded it from analyses. If more than one emotion score for an image met the cut-off threshold, we labelled it as showing "mixed emotion" and also excluded it from analyses. We conducted supplementary analyses with cut-off scores of 30 and 70 and found that these produced similar results (see Supplement S2). Note, Affdex also outputs likelihood scores from 0 to 100 for 21 AUs, which we used to calculate the prototypicality of facial expression stimuli (Krumhuber, Küster, Namba, Shah, et al., 2021).

Human Observers

Human observers were asked to label each image as showing one of the seven basic emotions, an "other emotion", or "no emotion". Each stimulus[2] was labelled by an average of 20.4 (range = 13 to 63) White adult human observers (total sample = 76 males, 169 females; aged 18-32 years, *M* = 27.4, *SD* = 3.6)[3] recruited from Prolific (www.prolific.com). The modal emotion labelling response among human observers served as the ground truth for each image. Stimuli were labelled as "no emotion" and excluded from the main analyses when the modal agreement was less than 50 percent. If an image had more than one modal response, we labelled it as "mixed emotion" and also excluded it from analyses. Note, we conducted supplementary analyses with minimum modal agreement rates at 30 and 70 percent and found that these produced similar results (Supplement S2).

Each observer completed one to three 30-minute testing sessions and renumerated £3.00 GBP per session. In each session, 200 randomly selected images were presented (image size on screen: 4.5 cm high x 4.3 cm wide) one at a time, with the nine labelling options below each image. Presentation and response times were unlimited.[4] If an observer recognised a person's

[2] One stimulus (1229_03_hf028.jpg) was accidentally repeated due to a programming error. Five observers labelled this stimulus twice, but all provided the same label both times.

[3] Data from an additional nine participants were excluded because they labelled fewer than 50 stimuli and thus had insufficient exposure to the range of expressions.

[4] Human observers also rated how genuine they thought each expression was using the procedure by Dawel et al. (2017).

face (e.g., because they followed the person on YouTube), they were asked to select a tenth option labelled “I know who this person is!”. Trials selecting this option were excluded to ensure observer judgments were not influenced by prior knowledge of the expressor (0.5 % of all trials excluded). This research was approved by The Australian National University Human Research Ethics Committee (protocol 2015/305) and carried out as per the Declaration of Helsinki.

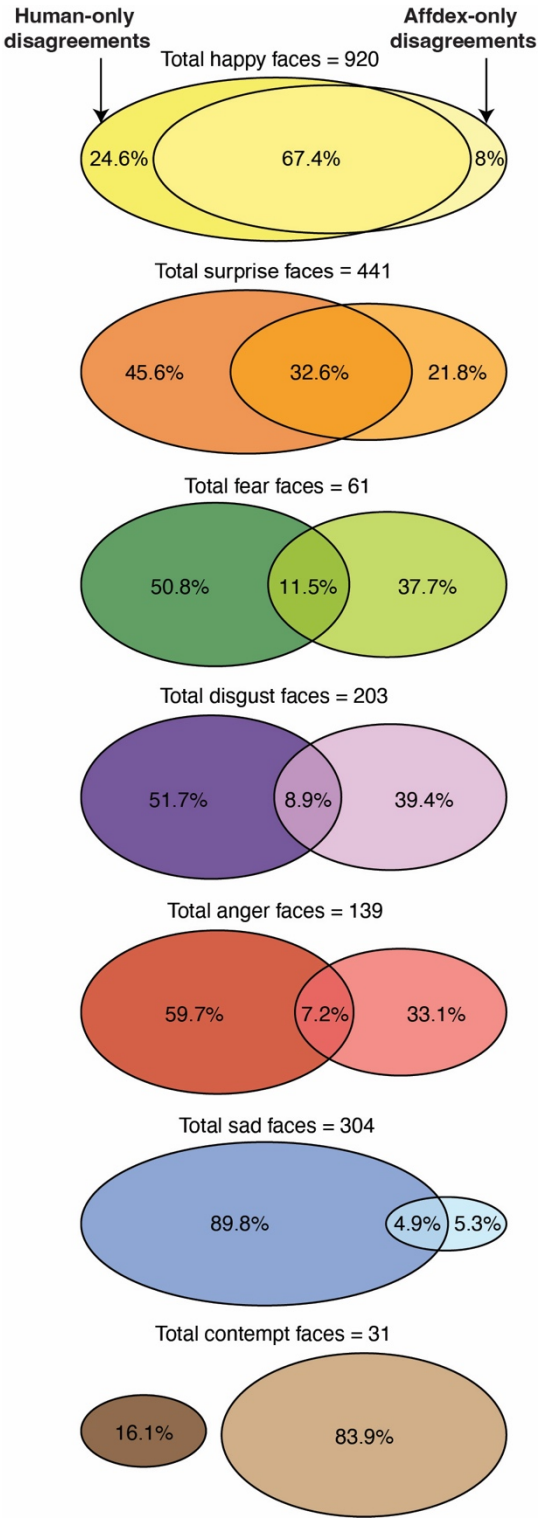
Results

Of the original 2,453 images, 1,921 (78%) met the cut-off thresholds for Affdex (likelihood score >= 50) or human classification (modal agreement >= 50%) or both.

