ORIGINAL RESEARCH
published: 05 May 2021
doi: 10.3389/fpsyg.2021.627561

Automatic Facial Expression Recognition in Standardized and Non-standardized Emotional Expressions

Theresa Küntzler[1]†, T. Tim A. Höfling[2]† and Georg W. Alpers[2]*

[1] *Department of Politics and Public Administration, Center for Image Analysis in the Social Sciences, Graduate School of Decision Science, University of Konstanz, Konstanz, Germany,* [2] *Department of Psychology, School of Social Sciences, University of Mannheim, Mannheim, Germany*

Emotional facial expressions can inform researchers about an individual's emotional state. Recent technological advances open up new avenues to automatic Facial Expression Recognition (FER). Based on machine learning, such technology can tremendously increase the amount of processed data. FER is now easily accessible and has been validated for the classification of standardized prototypical facial expressions. However, applicability to more naturalistic facial expressions still remains uncertain. Hence, we test and compare performance of three different FER systems (Azure Face API, Microsoft; Face++, Megvii Technology; FaceReader, Noldus Information Technology) with human emotion recognition (A) for standardized posed facial expressions (from prototypical inventories) and (B) for non-standardized acted facial expressions (extracted from emotional movie scenes). For the standardized images, all three systems classify basic emotions accurately (FaceReader is most accurate) and they are mostly on par with human raters. For the non-standardized stimuli, performance drops remarkably for all three systems, but Azure still performs similarly to humans. In addition, all systems and humans alike tend to misclassify some of the non-standardized emotional facial expressions as neutral. In sum, emotion recognition by automated facial expression recognition can be an attractive alternative to human emotion recognition for standardized and non-standardized emotional facial expressions. However, we also found limitations in accuracy for specific facial expressions; clearly there is need for thorough empirical evaluation to guide future developments in computer vision of emotional facial expressions.

Keywords: recognition of emotional facial expressions, software evaluation, human emotion recognition, standardized inventories, naturalistic expressions, automatic facial coding, facial expression recognition, specific emotions

Edited by:
Yizhang Jiang,
Jiangnan University, China

Reviewed by:
Yuanpeng Zhang,
Nantong University, China
Hillel Aviezer,
Hebrew University of Jerusalem, Israel

Correspondence:
Theresa Küntzler
theresa.kuentzler@uni-konstanz.de

† *These authors share first authorship*

Specialty section:
This article was submitted to Emotion Science, a section of the journal Frontiers in Psychology

Received: *11 November 2020*
Accepted: *11 March 2021*
Published: *05 May 2021*

Citation:
Küntzler T, Höfling TTA and Alpers GW (2021) Automatic Facial Expression Recognition in Standardized and Non-standardized Emotional Expressions. Front. Psychol. 12:627561. doi: 10.3389/fpsyg.2021.627561

1. INTRODUCTION

Detecting emotional processes in humans is important in many research fields such as psychology, affective neuroscience, or political science. Emotions influence information processing (e.g., Marcus et al., 2000; Meffert et al., 2006; Fraser et al., 2012; Soroka and McAdams, 2015), attitude formation (e.g., Lerner and Keltner, 2000; Marcus, 2000; Brader, 2005), and decision making (Clore et al., 2001; Slovic et al., 2007; Pittig et al., 2014). One well-established strategy to measure emotional reactions of individuals is to track their facial expressions (Scherer and Ellgring, 2007; Keltner and Cordaro, 2017). The classic approach to analyse emotional facial responses is either an expert observation such as the Facial Action Coding System (FACS) (Sullivan and Masters, 1988; Ekman and Rosenberg, 1997; Cohn et al., 2007) or direct measurement of facial muscle activity with electromyography (EMG) (Cohn et al., 2007). Both are, however, time-consuming with respect to both, application and analysis.

A potential alternative to facilitate, standardize, and scale research on facial expressions is automatic image-based Facial Expression Recognition (FER), which has recently emerged from computer vision technology. Using machine learning, algorithms are being developed that extract emotion scores from observed facial expressions (Goodfellow et al., 2015; Arriaga et al., 2017; Quinn et al., 2017), which is considerably more time and cost efficient compared to classical approaches (Bartlett et al., 1999). FER is easily accessible to researchers of all fields and is increasingly used by the scientific community. Applications can be found, for example, in psychology, where such algorithms are used to predict mental health from social media images (Yazdavar et al., 2020), to validate interventions for autism (Wu et al., 2019), or to screen for Parkinson's disease (Jin et al., 2020). A sociological example is the assessment of collective happiness in society from social media images (Abdullah et al., 2015). In political science, one example is the study of representation of politicians in the media using FER (Boxell, 2018; Peng, 2018; Haim and Jungblut, 2020). Furthermore, the technology is used in consumer and market research, for example to predict advertisement efficiency (Lewinski et al., 2014; Teixeira et al., 2014; Bartkiene et al., 2019).

1.1. Prototypical vs. Naturalistic Facial Expressions

Training and testing of FER tools is typically conducted on data sets, which contain prototypical and potentially exaggerated expressions (Dhall et al., 2012). The images of these inventories are created under standardized (detailed instructions for the actors) and well-controlled conditions (e.g., lighting, frontal face angle; Lewinski et al., 2014; Calvo et al., 2018; Stöckli et al., 2018; Beringer et al., 2019; Skiendziel et al., 2019). As a result, the classification performance of FER systems and its generalizability to non-standardized and more naturalistic facial expressions is uncertain.

For prototypical facial expressions, FER also corresponds well to human FACS coding (Bartlett et al., 1999; Tian et al., 2001; Skiendziel et al., 2019) and non-expert human classification (Bartlett et al., 1999; Lewinski, 2015; Calvo et al., 2018; Stöckli et al., 2018). Accuracy is high for static images (Lewinski et al., 2014; Lewinski, 2015; Stöckli et al., 2018; Beringer et al., 2019) as well as for dynamic facial expressions from standardized inventories (Mavadati et al., 2013; Zhang et al., 2014; Yitzhak et al., 2017; Calvo et al., 2018). There is also growing evidence that FER provides valid measures for most emotion categories if naive participants are instructed to pose intense emotional facial expressions in a typical lab setting with frontal face recording and good lighting condition (Stöckli et al., 2018; Beringer et al., 2019; Sato et al., 2019; Kulke et al., 2020). However, all of these studies present their participants prototypical facial expression and instruct them to mimic these visual cues. This might result in an overestimation of FER performance in comparison to non-standardized facial expressions and moreover truly naturalistic emotional facial expressions.

Previous research also documents systematic misclassification of different FER systems and emotion categories. For fear, studies find a consistently lower accuracy compared to other emotion categories (Lewinski et al., 2014; Stöckli et al., 2018; Skiendziel et al., 2019). Some studies also report a substantial decrease in accuracy for anger (Lewinski et al., 2014; Stöckli et al., 2018; Dupré et al., 2020), whereas Skiendziel et al. (2019) report an improvement of this measurement in their study. Less consistently, sadness (Lewinski et al., 2014; Skiendziel et al., 2019) and disgust are also found to be error prone (Skiendziel et al., 2019). In contrast, the facial expression of joy is systematically classified with the highest accuracy (Stöckli et al., 2018; Skiendziel et al., 2019; Dupré et al., 2020). When looking at confusion between emotions in prior studies, FaceReader shows a tendency toward increased neutral measures for all other emotions (Lewinski et al., 2014) and a tendency to misclassify fearful faces as surprise (Stöckli et al., 2018; Skiendziel et al., 2019). Studies that compared different FER systems consistently find a large variation in performance between systems (Stöckli et al., 2018; Dupré et al., 2020) which underlines the need for comparatives studies.

Besides a general lack of studies, that directly compare different FER systems, empirical validation of FER to recognize emotional facial expressions is limited to intensely posed expressions. In contrast to those images, naturalistic or spontaneous facial expressions show stronger variations and are often less intense in comparison to standardized facial expressions (Calvo and Nummenmaa, 2016; Barrett et al., 2019). For example Sato et al. (2019) find a strong decrease in FER performance if participants respond spontaneously to imagined emotional episodes. Höfling et al. (2020) report strong correlations of FER parameters and participants' emotion ratings that spontaneously respond to pleasant emotional scenes, but find no evidence for a valid FER detection of spontaneous unpleasant facial reactions. Other studies report a decrease in FER emotion recognition for more subtle and naturalistic facial expressions (Höfling et al., 2021) and find a superiority of humans to decode such emotional facial responses (Yitzhak et al., 2017; Dupré et al., 2020). However, the data sets applied are still comprised of images collected in a controlled lab setting, with little variation on lighting, camera angle, or age of the subject which might further decrease FER performance under less restricted recording conditions.

