Automated facial coding: validation of basic emotions and FACS AUs in FaceReader

Lewinski, P.; den Uyl, T.M.; Butler, C.

DOI
10.1037/npe0000028

Publication date
2014

Document Version
Author accepted manuscript

Published in
Journal of Neuroscience, Psychology, and Economics

Journal of Neuroscience, Psychology, and Economics. doi: 10.1037/npe0000028

Automated facial coding: Validation of basic emotions and FACS AUs in FaceReader

Lewinski, P., den Uyl, T. M., Butler, C.

Date submitted: 15 July 2014
Date resubmitted: 17 Oct 2014
Date accepted: 29 Oct 2014

Correction to Lewinski, den Uyl, and Butler (2014)

In the article “Automated Facial Coding: Validation of Basic Emotions and FACS AUs in FaceReader” by Peter Lewinski, Tim M. den Uyl, and Crystal Butler (Journal of Neuroscience, Psychology, and Economics, 2014, Vol. 7, No. 4, pp. 227–236. http://dx.doi.org/10.1037/npe0000028), after recomputing the results, the FaceReader FACS performance is actually higher than what was originally reported.

The average ADFES agreement index increased from 0.66 to 0.68, and the average WSEFEP index increased from 0.69 to 0.70. This means that FaceReader reached a FACS index of agreement of 0.69 on average in both datasets. An error was discovered in calculations while working on another project. It appeared that the annotations of the AU’s intensity (coded as “Not Active”, A, B, C, D or E) were extracted from both data sets (WSEFEP and ADFES) in the wrong way. Specifically, all images that were annotated with “A” intensity were counted as “Not Active.” Due to this error, a lower number of AUs appeared to be present in both data sets. This means that the numbers reported in the original article were incorrect. The other performance matrix changed, too. Therefore, a changelog for the readers appeared in:

Houser, D., & Weber, B. (2015). Correction to Lewinski, den Uyl, and Butler (2014): Automated Facial Coding: Validation of Basic Emotions and FACS AUs in FaceReader. *Journal of Neuroscience, Psychology, and Economics, 8*(1), 58-59. doi: 10.1037/npe0000033

The below version of the paper has those changes included.

Automated Facial Coding: Validation of Basic Emotions and FACS AUs in FaceReader

Peter Lewinski

The Amsterdam School of Communication Research ASCoR, Department of Communication Science, University of Amsterdam

Tim M. den Uyl

Vicarious Perception Technologies B.V., Amsterdam

Crystal Butler

Department of Computer Science, Graduate School of Arts and Sciences, New York University

The research leading to these results has received funding from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme FP7/2007-2013/ under REA grant agreement 290255. The findings have been presented as an abstract/poster at NEFCA Etmaal van de Communicatiewetenschap (24 Hours of Communication Sciences), Wageningen, the Netherlands, February 2014.

Peter Lewinski works - as a Marie Curie Research Fellow - for Vicarious Perception Technologies B.V., Amsterdam– an artificial intelligence company that develops FaceReader software for Noldus Information Technologies B.V. He is also a PhD Candidate in ASCoR. Tim den Uyl is a Machine Vision Engineer and Crystal Butler was an interne at the same company.

Correspondence concerning this article should be addressed to Peter Lewinski, The Amsterdam School of Communication Research ASCoR, Department of Communication Science, University of Amsterdam, Amsterdam 1018 WV., e-mail: p.lewinski@uva.nl or peter.lewinski@gmail.com

Abstract

In this paper, we validated automated facial coding (AFC) software – FaceReader (Noldus, 2014) - on two publicly available and objective datasets of human expressions of basic emotions. We present the matching scores (accuracy) for recognition of facial expressions and the Facial Action Coding System (FACS) index of agreement. In 2005, matching scores of 89% were reported for FaceReader. However, previous research used a version of FaceReader that implemented older algorithms (version 1.0) and did not contain FACS classifiers. In this study, we tested the newest version (6.0). FaceReader recognized 88% of the target emotional labels in the *Warsaw Set of Emotional Facial Expression Pictures* (WSEFEP) and *Amsterdam Dynamic Facial Expression Set* (ADFES). The software reached a FACS index of agreement of 0.69 on average in both datasets. The results of this validation test are meaningful only in relation to human performance rates for both basic emotion recognition and FACS coding. The human emotions recognition for the two datasets was 85% therefore; FaceReader is as good at recognizing emotions as humans. In order to receive FACS certification, a human coder must reach an agreement of 0.70 with the master coding of the final test. Even though FaceReader did not attain this score, action units (AUs) 1, 2, 4, 5, 6, 9, 12, 15 and 25 might be used with high accuracy. We believe that FaceReader has proven to be a reliable indicator of basic emotions in the past decade and has a potential to become similarly robust with FACS.