Human-Machine Classification Agreement

Figure 1 presents human and Affdex classification rates for each emotion (see Supplement S3 for confusion matrices). Agreement was calculated as the percent classified as that emotion by both humans and Affdex out of the total stimuli per category identified by either source. Strikingly, agreement was below one-third for all emotions except happiness (67.4%), and below 10% for disgust (8.9%), anger (7.2%), sadness (4.9%), and contempt (0%). Disagreements occurred on both sides, sometimes asymmetrically. For example, within the disagreed components human observers classified an additional 273 images as sad (89.8% of the total sad images), whereas Affdex detected only 16 additional sad expressions (5.3%). Alternatively, Affdex detected 26 contempt expressions (83.9%), whereas humans classified only five stimuli as contempt (16.1%).

Fig 1 *Percentage of stimuli classified as each emotion by humans (left) and Affdex (right), with human-Affdex agreement indicated by circle overlap*

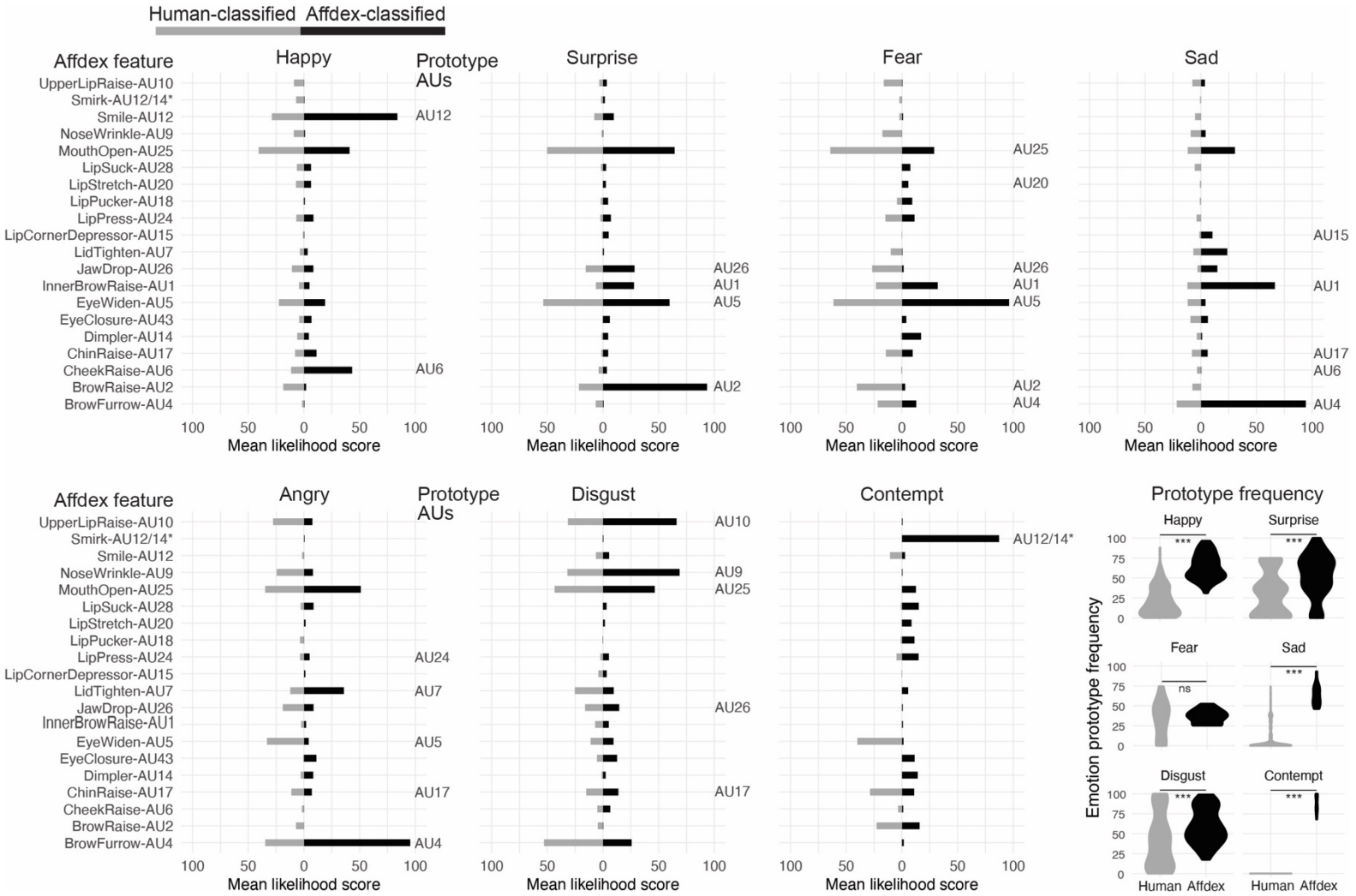


Action Units (AUs) for Disagreed-upon Classifications

Figure 2 shows the mean Affdex AU likelihood scores and prototype frequency for the disagreed classifications (i.e., the non-overlapping components of Figure 1). Findings show that AUs which signal a particular emotion according to BET (Ekman et al., 2002) were more likely to occur in the Affdex-only than human-only classifications, pointing towards Affdex's reliance on the presence of prototype-specific AUs. To evaluate the presence of emotion prototypes as predicted by BET, we used the method outlined by Krumhuber, Küster, Namba, Shah, et al. (2021). Briefly, we calculated a weighted prototypicality score for each stimulus by averaging the AU likelihood scores within a combination and multiplying this average by 1 if the AUs for a full prototype were present or 0.75 if the AUs were present for a major variant of that expression. Thus, a higher prototype score indicates greater evidence for the presence of an emotion prototype. Note, we were unable to perform this analysis for anger because Affdex does not provide scores for AU23, which BET posits as a key component of the anger prototype.

A 2 (classifier: Affdex-only, human-only) by 6 (emotion label) ANOVA on the prototype scores for the disagreed-upon stimuli revealed a significant main effect of classifier, $F(1, 1144) = 664.34$, $MSE = 475$, $p < .001$, indicating that Affdex-only classified stimuli were overall more prototypical than human-only classified ones. However, this effect was further qualified by a significant interaction between classifier and emotion label, $F(5, 1144) = 21.93$, $MSE = 475$, $p < .001$. Post hoc pairwise comparisons using Tukey's HSD method showed that stimuli classified by Affdex-only were significantly more prototypical than those classified by humans-only for all emotions (all ps $< .001$) except for fear ($p = .999$).