1.2. Aims, Overview, and Expectations

In summary, FER offers several advantages in terms of efficiency and we already know that it performs well on standardized, prototypical emotional facial expressions. Despite many advantages of FER application and their validity to decode prototypical facial expression, the quality of the expression measurement and its generalizability to less standardized facial expressions is uncertain. Because the underlying algorithms remain unclear to the research community, including the applied machine-learning and its specific training procedure, empirical performance evaluation is urgently needed. Hence, this paper has two main aims: First, we provide an evaluation and a comparison of three widely used systems that are trained to recognize emotional facial expressions (FaceReader, Face++, and the Azure Face API) and compare them with human emotion recognition data as a benchmark. Second, we evaluate the systems on acted standardized and non-standardized emotional facial expressions: The standardized facial expressions are a collection of four facial expression inventories created in a lab setting displaying intense prototypical facial expressions [The Karolinska Directed Emotional Faces (Lundqvist et al., 1998), the Radboud Faces Database (Langer et al., 2010), the Amsterdam Dynamic Facial Expression Set (Van der Schalk et al., 2011), and the Warsaw Set of Emotional Facial Expression (Olszanowski et al., 2015)]. To approximate more naturalistic emotional expressions, we use a data set of non-standardized facial expressions: The Static Facial Expressions in the Wild data set (Dhall et al., 2018), which is built from movie scenes and covers a larger variety of facial expressions, lighting, camera position, and actor ages.

FER systems provide estimations for the intensity of specific emotional facial expressions through two subsequent steps: The first step is face detection including facial feature detection and the second step is face classification into an emotion category. For face detection, we expect that different camera angles, but also characteristics of the face such as glasses or beards will increase FER face detection failures resulting in higher rates of drop out. We expect the standardized expressions to result in less drop out due to failures in face detection, since the camera angle is constantly frontal, and no other objects such as glasses obstruct the faces. Correspondingly, we expect more drop out in the non-standardized data set, which means there are more images where faces are not detected, since the variability of the facial expressions is higher. For the second step (i.e., emotion classification), we expect strong variation between emotion categories (e.g., increased performance for joy faces, decreased performance on fear faces). We further expect a tendency toward the neutral category and a misclassification of fear as surprise. As explained for the drop outs, we assume the non-standardized images to be more variable and therefore more difficult to classify. The overall performance on the non-standardized data is therefore expected to be lower. This research provides important information about the generalizability of FER to more naturalistic, non-standardized emotional facial expressions and moreover the performance comparison of specific FER systems.

2. MATERIALS AND METHODS

We use three different facial expression recognition tools and human emotion recognition data to analyze emotional facial expressions in more or less standardized facial expressions. As an approximation to standardized and non-standardized facial expressions we analyze static image inventories of actors who were instructed to display prototypical emotional expressions and, in addition, an inventory of actors displaying more naturalistic emotional facial expressions in movie stills. We extract probability parameters for facial expressions corresponding to six basic emotions (i.e., joy, anger, sadness, disgust, fear, surprise, and neutral) from all tools. As a benchmark, we collect data from human raters who rated subsets of the same images.

2.1. Images of Facial Expressions

We test the different FER tools as well as human facial recognition data on standardized and non-standardized emotional facial expressions displayed in still images. All selected inventories are publicly available for research and contain emotional facial expression images of the basic emotion categories. **Table 1** displays the emotion categories and image distributions for both data sets (i.e., standardized and non-standardized) including drop out rates specifically for the three FER tools.

Standardized facial expressions are a collection of images created in the lab with controlled conditions (i.e., good lighting, frontal head positions, directed view) displaying prototypical expressions of clearly defined emotions. In order to maximize image quantity and introduce more variability, the prototypical images consist of four databases: (1) The Karolinska Directed Emotional Faces contains images of 35 males and 35 females between 20 and 30 years old (Lundqvist et al., 1998). The present study uses all frontal images (resolution: 562 × 762). (2) The Radboud Faces Database, which contains images of facial expressions of 20 male and 19 female Caucasian Dutch adults (Langer et al., 2010). We used the subset of adult models looking straight into the camera with images taken frontal (resolution: 681 × 1,024). (3) The Amsterdam Dynamic Facial Expression Set, from which we used the still image set (resolution: 720 × 576). The models are distinguished between being Northern-European (12 models, 5 females) and Mediterranean (10 models, 5 of them female; Van der Schalk et al., 2011). (4) The Warsaw Set of Emotional Facial Expression offers images of 40 models (16 females, 14 males) displaying emotional facial expressions (Olszanowski et al., 2015). Images are taken frontal and the complete set is used in this study (resolution: 1,725 × 1,168). This results in an overall of 1,246 images evenly distributed over the relevant emotion categories.

Non-standardized facial expressions stem from a data set that was developed as a benchmark test for computer vision research for more naturalistic settings. The Static Facial Expressions in the Wild (SFEW) data set consists of stills from movie scenes that display emotions in the actors' faces. Examples of movies are "Harry Potter" or "Hangover" (Dhall et al., 2018). This study uses the updated version (Dhall et al., 2018). The data set was compiled using the subtitles for the deaf and hearing impaired and closed caption subtitles. These subtitles contain not only the spoken text, but additional information about surrounding sounds, such as laughter. The subtitles were automatically searched for words suggesting emotional content. Scenes resulting from this search were then suggested to trained human coders, who classified and validated the final selection of emotional facial expressions for this inventory (Dhall et al.,

TABLE 1 | Category distributions of test data and drop outs of Azure, Face++, and FaceReader.

	Standardized data							
	Neutral	**Happy**	**Sad**	**Fear**	**Angry**	**Surprise**	**Disgust**	**Overall**
Absolute frequency of images	178	178	178	178	178	178	178	1246
Relative frequency of images	14.3	14.3	14.3	14.3	14.3	14.2	14.2	100
	Drop out rates (percent per category)							
Face++	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Azure	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
FaceReader	0.6	0.0	0.0	1.1	1.1	2.2	1.2	0.88
	Non-standardized data							
	Neutral	**Happy**	**Sad**	**Fear**	**Angry**	**Surprise**	**Disgust**	**Overall**
Absolute frequency of images	236	270	245	143	254	151	88	1387
Relative frequency of images	17.0	19.5	17.7	10.3	18.3	10.9	6.3	100
	Drop out rates (percent per category)							
Face++	0.0	0.7	0.8	0.7	0.4	0.0	1.1	0.5
Azure	16.9	11.1	25.3	26.6	25.2	17.9	23.9	20.3
FaceReader	73.3	70.0	75.1	79.0	76.8	74.8	69.3	74.2

Percentages are rounded to the first decimal. The base of the percentage is the respective total of each category. Reading example: Azure did not find a face in 16.9% of the 236 neutral images of the non-standardized data and a total of 20.3% of the 1387 images dropped out because of no face detection.

2012). We use these images to rigorously test how well the systems perform on images that are not prototypical and not taken under standardized conditions (variable lighting and head positions). The inventory consists of 1,387 images (resolution: 720 × 576) which are unevenly distributed across emotion categories (minimum of 88 images for disgust and a maximum of 270 images for joy).

2.2. Facial Expression Recognition Tools

We test three FER tools: The Microsoft Azure Face API (Version 1.0, Microsoft), Face++ (Version 3.0, Megvii Technology) and FaceReader (Version 8.0, Noldus Information Technology). The first two are easily accessible APIs, which also offer a free subscription. FaceReader is a software to be installed locally on a computer and is well-established in the research community. Each of the systems allow to analyse faces in images, with functions such as face detection, face verification, and emotion recognition. They all provide probability scores for neutral, joy, sadness, anger, disgust, fear, and surprise. While scores of Azure and FaceReader are between 0 and 1, Face++ uses a scale from 1 to 100. We thus rescale Face++ scores to 0 to 1. FaceReader specifically provides an additional quality parameter and it is suggested to remove images, if the quality of face detection is too low. Therefore, we remove all images with a quality parameter below 70%.

2.3. Human Emotion Recognition

As a benchmark for the FER results we collected emotion recognition data of humans who each rate a random subsample of up to 127 of the 2,633 images each in an online study. Participants who rated less than 20 images are excluded for further analyses (17 participants rated between 20 and 126 pictures). This results in 101 participants (58 female, 42 male, 1 diverse, $M_{age} = 29.2$, $SD_{age} = 9.1$) who rated on average 116.1 (SD = 28.1) images. Twenty-five images were randomly not rated by any participants ($< 1\%$). Participants were instructed to classify facial expression as neutral, joy, sadness, anger, disgust, fear, surprise, or another emotion. Multiple choices were possible. In addition, the perceived genuineness of the expressed emotion was rated on a 7-point Likert scale (1 -very in-genuine, 7 - very genuine). All ratings are averaged per image to improve comparability to the metric provided by the FER tools. This results in percentages of emotion ratings and average values per image for the genuineness ratings.