Keywords: FaceReader; facial expressions; action units; FACS; basic emotions

Automated Facial Coding: Validation of Basic Emotions and FACS AUs in FaceReader

Manual facial coding – though precise – is a labor-intensive task. Due to recent advance, automated facial coding (AFC) is becoming more reliable and ubiquitous (Valstar, Mehu, Jiang, Pantic & Scherer, 2012). Software for AFC either directly FACS-codes facial movements or categorizes them into emotions or cognitive states. The FACS manual is a +700-page guide describing procedures for the manual, objective codification of facial behavior (Ekman & Friesen, 1978; Ekman, Friesen & Hager, 2002). The AFC software, along with other tools such as electrodermal response registration (for a review see Lajante, Droulers, Dondaine, & Amarantini, 2012); heart rate registration (e.g. Micu & Plummer, 2010); EEG (e.g. Cook, Warren, Pajot, Schairer & Leuchter, 2011) or eye-tracking (e.g. Ramsøy, Friis-Olivarius, Jacobsen, Jensen & Skov, 2012) is an accessible alternative for many researchers in consumer neuroscience.

The focus of this paper is to show the performance of FaceReader (Noldus, 2014) in the last tenant of validity and reliability of AFC software: recognition studies (e.g. Russell, 1994; Nelson and Russell, 2013). Analogous to human recognition studies, we provide one aggregated number that can be quoted in further research with FaceReader as an objective *accuracy* score (see in *Method* for definition). Every researcher using FaceReader invariably asks the questions – "how well does the software measure what it is supposed to measure?" In the current paper, we put forward the answer in the results sections.

FaceReader

FaceReader (Noldus, 2014) is the first commercially available AFC software still in existence. The software first finds a person's face, and then creates a 3D Active Appearance Model (AAM) (Cootes & Taylor, 2004) of a face. In the last stage, the AAM is used to compute scores of probability and intensity of facial expressions on a continuous scale from 0 to 1. For an algorithmic description of FaceReader, see van Kuilenburg, Wiering and den Uyl (2005).

FaceReader classifies people's emotions into discrete categories of basic emotions (Ekman, Sorenson, & Friesen, 1969; Ekman & Cordano, 2011). In previous research, accuracy (i.e. matching scores) of 89% (den Uyl & van Kuilenburg, 2005; van Kuilenburg, et al., 2005) was reported. In a standard FaceReader experiment, the facial data is gathered through an external remote webcam or one embedded into an existing eyetracker (e.g. Tobii or SMI). In addition, FaceReader Online can be integrated with Qualtrics and crowdsourcing platforms while analyzing facial data in a secure cloud, using people's own webcams[1]. The algorithms used in FaceReader Online are always up to date with the latest available version of FaceReader.

In the past few years, there has been an increase in academic research with FaceReader. FaceReader has proven useful in variety of contexts, such as emotion science (Chentsova-Dutton & Tsai, 2010), educational research (e.g. Terzis, Moridis & Economides, 2012; 2013; Chiu, Chou, Wu & Liaw, 2014) consumer behavior (e.g. Garcia-Burgos & Zamora, 2013; de Wijk, He, Mensink, Verhoeven & de Graaf, 2014; Danner, Sidorkina, Joechl & Duerrschmid, 2014), user-experience (e.g. Goldberg, 2014) and in marketing research (e.g. Lewinski, Fransen & Tan, 2014).

In previous research (van Kuilenburg, et al., 2005), matching scores were reported for FaceReader but the training and test dataset came from the same database, possibly inflating the recognition scores. The method used for testing the performance, leave-one-out cross-validation, was determined to be the best choice in 2005, as the authors had only a single database of annotated facial expressions at their disposal. In the current paper, we did not have this limitation anymore, as we had two annotated databases (ADFES and WSEFEP) available for testing that were not included in the FaceReader 6.0 training dataset. In addition, previous versions of FaceReader had older versions (1.0) of algorithms and did not contain FACS classifiers.

[1] For FaceReader Online see - www.facereader-online.com

Since version 1.0 was made public 10 years ago, versions 2.0, 3.0, 4.0, 5.0 have been made commercially available but were never re-validated. For this reason, we decided to test the newest version (6.0) in this study. In comparison to earlier versions, the main improvements in version 6.0, as relevant to academic research, are: (a) increased classification speed through code optimization, (b) increased robustness due to switching from 2D to 3D face modeling, c) improved accuracy based on an upgrade to 510 key identification points on the face instead of 55 key points. Version 6.0 can also analyze arousal and valence based on Russell's Circumplex Model of Affect (1980) as well as contempt, but the WSEFEP and ADFES datasets did not provide such labels and therefore we could not test these three new categories.

Validity and Reliability of AFC

We believe that there are some common misconceptions as to how to validate AFC software. We argue that the validity and reliability of AFC is based on (a) principles of computer algorithms, (b) psychological theories and (c) recognition studies. In this paper, we provide explicit evidence for the last point but we briefly explain the first two for the sake of clarity.

Computer algorithms code facial expressions according to a set of fixed rules that are invariably applied to each expression. The algorithms always follow this specific coding protocol, do not have personal biases (e.g. about gender, culture or age) and do not get tired. It is very unlikely that human coders will ever be able to reach the level of objectivity of AFC. The artificial intelligence that stands behind AFC simply does not have human free will and the unconstrained possibilities of making subjective choices. Consider that, as an example, that running AFC software twice on the same dataset will always give the same results.