Fig 2 *Mean AU scores (from Affdex) for the disagreed-upon classifications (non-overlapping components of Figure 1)*



Note. [a]Smirk = left OR right activation of AU12 or AU14. The feature descriptors for each AU are from Affdex. Prototype AUs are as per our prototype analyses, based on Ekman et al. (2002, p. 174). Supplement S4 presents AU profiles for agreed versus Affdex-only classifications.

Discussion

The present study makes a novel contribution to the literature by testing a BET-based emotion classifier (Affdex) on naturalistic, in-the-wild expressions with ground truth defined by agreement amongst human observers. Our major finding was that Affdex emotion classification is strikingly divergent from that of humans for naturalistic expressions. This applied to all basic emotions except happiness, across three stimulus selection thresholds, and despite our naturalistic stimuli being pre-selected to show the clearest point of expression. Furthermore, neither human observers nor Affdex were forced into classifying any emotion as present (i.e., stimuli could be classified as showing "no emotion" or "other emotion"). Expressions classified by Affdex as showing a particular emotion were more likely to be prototypical, aligning with the software's reliance on BET prototypes. Overall, these results point to significant limitations in machine ability to label naturalistic expressions in a human-like manner, highlighting the need for improved algorithms that can better handle the complexity and variability of real-world emotional expressions.

The classification divergence in the present study points to an urgent need for psychological research into other commercial and open-source software, particularly AI-based models like ChatGPT-40. The ability of AI to interact with people in a human-like manner requires that AI is benchmarked against human performance, which psychology offers specialized expertise in assessing. Alternatively, where the aim is for AI to predict human affect and behavior, it may be more fruitful to focus on individual AUs (e.g., McDuff et al., 2013) as there is only a weak correlation between full facial displays and felt emotion (Durán & Fernández-Dols, 2021; Tcherkassof & Dupré, 2021). The present study included unequal numbers of stimuli for each

emotion. However, the observed frequencies track previous observations for real-life interactions (Calvo et al., 2014). Future work should also consider dynamic expressions and contextual information, to improve the ecological validity of machine classification. Nonetheless, the present study provides compelling evidence that an influential BET-based machine classifier performs poorly against human classification of emotions in naturalistic expressions.