2.4. Analyses

First, we analyze the human raters' scores for perceived genuineness and emotion classification as a manipulation check for the two data sets of facial expressions. Differences between the genuineness of non-standardized vs. standardized facial expressions are tested statistically for all images as well as separately for all emotion categories utilizing independent *t*-tests. Correspondingly, we analyze the human emotion recognition data to provide a benchmark for the FER comparison. Again we statistically test for differences between non-standardized vs. standardized facial expressions for all emotion categories utilizing independent *t*-tests. In addition, we calculate one-sample *t*-tests against zero to estimate patterns of misclassification within human emotion recognition. Cohen's d is reported for all *t*-tests.

Second, we test the performance of face detection. As described above, FER is a two step process of first face detection and second emotion classification. To test performances on face detection, we check for how many images a specific tool gives no result (drop out rate).

Third, we calculate several indices of emotion classification (i.e., accuracy, sensitivity, and precision) for the three FER tools to report performance differences descriptively. In order to evaluate emotion classification, each algorithm's output is compared to the original coding of the intended emotional facial expression category (i.e., ground truth). The different tools return values for each emotion category. We define the category with the highest certainty as the chosen one, corresponding to a winner–takes–all principle[1]. A general indication of FER performance is the accuracy, which is the share of correctly identified images out of all images, where a face is processed (thus, excluding drop out)[2]. Other excellent measures to evaluate emotion classification are category specific sensitivity and precision. Sensitivity describes the share of correctly predicted images out of all images truly in the respective category. It is a measure of how well the tool does in detecting a certain category. Precision is the share of correctly predicted images out of all images predicted as one category. In other words, precision is a measure of how much we can trust the categorization of the tool. In order to identify patterns of classifications, we additionally build confusion matrices for the FER measurement and true categories.

Fourth, we report differences in emotion recognition performance between the three systems and human data with Receiver Operating Characteristic (ROC) analysis and statistical testing of the corresponding Area Under the Curve (AUC). ROC analysis is initially a two-class classification strategy. In order to apply the ROC rationale to a multi-class classification, we consider each probability given to a category as one observation. In other words, each image makes up for seven observations for each tool. The ROC curve plots a true positive share against a false positive share for varying probability thresholds above which a category is considered correct. A good classifier gives low probabilities to wrong classifications and high probabilities to correct classifications. This is measured by the AUC. Better classifiers give larger AUCs. We compare AUCs of the different algorithms pairwise, using a bootstrapping method with 2,000 draws (Robin et al., 2011).

Analyses are conducted in R (R Core Team, 2019), using the following packages (alphabetical order): caret (Kuhn, 2020), data.table (Dowle and Srinivasan, 2020), dplyr (Wickham et al., 2020), extrafont (Chang, 2014), ggplot2 (Wickham, 2016), httr (Wickham, 2020), jsonlite (Ooms, 2014), patchwork (Pedersen, 2020), plotROC (Sachs, 2017), pROC (Robin et al., 2011), purrr (Henry and Wickham, 2020), RColorBrewer (Neuwirth, 2014), stringr (Wickham, 2019), tidyverse (Wickham et al., 2019).

3. RESULTS

3.1. Human Raters: Genuineness of Facial Expressions

We test for differences between standardized and non-standardized facial expression inventories regarding their perceived genuineness (see **Figure 1A**). Analysis shows that the non-standardized facial expressions are perceived as much more genuine compared to the standardized facial expressions [standardized inventories: $M = 4.00$, $SD = 1.43$; non-standardized inventory: $M = 5.64$, $SD = 0.79$; $t(2606) = 36.58$, $p < 0.001$, $d = 1.44$]. In particular, non-standardized facial expressions are rated as more genuine for anger, $t(426) = 27.97$, $p < 0.001$, $d = 2.75$, sadness, $t(418) = 25.55$, $p < 0.001$, $d = 2.43$, fear, $t(317) = 21.10$, $p < 0.001$, $d = 2.38$, disgust, $t(263) = 18.10$, $p < 0.001$, $d = 2.36$, surprise, $t(322) = 16.02$, $p < 0.001$, $d = 1.79$, and joy, $t(441) = 5.58$, $p < 0.001$, $d = 0.54$, whereas among the standardized inventories neutral facial expressions are rated more genuine, $t(407) = 2.36$, $p = 0.019$, $d = 0.24$. These results support the validity of the selection of image test data—the standardized facial expressions are perceived less genuine compared to the non-standardized facial expressions.

3.2. Human Raters: Emotion Recognition

Next, we analyze the human emotion ratings (see **Figures 1B–H**). Comparisons against zero show that for most emotion categories, classifications are highest for the correct category. The only exception are non-standardized disgust faces that are more often categorized as angry, $t(87) = 7.99$, $p < 0.001$, $d = 0.85$, than disgusted, $t(87) = 4.40$, $p < 0.001$, $d = 0.47$. In addition, fearful faces are also misclassified (or at least co-classified) as surprise for standardized, $t(175) = 18.22$, $p < 0.001$, $d = 1.37$, and non-standardized facial expressions, $t(142) = 10.69$, $p < 0.001$, $d = 0.89$. A comparison between standardized and non-standardized data reveals a strong increase in neutral ratings for non-standardized emotion categories [disgust: $t(263) = 15.03$, $p < 0.001$, $d = 1.96$; surprise: $t(322) = 14.33$, $p < 0.001$, $d = 1.60$; fear: $t(317) = 9.54$, $p < 0.001$, $d = 1.07$; sadness: $t(418) = 9.01$, $p < 0.001$, $d = 0.89$; anger: $t(426) = 7.96$, $p < 0.001$, $d = 0.78$; joy: $t(441) = 4.26$, $p < 0.001$, $d = 0.41$]. Correspondingly, non-standardized facial expressions show a strong decrease in the correct emotion category compared to standardized facial expressions for some categories [disgust: $t(263) = 24.63$, $p < 0.001$, $d = 3.21$; surprise: $t(322) = 14.35$, $p < 0.001$, $d = 1.60$; sadness: $t(418) = 10.28$, $p < 0.001$, $d = 1.02$; neutral: $t(407) = 8.99$, $p < 0.001$, $d = 0.90$; anger: $t(426) = 8.03$, $p < 0.001$, $d = 0.79$; joy: $t(441) = 5.83$, $p < 0.001$, $d = 0.57$; fear: $t(317) = 3.79$, $p < 0.001$, $d = 0.43$]. Taken together, non-standardized compared to standardized facial expressions are perceived more often as neutral and less emotionally intense on average.

3.3. FER Systems: Drop Out

To evaluate the step of face detection, we report drop out rates separately for each FER tool in **Table 1**. Drop out for the standardized data is nearly non-existent, however, strong differences can be reported for the non-standardized data set. Azure returns no face detection for around 20% of the images.

[1] We also test different thresholds, but there is no reasonable performance improvement to be gained (see **Supplementary Figure 1**).

[2] Since this procedure leads to different samples for each algorithm, especially among the non-standardized data, we also compute the analysis for the subsample of non-standard images, which are recognized by all algorithms. The results are reported in **Supplementary Table 3**. Differences are minor, qualitatively the results remain the same.