Furthermore, as is the case with FaceReader, AFC is based on psychological theories and therefore the algorithms build upon preexisting knowledge. The FaceReader software estimates human affective states using methods determined by theories that are supported by thousands of

scholarly articles, and does not aim to make theoretical interpretations of its own. Prominently, FaceReader is based on more than 40 years of research on basic emotions, starting with the seminal paper by Ekman et al. (1969).

Design and Procedure

In this paper, across Validation 1 and 2, we validated FaceReader (Noldus, 2014) on two publicly available and objective datasets of human facial expressions of emotions. We used the *Warsaw Set of Emotional Facial Expression Pictures* (WSEFEP) (Olszanowski, Pochwatko, Kukliński, Ścibor-Rylski, & Ohme, 2008) and the *Amsterdam Dynamic Facial Expression Set* (ADFES) (van der Schalk, Hawk, Fischer, & Doosje, 2011).

FaceReader contains four different face models that are used to find the best fit for the face that is going to be analysed. These models are: (a) "General," the default face model; (b) "Children," a model for children between the ages of 3 and 10; (c) "East Asian," a model for East Asian faces, e.g. Japanese or Chinese; (d) "Elderly," a model for participants ages 60 and older. We set FaceReader to "General." The description in the FaceReader software itself states that "this model should work reasonably well under most circumstances for most people." We did *not* use any type (a priori or continuous) of participant calibration settings. For more information see the FaceReader reference manual, p. 53-54.

Validation 1 – Basic Emotions

Method

We calculated matching scores (accuracy) (see Ekman, et al., 1969; Russell, 1994) for recognition of prototypical facial expressions (Ekman, Friesen & Hager, 2002) of basic emotions (Ekman, et al., 1969; Ekman & Cordano, 2011). For basic emotion recognition, we adapted the definition of matching score for human recognition from Nelson and Russell (2013), specifically

"the percentage of observers who selected the predicted label" (p. 9). In the case of AFC software, observers become n = 1, i.e. the software itself, therefore we defined the matching score for the AFC software as *percentage of images that were recognized with the predicted label.*

Results

Accuracy for basic emotions. FaceReader recognized 88% of the target emotional labels in the 207 unique images in the *Warsaw Set of Emotional Facial Expression Pictures* (WSEFEP) (Olszanowski, et al., 2008) and 89% in the 154 unique images in the *Amsterdam Dynamic Facial Expression Set* (ADFES) (van der Schalk, et al., 2011). FaceReader failed to detect a face in 0.95% and 3.77% of the images, respectively.

***How specific emotions performed*.** FaceReader achieved a best recognition score (96%) of happiness for both ADFES and WSEFEP data sets. FaceReader performed the worst in correctly recognizing anger, with an overall average accuracy of 76%. The software classified neutral faces as neutral in 94% of cases. For general accuracy organized by basic emotions, see Table 1. For the confusion matrix for Table 1, which shows the number of false and true positives and negatives, see Table 2.

On average, FaceReader recognized female (89%) emotional faces better than male (86%). See Table 3 for an overview of the performance by gender. FaceReader best recognized the emotions of people of Dutch (91%), less so of Caucasian (88%) and worst for those of Turkish-Moroccan (86%) origin, see Table 4.

Across both datasets, FaceReader correctly recognized 89% of expressions on average, whereas human participants only recognized 85%. We manually computed the average human accuracy for WSEFEP from the original dataset made available by Olszanowski et al. (2008) and we took the original raw (%) values from Table 2 from Study 1 by van der Schalk et al. (2011). See Table 5 for a detailed overview.

Table 1

FaceReader Accuracy – Specific Basic Emotions: Overall

Emotion	Database	Number	Matched	Accuracy	Average
Neutral	ADFES	22	21	95%	94%
	WSEFEP	30	28	93%	
Happiness	ADFES	23	22	96%	96%
	WSEFEP	30	29	97%	
Sadness	ADFES	23	22	96%	86%
	WSEFEP	30	23	77%	
Anger	ADFES	25	19	76%	76%
	WSEFEP	30	23	77%	
Surprise	ADFES	18	17	94%	94%
	WSEFEP	27	25	93%	
Fear	ADFES	21	16	76%	82%
	WSEFEP	32	28	88%	
Disgust	ADFES	22	20	91%	92%
	WSEFEP	28	26	93%	
Total	ADFES	154	137	89%	88%
	WSEFEP	207	182	88%	

Note. Number = number of images of specific emotion in the dataset; Matched = number of images of specific emotion in the dataset that FaceReader classified properly. See Table 2 for confusion matrix.