References

Aviezer, H., Hassin, R. R., Ryan, J., Grady, C., Susskind, J., Anderson, A., Moscovitch, M., & Bentin, S. (2008). Angry, disgusted, or afraid? Studies on the malleability of emotion perception. *Psychological Science*, *19*(7), 724–732. https://doi.org/10.1111/j.1467-9280.2008.02148.x

Calvo, M. G., Gutiérrez-García, A., Fernández-Martín, A., & Nummenmaa, L. (2014). Recognition of facial expressions of emotion is related to their frequency in everyday life. *Journal of Nonverbal Behavior*, *38*(4), 549–567. https://doi.org/10.1007/s10919-014-0191-3

Cowen, A. S., & Keltner, D. (2020). What the face displays: Mapping 28 emotions conveyed by naturalistic expression. *American Psychologist*, *75*(3), 349–364. https://doi.org/10.1037/amp0000488

Dawel, A., Miller, E. J., Horsburgh, A., & Ford, P. (2022). A systematic survey of face stimuli used in psychological research 2000–2020. *Behavior Research Methods*, *54*(4), 1889–1901. Scopus. https://doi.org/10.3758/s13428-021-01705-3

Durán, J. I., & Fernández-Dols, J.-M. (2021). Do emotions result in their predicted facial expressions? A meta-analysis of studies on the co-occurrence of expression and emotion. *Emotion*, *21*(7), 1550–1569. https://doi.org/10.1037/emo0001015

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion*, *6*(3/4), 169–200.

Ekman, P., Friesen, W. V., & Hager, J. C. (2002). *Facial Action Coding System: Investigator's Guide*. Research Nexus division of Network Information Research Corporation.

iMotions. (2018). *Facial Expressions Affdex Methods Guide*. iMotions. help.imotions.com/hc/en-us/articles/115000512165-Facial-Expressions-Affdex-Methods-Guide

Kret, M. E. (2015). Emotional expressions beyond facial muscle actions. A call for studying autonomic signals and their impact on social perception. *Frontiers in Psychology*, *6*(May), 1–10. https://doi.org/10.3389/fpsyg.2015.00711

Krumhuber, E. G., Küster, D., Namba, S., Shah, D., & Calvo, M. G. (2021). Emotion recognition from posed and spontaneous dynamic expressions: Human observers versus machine analysis. *Emotion*, *21*(2), 447–451. https://doi.org/10.1037/emo0000712

Krumhuber, E. G., Küster, D., Namba, S., & Skora, L. (2021). Human and machine validation of 14 databases of dynamic facial expressions. *Behavior Research Methods*, *53*, 686–701. https://doi.org/10.3758/s13428-020-01443-y

Küster, D., Baker, M., & Krumhuber, E. G. (2021). PDSTD - The Portsmouth Dynamic Spontaneous Tears Database. *Behavior Research Methods*, *November*. https://doi.org/10.3758/s13428-021-01752-w

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. https://doi.org/10.1037/npe0000028

Littlewort, G., Whitehill, J., Wu, T., Fasel, I., Frank, M., Movellan, J., & Bartlett, M. (2011). The computer expression recognition toolbox (CERT). *Face and Gesture 2011*, 298–305. https://doi.org/10.1109/FG.2011.5771414

McDuff, D., El Kaliouby, R., Kodra, E., & Larguinat, L. (2013). Do emotions in advertising drive sales ? In *Affectiva* (pp. 1–13).

McDuff, D., Mahmoud, A., Mavadati, M., Amr, M., Turcot, J., & El Kaliouby, R. (2016). AFFDEX SDK: A cross-platform real-time multi-face expression recognition toolkit. *Conference on*

Human Factors in Computing Systems - Proceedings, *07-12-May-*, 3723–3726. https://doi.org/10.1145/2851581.2890247

Schmidt, K. L., Liu, Y., & Cohn, J. F. (2006). The role of structural facial asymmetry in asymmetry of peak facial expressions. *Laterality*, *11*(6), 540–561. https://doi.org/10.1080/13576500600832758

Tcherkassof, A., & Dupré, D. (2021). The emotion–facial expression link: Evidence from human and automatic expression recognition. *Psychological Research*, *85*(8), 2954–2969. https://doi.org/10.1007/s00426-020-01448-4

Thorstenson, C. A., Pazda, A. D., & Krumhuber, E. G. (2021). The influence of facial blushing and paling on emotion perception and memory. *Motivation and Emotion*, *45*(6), 818–830. https://doi.org/10.1007/s11031-021-09910-5

Yitzhak, N., Gilaie-dotan, S., & Aviezer, H. (2018). Neuropsychologia The contribution of facial dynamics to subtle expression recognition in typical viewers and developmental visual agnosia. *Neuropsychologia*, *117*(October 2017), 26–35. https://doi.org/10.1016/j.neuropsychologia.2018.04.035

Zeng, Z., Pantic, M., Roisman, G. I., & Huang, T. S. (2009). A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(1), 39–58. https://doi.org/10.1109/TPAMI.2008.52

Zloteanu, M., & Krumhuber, E. G. (2021). Expression authenticity: The role of genuine and deliberate displays in emotion perception. *Frontiers in Psychology*, *11*, 611248. https://doi.org/10.3389/fpsyg.2020.611248