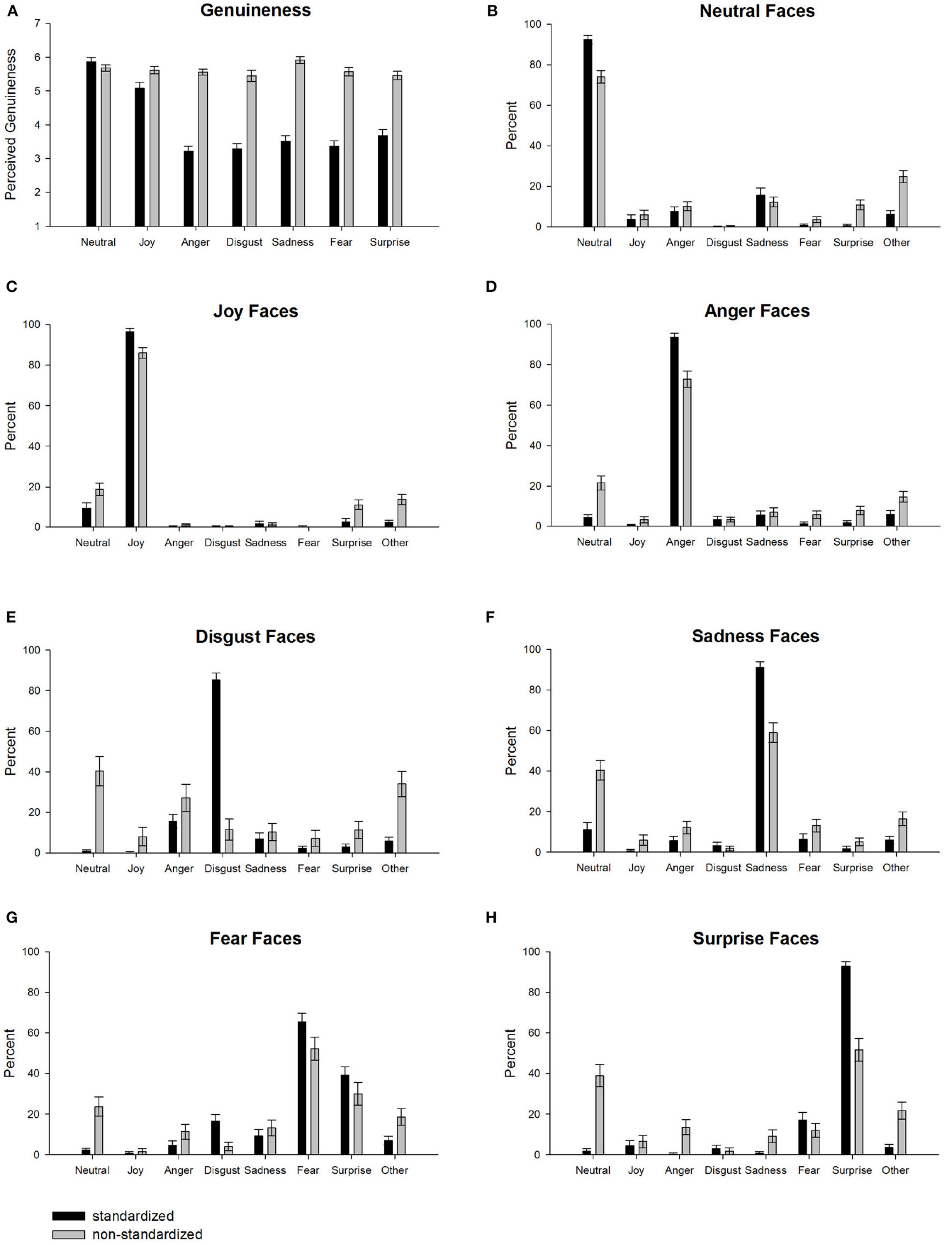

FIGURE 1 | Averaged human ratings separately for basic emotion categories for standardized (black bars) and non-standardized facial expressions (gray bars). **(A)** Depicts mean genuineness ratings ranging from 1 (very in-genuine) to 7 (very genuine). **(B–H)** Depict mean emotion ratings (percent) for **(B)** neutral, **(C)** joy, **(D)** anger, **(E)** disgust, **(F)** sadness, **(G)** fear, and **(H)** surprise expressions. Error bars are 95% confidence intervals.

TABLE 2 | Sensitivity, precision, and accuracy of Azure, Face++, and FaceReader separately for emotion categories.

	Azure				**Face++**				**FaceReader**			
	Stand.		**Non-Stand.**		**Stand.**		**Non-Stand.**		**Stand.**		**Non-Stand.**	
	Sens	**Prec**	**Sens**	**Prec**	**Sens**	**Prec**	**Sens**	**Prec**	**Sens**	**Prec**	**Sens**	**Prec**
Neutral	1.00	0.63	0.94	0.38	0.94	0.70	0.40	0.34	0.99	0.92	0.68	0.2
Joy	1.00	0.98	0.85	0.88	0.99	0.96	0.48	0.76	1.00	0.99	0.42	0.92
Anger	0.51	0.91	0.38	0.87	0.49	0.84	0.15	0.36	0.96	0.99	0.14	0.42
Disgust	0.85	0.98	0.10	0.50	0.89	0.77	0.16	0.17	0.97	0.99	0.15	0.17
Sadness	0.88	0.75	0.48	0.77	0.81	0.75	0.19	0.40	0.98	0.97	0.16	0.32
Fear	0.46	0.99	0.03	0.33	0.40	0.95	0.18	0.18	0.88	0.97	0.00	0.00
Surprise	0.98	0.73	0.56	0.43	0.97	0.71	0.66	0.20	0.98	0.93	0.34	0.33
Average	0.81	0.85	0.48	0.59	0.79	0.81	0.32	0.35	0.97	0.97	0.27	0.34
Accuracy	0.81		0.57		0.79		0.32		0.97		0.31	

Stand., standardized data; Non-Stand., non-standardized data; Sens., sensitivity; Prec., precision.

For FaceReader, the drop out is even higher with 74%[3]. This result partially confirms our expectations, as for Azure and FaceReader the drop out in the non-standardized data is much higher than among the standardized data. In contrast, Face++ shows superior face detection with nearly no drop out for the non-standardized data. See **Supplementary Table 1** for statistical comparison of the drop out rates.

3.4. FER Systems: Emotion Recognition

To descriptively compare classification performance, we report accuracies for each tool on each data set, along with category specific sensitivity and precision (**Table 2**). Details on the statistical comparisons can be found in **Supplementary Table 2**[4]. As expected, accuracy is better for all tools on the standardized data. FaceReader performs best, with 97% of the images classified correctly. The difference to both Azure and Face++ is significant ($p < 0.001$). Azure and Face++ perform similarly, $p = 0.148$, both put around 80% of the images in the correct category. For the non-standardized data, accuracy is much lower. Azure performs best, still correctly classifying 56% of the images. FaceReader and Face++ both correctly classify only about one third of the non-standardized images which constitutes a significant decrease of accuracy compared to Azure ($p < 0.001$).

Looking at the specific emotion categories and their performance indices, joy expressions are classified best. For the standardized data, both sensitivity and precision are or nearly are all 1. Also for the non-standardized data, the joy category is classified best. However, Azure is the only software with overall acceptable performance. In the standardized angry category, all tools show high precision, however Azure and Face++ lack in sensitivity. For the non-standardized angry category, only Azure's precision is acceptable. Face++, and FaceReader do not perform reliably. Performance on the other categories on the standardized data resembles each other: FaceReader clearly outperforms the other tools. In contrast, for the non-standardized facial expressions, Azure performs best, although the values are substantially decreased in comparison to standardized facial expressions.

To study confusion rates between categories, **Figure 2** depicts confusion matrices between the true labels and the highest rated emotion by each software. In the standardized data, all three tools show the pattern of classifying fearful expressions as surprise or sadness. The confusion between fear and surprise is expected, whereas the confusion of fear with sadness is new. Additionally, Azure and Face++ show a tendency to misclassify anger, sadness and fear as neutral. For FaceReader, this tendency is observable to a smaller extent. This reflects partially the expected tendency toward a neutral expression. In the non-standardized data set, all applications show a pronounced tendency toward the neutral category. Additionally, Face++ shows a trend toward surprise, sadness and fear. To a smaller extend, the misclassification to surprise and sadness is problematic in Azure and FaceReader alike.

3.5. Humans vs. FER: Comparison of Emotion Recognition

To directly compare all sources of emotion recognition, we calculate ROC curves and report them in **Figure 3** along with the corresponding AUCs. ROC curves for specific emotion categories are shown in **Supplementary Figure 2** and corresponding statistical comparisons are reported in **Supplementary Table 5**[5].

For the standardized facial expressions (see **Figure 3A**), humans, overall, recognize them significantly better than Azure, $p = 0.035$, and Face++, $p < 0.001$. However, FaceReader performs significantly better than humans on such facial expressions, $p < 0.001$. While the same pattern holds true for

[3] Twenty percent of the images have a quality that is too low for FaceReader to reliably detect emotions and we therefore exclude these from the analysis, in 54% no face is found by FaceReader.

[4] Since drop out rates differ strongly between the algorithms, especially among the naturalistic data, we also compute the analysis for the subset of naturalistic images, which are recognized by all algorithms. Differences are minor with corresponding patterns (see also **Supplementary Table 4**). Additionally, we report the shares of correctly identified images based on all images in **Supplementary Table 3**.

[5] AUC analysis for the subset of non-standardized data passed by all algorithms yields the same results.