Table 2

Confusion Matrix for Table 1

		FaceReader classification							
		Neutral	Happiness	Sadness	Anger	Surprise	Fear	Disgust	Total (Target)
Target-label	Neutral	49	0	1	0	0	1	1	52
	Happiness	0	51	0	0	1	0	1	53
	Sadness	6	0	45	1	0	0	1	53
	Anger	9	0	3	42	0	0	1	55
	Surprise	0	0	1	0	42	2	0	45
	Fear	4	1	1	0	3	44	0	53
	Disgust	2	0	1	1	0	0	46	50
	Total (FR)	70	52	52	44	46	47	50	361

Note. Total (FR) is the number of times FaceReader classified the basic emotion per target-label category. *Total (Target)* is number of times the basic emotion target-label is present.

Table 3

Facereader Accuracy – Specific Basic Emotions: Gender

Emotion	Database	Gender	Number	Matched	Accuracy	Average
Neutral	ADFES	Male	12	12	100%	
		Female	10	9	90%	93%*
	WSEFEP	Male	14	12	86%	
		Female	16	14	88%	89%†
Happiness	ADFES	Male	12	12	100%	
		Female	11	10	91%	100%*
	WSEFEP	Male	14	14	100%	
		Female	16	15	94%	92%†
Sadness	ADFES	Male	13	12	92%	
		Female	10	9	90%	82%*
	WSEFEP	Male	14	10	71%	
		Female	16	13	81%	86%†
Anger	ADFES	Male	15	12	80%	
		Female	10	7	70%	76%*
	WSEFEP	Male	14	10	71%	
		Female	16	13	81%	76%†
Surprise	ADFES	Male	9	9	100%	
		Female	9	8	89%	92%*
	WSEFEP	Male	12	10	83%	
		Female	15	15	100%	94%†
Fear	ADFES	Male	10	7	70%	
		Female	11	9	82%	73%*
	WSEFEP	Male	16	12	75%	
		Female	16	15	94%	88%†
Disgust	ADFES	Male	12	10	83%	
		Female	10	10	100%	88%*
	WSEFEP	Male	14	13	93%	
		Female	14	13	93%	96%†
Total	ADFES	Male	83	74	89%	
		Female	71	62	87%	86%*
	WSEFEP	Male	98	81	83%	
		Female	109	98	90%	89%†
Average	All	Male	181	155	86%	
		Female	180	160	89%	

Note. Number = number of images of specific emotion in the dataset; Matched = number of images of specific emotion in the dataset that FaceReader properly classified. * - male, † - female

Table 4

Facereader Accuracy – Specific Basic Emotions: Ethnicity

Emotion	Ethnicity	Number	Matched	Accuracy
Neutral	Dutch	12	11	92%
	T-M	10	10	100%
	Caucasian	30	28	93%
Happiness	Dutch	12	12	100%
	T-M	11	10	91%
	Caucasian	30	29	97%
Sadness	Dutch	12	12	100%
	T-M	11	10	91%
	Caucasian	30	23	77%
Anger	Dutch	13	10	77%
	T-M	12	8	67%
	Caucasian	30	23	77%
Surprise	Dutch	11	11	100%
	T-M	7	6	86%
	Caucasian	27	25	93%
Fear	Dutch	13	10	77%
	T-M	8	6	75%
	Caucasian	32	28	88%
Disgust	Dutch	12	11	92%
	T-M	10	9	90%
	Caucasian	28	26	93%
Total	Dutch	85	77	91%
	T-M	69	59	86%
	Caucasian	207	182	88%

Note. T-M = Turkish-Moroccan; Number = number of images of specific emotion in the dataset; Matched = number of images of specific emotion in the dataset that FaceReader properly classified.

Table 5

Facereader vs. Human Accuracy

Emotion	Database	FR	Human	Average
Neutral	ADFES	95%	(-)	94%*
	WSEFEP	93%	67%	67%†
Happiness	ADFES	96%	91%	97%*
	WSEFEP	97%	87%	89%†
Sadness	ADFES	96%	82%	87%*
	WSEFEP	77%	88%	85%†
Anger	ADFES	76%	88%	77%*
	WSEFEP	77%	87%	88%†
Surprise	ADFES	94%	89%	94%*
	WSEFEP	93%	89%	89%†
Fear	ADFES	76%	84%	82%*
	WSEFEP	88%	69%	77%†
Disgust	ADFES	91%	86%	92%*
	WSEFEP	93%	91%	84%†
Total	ADFES	89%	87%	89%*
	WSEFEP	88%	82%	85%†

Note. FR = FaceReader, * - FaceReader, † - Human. We computed manually the average human accuracy for WSEFEP from WSEFEP dataset made available by Olszanowski et al. (2008) and we took the original raw (%) values from Table 2 from Study 1 by van der Schalk et al. (2011).

We also computed the matching score (accuracy) for the *Karolinska Directed Emotional Faces* (KDEF) dataset (Lundqvist, Flykt, & Öhman, 1998) for FaceReader 6.0, which correctly recognized 86% of basic emotions on average. In 2005, FaceReader 1.0 correctly recognized 89% of emotions correctly (den Uyl & van Kuilenburg, 2005; van Kuilenburg, et al., 2005) but as mentioned in the introduction already, the comparison method used in 2005 was not as conservative as the approach in this paper. Therefore, the direct comparison between FaceReader 1.0 and 6.0 on the same dataset as the one used in 2005 indicates that the previous version is better by 3%. However, it must be highlighted that FaceReader 1.0 was specifically trained to deal well with KDEF dataset while FaceReader 6.0 now has much more robust and well-trained classifiers that perform just as well, if not better, on a much more diverse and thus generalizable set of images.