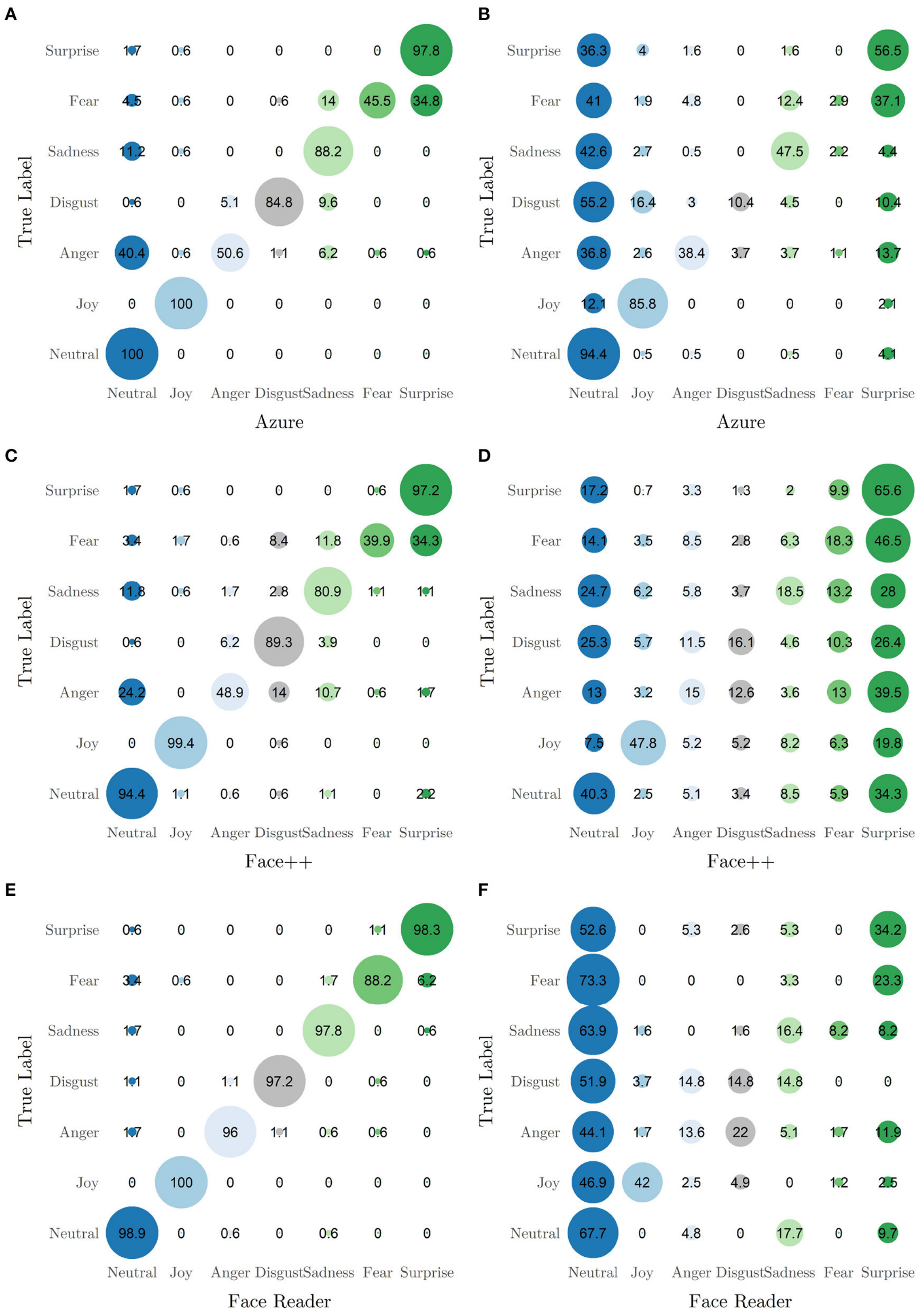

FIGURE 2 | Confusion matrices indicating classification performance on standardized (left panels) and non-standaridzed data (right panels): **(A)** standardized data by Azure, **(B)** non-standardized data by Azure, **(C)** standardized data by Face++, **(D)** non-standardized data by Face++, **(E)** standardized data by FaceReader and **(F)** non-standardized data by FaceReader. Numbers indicate percentages to the base of the true category. Reading example: From the standardized data Azure classifies 4.5% of the truly fearful expressions as neutral. The 45.5% of the fearful images are classified correctly.

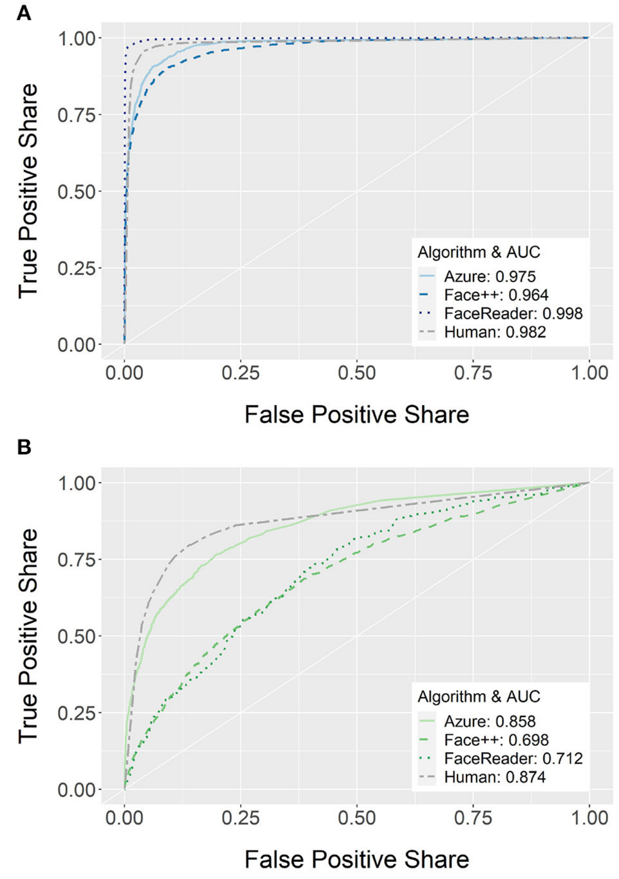

FIGURE 3 | Classification performance depicted as Receiver Operating Characteristic (ROC) curves and corresponding Area under the Curve (AUC) for overall emotion recognition performance for the three FER systems (Azure, Face++, and FaceReader) and human raters. Separately for **(A)** standardized facial expressions and **(B)** non-standardized facial expressions separately. The white diagonal line indicates classification performance by chance.

fear faces (Azure: $p = 0.003$; FaceReader: $p < 0.001$, Face++: $p < 0.001$), all algorithms perform significantly better than humans for neutral (Azure: $p < 0.001$; FaceReader: $p < 0.001$, Face++: $p < 0.001$), joy (Azure: $p = 0.023$; FaceReader: $p = 0.024$, Face++: $p = 0.027$), and surprise expressions (Azure: $p = 0.012$; FaceReader: $p = 0.012$, Face++: $p = 0.013$). Also, for standardized facial expressions of disgust, FaceReader, $p = 0.002$, and Face++, $p = 0.023$, perform better compared to humans while Azure is comparable to humans, $p = 0.450$. Regarding anger, FaceReader, and humans perform comparably, $p = 0.353$, and both outperform Azure and Face++, $p < 0.001$. Finally, FaceReader shows better classification of sad faces compared to Azure, $p = 0.078$, Face++, $p < 0.001$, and humans, $p = 0.021$.

For the non-standardized facial expressions (see **Figure 3B**), humans overall show similar performance to Azure, $p = 0.058$, and both perform better than FaceReader, $p < 0.001$, and Face++, $p < 0.001$. While this pattern is the same for joy (Azure: $p = 0.554$; FaceReader: $p < 0.001$, Face++: $p < 0.001$) and sadness (Azure: $p = 0.448$; FaceReader: $p < 0.001$, Face++: $p < 0.001$), humans outperform all algorithms in the detection of anger (Azure: $p < 0.001$; FaceReader: $p < 0.001$, Face++: $p < 0.001$) and fear facial expressions (Azure: $p < 0.001$; FaceReader: $p < 0.001$, Face++: $p < 0.001$). In contrast, Azure performs better than humans regarding neutral, $p < 0.001$, and disgust faces, $p < 0.001$, while FaceReader (neutral: $p < 0.001$; disgust: $p = 0.002$) and Face++ (neutral: $p = 0.001$; disgust: $p = 0.023$) show equal or worse performance compared to humans. Finally, Azure, $p = 0.006$, and Face++, $p < 0.001$, performs better than humans in the detection of non-standardized surprise facial expressions where FaceReader performs similar to humans, $p = 0.535$.