Validation 2 – FACS AUs

Method

Human inter-coder reliability. We needed first to assess the reliability of the manual human coding of the two datasets. Therefore, we calculated the agreement between the two FACS coders using the *Agreement Index,* as described by Ekman et al. (2002) in the FACS Manual who based his formula on Wexler (1972). This index is computed for every annotated image according to the following formula:

$$\frac{\text{(Number of AUs that both coders agree upon)} * 2}{\text{The total number of AUs scored by the two coders}}$$

For example, if an image was coded as 1+2+5+6+12 by one coder and as 5+6+12 by the other, the agreement index would be: $3 * 2 / 8 = 0.75$. Note that the intensity of the Action Unit

(AU) classification is ignored for the calculation of the agreement index, with the focus on whether the AU is active or not.

FaceReader FACS agreement index. In the results section of Validation 2, we used the same *Agreement Index* to demonstrate performance of FaceReader FACS. Therefore, we will compare the score of a pair of certified human coders and FaceReader FACS automated coding. It is an overall measure of accuracy in FACS coding.

Evaluation metrics. In order to evaluate the FaceReader performance for specific AUs, we provide metrics of *presence*; *recall*; *precision*; *F1* and *2AFC*. Those metrics are usually reported in AFC research when studying FACS performance, and we provide a brief description for the terms used in the results section in Tables 5 and 7 for the sake of clarity. An *AU* is the action unit number from the FACS manual (Ekman, Friesen & Hager, 2002). For a description (a name) of each of the Action Units (AUs), see Table 8. *Present* is the number of times an AU was coded in the dataset. *Recall* denotes the ratio of annotated AUs that were detected by FaceReader. For example, a recall of 0.84 for AU1 indicates, that 84% of the annotated images with AU1 are classified as such by FaceReader. *Precision* is a ratio of how often FaceReader is correct when classifying an AU as present. For example, in the case of AU1 the FaceReader classification is correct 83% of the time. A trade-off exists between the *recall* and *precision* measures, and a good classifier ought to have a reasonable score on both measures. The *F1* measure summarizes this trade-off in a single value. It is computed using the formula: 2 * ((precision * recall) / (precision + recall)). *Accuracy* simply represents the percentage of correct classifications. It is computed by dividing the number of correctly classified images (both positive and negative) by the total number of images. The *2AFC* descriptor represents a ‘two-alternative forced choice’. To compute this measure for an AU, FaceReader is presented with every possible combination of positive (present) and negative (not present) images in pairs of two. For each of these combinations of

positive and negative images, FaceReader should be able to determine which of the two images contains the AU in question. The consequence of this evaluation method is that – contrary to the other measures - a random classifier would score correctly around 50% of the time. This gives the *2AFC* measure a very intuitive interpretation. For example, a *2AFC* score of 0.93 for AU1 indicates that this classifier scores around 43% better than chance.

Results

Accuracy of FACS. The software reached a FaceReader FACS index of agreement with the two human coders of 0.70 over the 177 unique images of WSEFEP and an agreement of 0.68 over the 205 unique images of ADFES. For the WSEFEP and ADFES database, we used all available images of the basic emotions to compute the matching scores and we removed all base-line images to compute the FACS index of agreement. See Table 6 and Table 7 for a detailed overview. We provide the final two databases, including FACS AUs manual annotation, for the WSEFEP and ADFES in Supplementary Material.

How specific AUs performed. To evaluate and compare the quality of individual AU classifiers based on validation in our test, we decided to discriminate between AU classifiers based on the *F1* measure. The *F1* measure is a suitable metric for this purpose because it combines the important *recall* and *precision* measures, and displays the largest differences between the AU classifiers. See Tables 6 and 7 for the *F1* measure and all other FACS evaluation metrics.

Table 6

WSEFEP - The Performance of Action Units for Six Evaluation Metrics

AU	Present	Recall	Precision	F1	Accuracy	2AFC
1	80	0.81	0.87	0.84	0.86	0.93
2	47	0.91	0.72	0.80	0.88	0.95

4	69	0.64	0.69	0.66	0.75	0.82
5	71	0.76	0.96	0.85	0.89	0.91
6	42	0.95	0.67	0.78	0.88	0.97
7	51	0.90	0.50	0.64	0.71	0.84
9	26	0.73	0.76	0.75	0.93	0.94
10	11	0.91	0.16	0.27	0.69	0.85
12	29	1.00	0.78	0.88	0.95	1.00
15	33	0.70	0.70	0.70	0.89	0.89
17	67	0.51	0.85	0.64	0.78	0.88
20	20	0.40	0.32	0.36	0.84	0.77
23	19	0.47	0.39	0.43	0.86	0.88
24	20	0.60	0.50	0.55	0.89	0.85
25	105	0.89	0.91	0.90	0.88	0.96
26	34	0.85	0.62	0.72	0.87	0.94
Average		0.75	0.65	0.67	0.85	0.90

Note. For a description (a name) of each of the Action Units (AUs) see Table 8.

Table 7

ADFES - The Performance of Action Units for Six Evaluation Metrics

AU	Present	Recall	Precision	F1	Accuracy	2AFC
1	92	0.75	0.84	0.79	0.82	0.89
2	64	0.78	0.86	0.82	0.89	0.92
4	82	0.74	0.82	0.78	0.83	0.88
5	60	0.73	0.80	0.77	0.87	0.90
6	51	0.78	0.70	0.74	0.86	0.88
7	54	0.67	0.44	0.53	0.69	0.78
9	22	0.91	0.83	0.87	0.97	0.96
10	16	0.75	0.29	0.42	0.84	0.83
12	57	0.54	0.91	0.68	0.86	0.91
14	53	0.72	0.73	0.72	0.86	0.87
15	26	0.77	0.74	0.75	0.94	0.93
17	56	0.54	0.86	0.66	0.85	0.91
20	18	0.61	0.50	0.55	0.91	0.90
23	13	0.62	0.42	0.50	0.92	0.85
24	26	0.42	0.50	0.46	0.87	0.75
25	77	0.96	0.83	0.89	0.91	0.99
26	24	0.62	0.71	0.67	0.93	0.88
Average		0.70	0.69	0.68	0.87	0.88

Note. For a description (a name) of each of the Action Units (AUs) see Table 8.

Table 8

Names of Facial Action Coding System (FACS) Action Units (AUs)

AU	Name
1	Inner Brow Raise
2	Outer Brow Raise
4	Brow Lowerer
5	Upper Lid Raise
6	Cheek Raise
7	Lids Tight
9	Nose Wrinkle
10	Upper Lip Raiser
12	Lip Corner Puller
14	Dimpler
15	Lip Corner Depressor
17	Chin Raiser
20	Lip Stretch
23	Lip Tightener
24	Lip Presser
25	Lips Part
26	Jaw Drop

Note. The names of AUs are provided after FACS manual (Ekman, Friesen & Hager, 2002), p. 526.

*WSEFEP***.** Based on the *F1* measure the best classifiers – those that might already be good enough to pass the FACS test[2] – are AUs 1, 2, 5, 6, 9, 12 and 25 (*F1*: 0.74 - 1.00). The classifiers that performed reasonably well are AUs 4, 7, 15, 17 and 26 (*F1*: 0.64 - 0.72). The AUs that performed less well are AUs 10, 20, 23 and 24 (*F1*: 0.27 - 0.55).

ADFES. Based on the *F1* measure the best classifiers – those that might be already be good enough to pass the FACS test – are AUs 1, 2, 4, 5, 6, 9, 15 and 25 (*F1*: 0.74 - 1.00). The classifiers that performed reasonably well are AUs 12, 14, 17 and 26 (*F1*: 0.64 - 0.72). The AUs

[2] Certified FACS coders are informed after passing the test that the score of 0.70 is needed to past the FACS exam, see Discussion

that performed less well are AUs 7, 10, 20, 23 and 24 (*F1*: 0.27 - 0.55). FaceReader can also classify AUs 14, 18, 27 and 43, but these were not sufficiently present in the dataset to evaluate their performance.

Discussion

Humans vs. FaceReader

Basic emotions accuracy. The results of this validation test are meaningful only in relation to human performance rates for both basic emotion recognition and FACS coding. We computed the accuracy of basic emotions recognition by humans for the two datasets. For ADFES it is 87% (van der Schalk et al., 2011, Table 2) and for WSEFEP it is 82% (Olszanowski et al., 2008, original dataset), as shown in Table 5. As described earlier, FaceReader recognized 88% of the target emotional labels in WSEFEP and 89% in ADFES, for an 89% weighted average.

Such results for humans do not come as a surprise, as in the meta-analysis of recognition studies of facial expressions in humans, Russell (1994) and Nelson and Russell (2013) never reported accuracy higher than 90% and often as low as 60-80%. Therefore, FaceReader accuracy in detecting basic emotions is the same as participants' judgments of the two tested databases and within the score ranges reported in the literature of human emotion perception.

FACS accuracy. In order to pass the FACS certification exam, a human coder must reach an agreement of 0.70 with the master (i.e. criterion) coding of the final test. However, the FACS manual explicitly warns that in order to reach 0.80 – 0.90 agreement scores, a human coder must practice for at least one thousand hours after passing the bar. As mentioned, FaceReader reached a FACS index of agreement of 0.70 over WSEFEP and an agreement of 0.68 over ADFES, hence a 0.69 weighted average. Therefore, FaceReader performance has fallen short of reaching the

agreement of 0.70 but has performed reasonably well, especially in comparison to inexperienced human coders and other AFC systems (for a review see Valstar, et al., 2012).