Taken together, for most emotion categories there is at least one FER system that performs equally well or better compared to humans. The only exceptions are non-standardized expressions of fear and anger, where humans clearly outperform all FER systems. FaceReader shows particularly good performance for standardized facial expressions and Azure performs better on non-standardized facial expressions.

4. DISCUSSION

In this paper, we evaluate and compare three widely used FER systems, namely Azure, Face++ and FaceReader, and human emotion recognition data. For the performance comparison, we use two different kinds of emotional facial expression data sets: First, a standardized data set comprised of lab generated images displaying intense, prototypical facial expressions of emotions under very good recording conditions (i.e., lighting, camera angle). Second, we test a non-standardized set, which contains facial expressions from movie scenes depicting emotional faces as an approximation for more naturalistic, spontaneous facial expressions (Dhall et al., 2018). The non-standardized facial expressions constitute an especially difficult test case, since it contains large variation in the expressions itself, the surrounding circumstances and the displayed person's characteristics.

Overall, the three classifiers as well as humans perform well on standardized facial expressions. However, we observe large variation and a general decrease in performance for the non-standardized data, in line with previous work (Yitzhak et al., 2017; Dupré et al., 2020). Although emotion recognition performance is generally lower for such facial expressions, FER tools perform similarly or better than humans for most emotion categories of non-standardized (except for anger and fear) and standardized facial expressions. Facial expressions of joy are detected best among the emotion categories in both standardized and non-standardized facial expressions, which also replicates existing findings (Stöckli et al., 2018; Höfling et al., 2021). However, FER performance varies strongly between systems and emotion categories. Depending on the data and on which emotions one aims at classifying, one algorithm might be better suited than the other: Face++ shows almost no drop out in face detection even under the non-standardized condition, FaceReader shows excellent performance for standardized prototypical facial expressions and outperforms humans, and Azure shows superior overall performance on non-standardized facial expressions among all FER tools.

4.1. Implications for Application

From our data, we can derive three broad implications. First, all FER tools perform much better on the standardized, prototypical data, than on the non-standardized, more naturalistic data. This might indicate over fitting on standardized data. Second, FER systems and human coders can detect some emotion categories better than others, resulting in asymmetries in classification performance between emotion categories. This indicates that the detection of certain emotional facial expressions is generally more error prone than others. Third, we can identify performance problems that are specific to FER tools.

First, as expected, all FER systems perform better on the standardized compared to non-standardized and more naturalistic facial expressions. This is the case for both face detection and emotion classification. Within the standardized data, face detection is near to prefect for all systems and shows almost no drop out based on face detection failures. Regarding the emotion classification, FaceReader outperforms Face++, Azure, and even human coders. Within the non-standardized data, face detection is observed to be problematic for Azure and FaceReader. Judging the classification performance on the non-standardized data set, all three classifiers show a large overall decrease in accuracy, whereby Azure is most accurate compared to Face++ and FaceReader. In particular, all FER systems, and less pronounced in humans, show a misclassification of emotional facial expressions as neutral facial expressions for the non-standardized data. This is an important observation not shown by Dupré et al. (2020), since they have not reported confusions with the neutral category. We suspect the neutral classification due to the expressions in acted films being less intense compared to standardized, lab generated data. Hence, the vastly better performance on standardized, prototypical facial expressions which were generated under controlled conditions may indicate limitations of FER systems to more naturalistic and more subtle emotional facial expressions.

Second, we observe that FER and human performance reflect varying underlying difficulties in the classification of different emotions. In other words, certain emotions are harder to detect than others, for example because of more subtle expressions or less distinct patterns. This evolves from shared classification error patterns between the three algorithms which corresponds to prior research on other algorithms and human recognition performance. In our data, joy is recognized best and fear is among the most difficult to classify which is in line with prior FER (Stöckli et al., 2018; Skiendziel et al., 2019; Dupré et al., 2020) and human emotion recognition research (Nummenmaa and Calvo, 2015; Calvo and Nummenmaa, 2016). Anger has been found to be difficult to classify in some studies (Stöckli et al., 2018; Dupré et al., 2020), but not in others (Skiendziel et al., 2019). With regards to our findings, angry faces can be classified with low sensitivity, but high precision. Sadness and disgust are reported to be difficult to detect in other studies (Lewinski et al., 2014; Skiendziel et al., 2019). Fear is regularly misclassified as surprise, as found in other studies with FER (Stöckli et al., 2018; Skiendziel et al., 2019) and humans alike (Palermo and Coltheart, 2004; Calvo and Lundqvist, 2008; Tottenham et al., 2009; Calvo et al., 2018). For the non-standardized data, FER performance on disgust is among the lowest for all classifiers which corresponds to human recognition data in the present study. In line with previous research, the pronounced performance drop for many non-standardized images compared to standardized emotion categories (Yitzhak et al., 2017; Dupré et al., 2020) might indicate that the FER systems are not trained on detecting the full variability of emotional facial expressions. Importantly, these results reflect that FER simulates human perception and also shows similar classification errors.

Third, we make a series of observations, that specific FER systems misclassify certain emotion categories, which is not shared by human coders. In our data, fear is also misclassified as sadness by Azure in standardized and non-standardized facial expressions. For the non-standardized data, we also report a general tendency to misclassify surprise expressions, that is not evident in other studies. Especially the misclassification toward surprise in the non-standardized data might be explained by an open mouth due to speaking in movies, for which the applications do not account. In addition, Face++ misclassifies any emotion in the non-standardized data as fear and to a lesser extend as sadness. Regarding FaceReader, we observe a pronounced misclassification of naturalistic facial expressions as neutral. These findings indicate misclassification pattern specific for the three investigated FER systems which possibly reflect differences in their machine-learning architecture, training material and validation procedure.

4.2. Limitations and Outlook

This study has some limitations. Most obviously, we compare three representative and not all available software systems on the market. While we choose software that is widely used, other algorithms will need to be examined in a similar fashion. For example, Beringer et al. (2019) find that FACET shows a certain resilience to changes in lighting and camera angle on lab generated data. We could not see in this study if this resilience transfers to an even harder task.

To approximate more naturalistic facial expressions, we utilize images from movie stills as the non-standardized data set. While this is convenient and emotional expressions are already classified and evaluated, these images are of course also posed by actors. However, good acting is generally thought of as a realistic portrayal of true affect. Our ratings of genuineness appear to support our distinction of standardized and non-standardized facial expressions. In addition, our human recognition data provide further validation of emotion categorization of this particular facial expression inventory. Even though acted portrays of emotional facial expressions differ between prototypical inventories and movies, which is in line with previous research (Carroll and Russell, 1997), these acted facial expressions are only approximations for true emotional expressions. Moreover, movie stimuli may be rated as more authentic compared to the prototypical data, due to many reasons like the variation in head orientations, lighting, backgrounds, and familiarity with the actors or movie plot. Hence, facial expressions of true emotion require an additional criterion of emotional responding like ratings of currently elicited emotions.

Furthermore, we argue that FER would be most useful in categorizing spontaneous and naturalistic facial expressions in different contexts. The SFEW data set serves as an approximation for this. However, it is unclear whether the displayed emotional facial expressions are grounded in emotional processing or just simulated. For example, Höfling et al. (2020) elicited spontaneous emotional responses by presenting emotional scenes to their participants and found FER detects changes in facial expressions only for pleasant emotional material. Hence, more data sets are needed to test different naturalistic settings and foster development in this area.

Beyond the bias in FER toward prototypical expressions under good condition, there are other sources of systemic error that we did not address, such as biases against race, gender, age, or culture (Zou and Schiebinger, 2018; Aggarwal et al., 2019; Wellner and Rothman, 2020). For example, it has been shown that automated facial analysis to classify gender works less well for people with a darker skin tone (Buolamwini and Gebru, 2018). Many training data sets are concentrated on Northern America and Europe (Shankar et al., 2017), which partially causes the biases and at the same time makes it difficult to detect them. Future research should take these variables into account to evaluate measurement fairness independent of specific person characteristics.

5. CONCLUSION

This study contributes to the literature by comparing the accuracy of three state-of-the-art FER systems to classify emotional facial expressions (i.e., FaceReader, Azure, Face++). We show that all systems and human coders perform well for standardized, prototypical facial expressions. When challenged with non-standardized images, used to approximate more naturalistic expressions collected outside of the lab, performance of all systems as well as human coders drops considerably. Reasons for this are substantial drop out rates and a decrease in classification accuracy specific to FER systems and emotion categories. With only a short history, FER is already a valid research tool for intense and prototypical emotional facial expressions. However, limitations are apparent in the detection of non-standardized facial expressions as they may be displayed in more naturalistic scenarios. Hence, further research is urgently needed to increase the potential of FER as a research tool for the classification of non-prototypical and more subtle facial expressions. While the technology is, thus, a promising candidate to assess emotional facial expressions on a non-contact basis, researchers are advised to interpret data from non-prototypical expressions in non-restrictive settings (e.g., strong head movement) carefully.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

TK conceived and designed the study, contributed Face++ and Azure data, conducted the analysis and interpretation of the results, and also drafted the work. TH contributed FaceReader data and collected data from human raters. TH and GA contributed to the interpretation of the results and writing of the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This publication was funded by the open access publication fund of the University of Konstanz.

ACKNOWLEDGMENTS

We would like to express our gratitude to those, who commented on earlier version of this paper, especially Susumu Shikano, Karsten Donnay, Sandra Morgenstern, Kiela Crabtree, Sarah Shugars, and the participants of the Image Processing for Political Research workshop at APSA 2020. We also would like to thank Timo Kienzler for technical support. We thank the two reviewers for their constructive suggestions, which further improved the paper.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.627561/full#supplementary-material

REFERENCES

Abdullah, S., Murnane, E. L., Costa, J. M. R., and Choudhury, T. (2015). "Collective smile: measuring societal happiness from geolocated images," in *CSCW '15: Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, Tallinn, Estonia. 361–374. doi: 10.1145/2675133.2675186

Aggarwal, A., Lohia, P., Nagar, S., Dey, K., and Saha, D. (2019). "Black box fairness testing of machine learning models," in *ESEC/FSE 2019: Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, Tallinn, Estonia. 625–635. doi: 10.1145/3338906.3338937

Arriaga, O., Valdenegro-Toro, M., and Plöger, P. (2017). Real-time convolutional neural networks for emotion and gender classification. *CoRR, abs/1710.07557*.

Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., and Pollak, S. D. (2019). Emotional expressions reconsidered: challenges to inferring emotion from human facial movements. *Psychol. Sci. Publ. Interest* 20, 1–68. doi: 10.1177/1529100619832930

Bartkiene, E., Steibliene, V., Adomaitiene, V., Juodeikiene, G., Cernauskas, D., Lele, V., et al. (2019). Factors affecting consumer food preferences: food taste and depression-based evoked emotional expressions with the use of face reading technology. *BioMed Res. Int*. 2019:2097415. doi: 10.1155/2019/2097415

Bartlett, M., Hager, J. C., Ekman, P., and Sejnowski, T. J. (1999). Measuring facial expressions by computer image analysis. *Psychophysiology* 36, 253–263. doi: 10.1017/S0048577299971664

Beringer, M., Spohn, F., Hildebrandt, A., Wacker, J., and Recio, G. (2019). Reliability and validity of machine vision for the assessment of facial expressions. *Cogn. Syst. Res*. 56, 119–132. doi: 10.1016/j.cogsys.2019.03.009

Boxell, L. (2018). Slanted Images: Measuring Nonverbal Media Bias. *Munich Personal RePEc Archive Paper No. 89047*. Munich: Ludwig Maximilian University of Munich.

Brader, T. (2005). Striking a responsive chord: how political ads motivate and persuade voters by appealing to emotions. *Am. J. Polit. Sci*. 49, 388–405. doi: 10.1111/j.0092-5853.2005.00130.x

Buolamwini, J., and Gebru, T. (2018). Gender shades: intersectional accuracy disparities in commercial gender classification. *Proc. Mach. Learn. Res*. 81, 1–15.

Calvo, M., and Lundqvist, D. (2008). Facial expressions of emotion (kdef): identification under different display-duration conditions. *Behav. Res. Methods* 40, 109–115. doi: 10.3758/BRM.40.1.109

Calvo, M. G., Fernández-Martín, A., Recio, G., and Lundqvist, D. (2018). Human observers and automated assessment of dynamic emotional facial expressions: Kdef-dyn database validation. *Front. Psychol*. 9:2052. doi: 10.3389/fpsyg.2018.02052

Calvo, M. G., and Nummenmaa, L. (2016). Perceptual and affective mechanisms in facial expression recognition: an integrative review. *Cogn. Emot*. 30, 1081–1106. doi: 10.1080/02699931.2015.1049124

Carroll, J. M., and Russell, J. A. (1997). Facial expressions in hollywood's portrayal of emotion. *J. Pers. Soc. Psychol*. 72, 164–176. doi: 10.1037/0022-3514.72.1.164

Chang, W. (2014). *extrafont: Tools for Using Fonts*. R package version 0.17.

Clore, G. L., Gasper, K., and Garvin, E. (2001). "Affect as information," in *Handbook of Affect and Social Cognition*, ed J. Forgas (Mahwah, New Yersey:Psychology Press), 121–144.

Cohn, J. F., Ambadar, Z., and Ekman, P. (2007). "Observer-based measurement of facial expression with the facial action coding system," in *Handbook of Emotion Elicitation and Assessment*, eds J. Coan and J. Allen (Oxford:Oxford University Press), 222–238.

Dhall, A., Goecke, R., Lucey, S., and Gedeon, T. (2012). Collecting large, richly annotated facial-expression database from movies. *IEEE MultiMed*. 19, 34–41. doi: 10.1109/MMUL.2012.26

Dhall, A., Kaur, A., Goecke, R., and Gedeon, T. (2018). "Emotiw 2018: audio-video, student engagement and group-level affect prediction," in *ICMI' 18* Boulder, CO, 653–656. doi: 10.1145/3242969.3264993

Dowle, M., and Srinivasan, A. (2020). *data.table: Extension of 'data.frame'*. R Package Version 1.13.2.

Dupré, D., Krumhuber, E. G., Küster, D., and McKeown, G. J. (2020). A performance comparison of eight commercially available automatic classifiers for facial affect recognition. *PLoS ONE* 15:e231968. doi: 10.1371/journal.pone.0231968

Ekman, P., and Rosenberg, E. L. (1997). *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*. New York, NY:Oxford University Press.

Fraser, K., Ma, I., Teteris, E., Baxter, H., Wright, B., and McLaughlin, K. (2012). Emotion, cognitive load and learning outcomes during simulation training. *Med. Educ*. 46, 1055–1026. doi: 10.1111/j.1365-2923.2012.04355.x

Goodfellow, I. J., Erhan, D., Carrier, P., Courville, A., Mirza, M., Hamner, B., et al. (2015). Challenges in representation learning: a report on three machine learning contests. *Neural Netw*. 64, 59–63. doi: 10.1016/j.neunet.2014.09.005

Haim, M., and Jungblut, M. (2020). Politicians' self-depiction and their news portrayal: Evidence from 28 countries using visual computational analysis. *Polit. Commun*. doi: 10.1080/10584609.2020.1753869

Henry, L., and Wickham, H. (2020). *purrr: Functional Programming Tools*. R package version 0.3.4.

Höfling, T. T. A., Alpers G. W., Gerdes, A. B. M., and Föhl, U. (2021). Automatic facial coding versus electromyography of mimicked, passive, and inhibited facial response to emotional faces. *Cogn. Emot*. doi: 10.1080/02699931.2021.1902786

Höfling, T. T. A., Gerdes, A. B. M., Föhl, U., and Alpers, G. W. (2020). Read my face: automatic facial coding versus psychophysiological indicators of emotional valence and arousal. *Front. Psychol*. doi: 10.3389/fpsyg.2020.01388

Jin, B., Qu, Y., Zhang, L., and Gao, Z. (2020). Diagnosing Parkinson disease through facial expression recognition: video analysis. *J. Med. Intern. Res*. 22:e18697. doi: 10.2196/18697

Keltner, D., and Cordaro, D. T. (2017). "Understanding multimodal emotional expressions," in *The Science of Facial Expression*, eds J. Russel and J. Fernandez Dols (New York, NY:Oxford University Press), 57–76. doi: 10.1093/acprof:oso/9780190613501.003.0004

Kuhn, M. (2020). *caret: Classification and Regression Training*. R package version 6.0-86.