Basic emotions vs. FACS

As suggested by one of the reviewers, FaceReader seems to perform better in recognizing basic emotions than FACS Action Units. FaceReader performs as well as humans in emotion recognition, but not in Action Unit recognition. We can think of at least three reasons why this might be the case. The first and most compelling argument comes from simple probability calculation. While the FaceReader's FACS module has to correctly classify a facial expression into a combination of 17 possible categories (defined as AUs), the basic emotion module has to classify an expression into only one of six possible discrete categories (defined as basic emotions). A higher number of possible classification categories means higher error rates. The second reason is that FACS is an *expert* coding system while basic emotion coding is more of a *naive* coding system. Most humans can recognize basic emotions (Ekman, et al., 1969) but recognizing action units requires extensive and specialized training. In other words, applying FACS is a far more complex task than basic emotion coding. The third reason is that automated facial coding of basic emotions has been possible since 2005 (den Uyl & van Kuilenburg, 2005; van Kuilenburg, et al., 2005) while the capability to do FACS coding was added to FaceReader in 2012. The emotion coding algorithms have had more time to mature and undergo further refinement.

Limitations

We recognize the possibility that the average FaceReader FACS index of agreement of 0.69 may be inflated due to frontal, close-up, posed photographs of superior quality, not normally found in datasets of spontaneous (and more ecologically valid) facial expressions. Importantly, the same argument is not plausible for the FaceReader basic emotions accuracy score because

humans perform identically on the same, posed material and they did not reach a 100% recognition score either.

Conclusions

In general, we believe that FaceReader has proven to be a reliable indicator of facial expressions of basic emotions in the past decade and has the potential to become similarly robust with FACS coding. For version 6.0 of FaceReader, researchers may report a general 88% basic emotion accuracy score and use values from Table 1 for specific emotions. For FACS accuracy, the FaceReader index of agreement is 0.69 and performance on specific AUs can be quoted from Table 6 and 7.

Further, FaceReader categorization of basic emotions is reliable and does not need human correction. However, the beta FACS module could be used for semi-automated coding, as in GeFACT (With & Delplanque, 2013), where a human FACS certified coder corrects the software's FACS output to reach acceptable (i.e. above 0.70) levels of agreement. It is also anticipated that future versions of FaceReader may provide better performance, particularly in light of the revival of deep learning in artificial neural networks (LeCun, Bottou, Bengio & Haffner, 1998; Le, Ranzato, Monga, Devin & Chen 2013).

References

Chentsova-Dutton, Y. E., & Tsai, J. L. (2010). Self-focused attention and emotional reactivity: the role of culture. *Journal of Personality and Social Psychology, 98*(3), 507-519. doi: 10.1037/a0018534

Chiu, M. H., Chou, C. C., Wu, W. L., & Liaw, H. (2014). The role of facial microexpression state (FMES) change in the process of conceptual conflict. *British Journal of Educational Technology, 45*(3), 471-486. doi: 10.1111/bjet.12126

Cook, I. A., Warren, C., Pajot, S. K., Schairer, D., & Leuchter, A. F. (2011). Regional brain activation with advertising images. *Journal of Neuroscience, Psychology, and Economics*, *4*(3), 147. doi: 10.1037/a0024809

Cootes, T. F., & Taylor, C. J. (2004). Statistical models of appearance for computer vision. *Imaging Science and Biomedical Engineering, University of Manchester,* Manchester M13 9PT, UK March, 8.

Danner, L., Sidorkina, L., Joechl, M., & Duerrschmid, K. (2014). Make a face! Implicit and explicit measurement of facial expressions elicited by orange juices using face reading technology. *Food Quality and Preference, 32*, 167-172. doi: 10.1016/j.foodqual.2013.01.004

de Wijk, R. A., He, W., Mensink, M. G., Verhoeven, R. H., & de Graaf, C. (2014). ANS responses and facial expressions differentiate between the taste of commercial breakfast drinks. *PloS one, 9*(4), e93823. doi: 10.1371/journal.pone.0093823

den Uyl, M., & van Kuilenberg, H. (2005). The FaceReader: Online facial expression recognition. In L.P.J.J. Noldus, F. Grieco, L.W.S. Loijens and P.H. Zimmerman (Eds.), *Proceedings of Measuring Behavior 2005, 5th International Conference on Methods and*

Techniques in Behavioral Research (pp. 589-590). Wageningen, the Netherlands: Noldus Information Technology.

Ekman, P., & Cordano, D. (2011). What is meant by calling emotions basic. *Emotion Review, 3*(4), 364-370. doi: 10.1177/1754073911410740

Ekman, P., & Friesen, W. V. (1978). *Facial action coding system: A technique for the measurement of facial movement.* Palo Alto, CA: Consulting Psychologists Press.

Ekman, P., Friesen W.V., & Hager J. C. (2002). *Facial action coding system: The manual*. Salt Lake City, UT: Research Nexus.