Kulke, L., Feyerabend, D., and Schacht, A. (2020). A comparison of the affectiva imotions facial expression analysis software with EMG for identifying facial expressions of emotion. *Front. Psychol*. 11:329. doi: 10.3389/fpsyg.2020.00329

Langer, O., Dotsch, R., Bijlstra, G., Wigboldus, D. H. J., Hawk, S. T., and van Knippenberg, A. (2010). Presentation and validation of the radboud faces database. *Cogn. Emot*. 24, 1377–1388. doi: 10.1080/02699930903485076

Lerner, J. S., and Keltner, D. (2000). Beyondvvalence: Toward a model of emotion-specific influences on judgement and choice. *Cogn. Emot*. 14, 473–493. doi: 10.1080/026999300402763

Lewinski, P. (2015). Automated facial coding software outperforms people in recognizing neutral faces as neutral from standardized datasets. *Front. Psychol*. 6:1386. doi: 10.3389/fpsyg.2015.01386

Lewinski, P., den Uyl, T., and Butler, C. (2014). Automated facial coding: validation of basic emotions and facs aus in facereader. *J. Neurosci. Psychol. Econ*. 7, 227–236. doi: 10.1037/npe0000028

Lundqvist, D., Flykt, A., and Öhman, A. (1998). *The Karolinska Directed Emotional Faces - KDEF*. CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet. ISBN 91-630-7164-9. doi: 10.1037/t27732-000

Marcus, G. E. (2000). Emotions in politics. *Annu. Rev. Polit. Sci*. Chicago. 3, 221–250. doi: 10.1146/annurev.polisci.3.1.221

Marcus, G. E., Neuman, W. R., and MacKuen, M. (2000). *Affective Intelligence and Political Judgement*. The University of Chicago Press.

Mavadati, M. S., Mahoor, M. H., Bartlett, K., Trinh, P., and Cohn, J. F. (2013). Disfa: A spontaneous facial action intensity database. *IEEE Trans. Affect. Comput*. 4, 151–160. doi: 10.1109/T-AFFC.2013.4

Meffert, M. F., Chung, S., Joiner, A. J., Waks, L., and Garst, J. (2006). The effects of negativity and motivated information processing during a political campaign. *J. Commun*. 56, 27–51. doi: 10.1111/j.1460-2466.2006.00003.x

Neuwirth, E. (2014). *RColorBrewer: ColorBrewer Palettes*. Long Beach, CA:R package version 1.1-2.

Nummenmaa, L., and Calvo, M. G. (2015). Dissociation between recognition and detection advantage for facial expressions: a meta-analysis. *Emotion* 15, 243–256. doi: 10.1037/emo0000042

Olszanowski, M., Pochwatko, G., Kuklinski, K., Scibor-Rylski, M., Lewinski, P., and Ohme, R. (2015). Warsaw set of emotional facial expression pictures: a validation study of facial display photographs. *Front. Psychol*. 5:1516. doi: 10.3389/fpsyg.2014.01516

Ooms, J. (2014). The jsonlite package: a practical and consistent mapping between JSON data and R objects. *arXIv [Preprint] arXiv*:1403.2805 *[stat.CO]*.

Palermo, R., and Coltheart, M. (2004). Photographs of facial expression: accuracy, response times, and ratings of intensity. *Behav. Res. Methods Instrum. Comput*. 36, 634–638. doi: 10.3758/BF03206544

Pedersen, T. (2020). *patchwork: The Composer of Plots*. R package version 1.1.0.

Peng, Y. (2018). Same candidates, different faces: uncovering media bias in visual portrayals of presidential candidates with computer vision. *J. Commun*. 65, 920–941. doi: 10.1093/joc/jqy041

Pittig, A., Schulz, A. R., Craske, M. G., and Alpers, G. W. (2014). Acquisition of behavioral avoidance: task-irrelevant conditioned stimuli trigger costly decisions. *J. Abnorm. Psychol*. 123, 314–329. doi: 10.1037/a0036136

Quinn, M. A., Sivesind, G., and Reis, G. (2017). *Real-time Emotion Recognition From Facial Expressions*. Available online at: http://cs229.stanford.edu/proj2017/final-reports/5243420.pdf

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J., et al. (2011). proc: an open-source package for r and s+ to analyze and compare roc curves. *BMC Bioinformatics* 12:77. doi: 10.1186/1471-2105-12-77

Sachs, M. C. (2017). plotROC: A tool for plotting roc curves. *J. Stat. Softw. Code Snipp*. 79, 1–19. doi: 10.18637/jss.v079.c02

Sato, W., Hyniewska, S., Minemoto, K., and Yoshikawa, S. (2019). Facial expressions of basic emotions in Japanese laypeople. *Front. Psychol*. 10:259. doi: 10.3389/fpsyg.2019.00259

Scherer, K., and Ellgring, H. (2007). Multimodal expression of emotion: Affect programs or componential appraisal patterns? *Emotion* 7, 158–171. doi: 10.1037/1528-3542.7.1.158

Shankar, S., Halpern, Y., Breck, E., Atwood, J., Wilson, J., and Sculley, D. (2017). "No classification without representation: assessing geodiversity issues in open data sets for the developing world," in *31st Conference on Neural Information Processing Systems (NIPS 2017)*.

Skiendziel, T., Rösch, A. G., and Schultheiss, O. C. (2019). Assessing the convergent validity between the automated emotion recognition software noldus facereader 7 and facial action coding system scoring. *PLoS ONE*. 14:e0223905. doi: 10.1371/journal.pone.0223905

Slovic, P., Finucane, M., Peters, E., and MacGregor, D. G. (2007). The affect heuristic. *Eur. J. Oper. Res*. 177, 1333–1352. doi: 10.1016/j.ejor.2005.04.006

Soroka, S., and McAdams, S. (2015). News, politics and negativity. *Polit. Commun*. 32, 1–22. doi: 10.1080/10584609.2014.881942

Stöckli, S., Schulte-Mecklenbeck, M., Borer, S., and Samson, A. C. (2018). Facial expression analysis with affdex and facet: a validation study. *Behav. Res. Methods* 50, 1446–1460. doi: 10.3758/s13428-017-0996-1

Sullivan, D. G., and Masters, R. D. (1988). "happy warriors": Leaders' facial displays, viewers' emotions, and political support. *Am. J. Polit. Sci*. 32, 345–368. doi: 10.2307/2111127

Teixeira, T., Picard, R., and el Kaliouby, R. (2014). Why, when, and how much to entertain consumers in advertisements? A web-based facial tracking field study. *Market. Sci*. 33, 809–827. doi: 10.1287/mksc.2014.0854

Tian, Y.-l, Kanade, T., and Cohn, J. F. (2001). Recognizing action units for facial expression analysis. *IEEE Trans. Pattern Anal. Mach. Intell*. 23, 97–115. doi: 10.1109/34.908962

Tottenham, N., Tanaka, J., Leon, A., McCarry, T., Nurse, M., Hare, T., et al. (2009). The nimstim set of facial expressions: judgments from untrained research participants. *Psychiatry Res*. 168, 242–249. doi: 10.1016/j.psychres.2008.05.006

Van der Schalk, J., Hawk, S., Fischer, A., and Doosja, B. (2011). Moving faces, looking places: validation of the Amsterdam dynamic facial expression set (ADFES). *Emotion* 11, 907–920. doi: 10.1037/a0023853

Wellner, G., and Rothman, T. (2020). Feminist AI: can we expect our ai systems to become feminist? *Philos. Technol*. 33, 191–205. doi: 10.1007/s13347-019-00352-z

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. New York, NY: Springer-Verlag. doi: 10.1007/978-3-319-24277-4_9

Wickham, H. (2019). *stringr: Simple, Consistent Wrappers for Common String Operations*. R package version 1.4.0.

Wickham, H. (2020). *httr: Tools for Working With URLs and HTTP*. R package version 1.4.2.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., et al. (2019). Welcome to the tidyverse. *J. Open Source Softw*. 4:1686. doi: 10.21105/joss.01686

Wickham, H., François, R., Lionel, H., and Müller, K. (2020). *dplyr: A Grammar of Data Manipulation*. R package version 1.0.2.

Wu, F., Lin, S., Cao, X., Zhong, H., and Zhang, J. (2019). "Head design and optimization of an emotionally interactive robot for the treatment of autism," in *CACRE2019: Proceedings of the 2019 4th International Conference on Automation, Control and Robotics Engineering*, 1–10. doi: 10.1145/3351917.3351992

Yazdavar, A., Mahdavinejad, M., Bajaj, G., Romine, W., Sheth, A., Monadjemi, A., et al. (2020). Multimodal mental health analysis in social media. *PLoS ONE* 15:e0226248. doi: 10.1371/journal.pone.0226248

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

Zhang, X., Yin, L., Cohn, J. F., Canavan, S., Reale, M., Horowitz, A., et al. (2014). BP4d-spontaneous: a high-resolution spontaneous 3D dynamic facial expression database. *Image Vis. Comput*. 32, 692–706. doi: 10.1016/j.imavis.2014.06.002

Zou, J., and Schiebinger, L. (2018). Ai can be sexist and racist-it's time to make it fair. *Nature* 559, 324–326. doi: 10.1038/d41586-018-05707-8

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.