Ekman, P., Sorenson, E. R., & Friesen, W. V. (1969). Pan-cultural elements in facial displays of emotion. *Science, 164*(3875), 86-88. doi:10.1126/science.164.3875.86

Garcia-Burgos, D., & Zamora, M. C. (2013). Facial affective reactions to bitter-tasting foods and body mass index in adults. *Appetite, 71*(1), 178-186. doi: 10.1016/j.appet.2013.08.013

Goldberg, J. H. (2014). Measuring Software Screen Complexity: Relating Eye Tracking, Emotional Valence, and Subjective Ratings. *International Journal of Human-Computer Interaction, 30*(7), 518-532. doi: 10.1080/10447318.2014.906156

Lajante, M., Droulers, O., Dondaine, T., & Amarantini, D. (2012). Opening the “black box” of electrodermal activity in consumer neuroscience research. *Journal of Neuroscience, Psychology, and Economics, 5*(4), 238-249. doi: 10.1037/a0030680

Le, Q. V., Ranzato, M., Monga R., Devin, M., & Chen, K. (2013, May). Building high-level features using large scale unsupervised learning. *Proceedings of the 29th IEEE International Conference on Machine Learning* (pp. 8595-8598). Edinburg, Scotland, UK.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE, 86*(11), 2278-2324.

Lewinski, P., Fransen, M. L., & Tan, E.S.H. (2014). Predicting advertising effectiveness by facial expressions in response to amusing persuasive stimuli. *Journal of Neuroscience, Psychology, and Economics, 7*(1), 1-14. doi: 10.1037/npe0000012

Lundqvist, D., Flykt, A., & Öhman, A. (1998). The Karolinska directed emotional faces (KDEF). *CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet.*

Micu, A. C., & Plummer, J. T. (2010). Measurable emotions: How television ads really work patterns of reactions to commercials can demonstrate advertising effectiveness. *Journal of Advertising Research*, *50*(2), 137-153. doi: 10.2501/S0021849910091300

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

Noldus. (2014) FaceReader: Tool for automatic analysis of facial expression: Version 6.0. Wageningen, the Netherlands: Noldus Information Technology B.V.

Olszanowski, M., Pochwatko, G., Kukliński, K., Ścibor-Rylski, M., & Ohme, R. (2008, June). *Warsaw set of emotional facial expression pictures - Validation study*. Opatija, Croatia: EAESP General Meeting.

Ramsøy, T. Z., Friis-Olivarius, M., Jacobsen, C., Jensen, S. B., & Skov, M. (2012). Effects of perceptual uncertainty on arousal and preference across different visual domains. *Journal of Neuroscience, Psychology, and Economics*, *5*(4), 212-226. doi: 10.1037/a0030198

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*(6), 1161-1178. doi: 10.1037/h0077714

Russell, J. A. (1994). Is there universal recognition of emotion from facial expressions? A review of cross-cultural studies. *Psychological Bulletin, 115,* 102–141. doi: 10.1037/0033-2909.115.1.102

Terzis, V., Moridis, C. N., & Economides, A. A. (2012). The effect of emotional feedback on behavioral intention to use computer based assessment. *Computers & Education*, *59*(2), 710-721. doi: 10.1016/j.compedu.2012.03.003

Terzis, V., Moridis, C. N., & Economides, A. A. (2013). Measuring instant emotions based on facial expressions during computer-based assessment. *Personal and Ubiquitous Computing*, *17*(1), 43-52. doi: 10.1007/s00779-011-0477-y

Valstar, M. F., Mehu M., Jiang B., Pantic M., & Scherer K. R. (2012). Meta-analysis of the first facial expression recognition challenge. *IEEE Transactions on Systems, Man, and Cybernetics – Part B: Cybernetics*, *42*(4), 966-979. doi: 10.1109/TSMCB.2012.220067

van der Schalk, J., Hawk, S. T., Fischer, A. H., & Doosje, B. (2011). Moving faces, looking places: Validation of the Amsterdam Dynamic Facial Expression Set (ADFES). *Emotion, 11*(4), 907-920. doi: 10.1037/a0023853

van Kuilenburg, H., Wiering, M., den Uyl, M. (2005). A model based method for facial expression recognition. In D. Hutchison, T. Kanade, J. Kittler, J.M. Kleinberg, F. Matern, J.C. Mitchell, M. Noar, G. Weikum… (Eds.), *Lectures Notes in Computer Science: Vol. 3720. Machine Learning: ECML 2005* (pp. 194-205). Berlin, Germany: Springer-Verlag. doi: 10.1007/11564096_22

Wexler, D. (1972). Method for unitizing protocols of descriptions of emotional states. *Journal of Supplemental Abstracts Service, 2*, 116.

With, S., & Delplanque, S. (2013, August). *The Geneva Facial Action Toolbox (GeFACT) A matlab® toolbox for processing facial action units' time series*. Poster presented at the 2013 biennial meeting of International Society for Research on Emotions, Berkeley, California. Abstract retrieved from

http://www.isre2013.org/online/core_routines/view_abstract_no.php

show_close_window=yes&abstractno=78

Supplementary Material

http://dx.doi.org/10.1037/npe0000028.supp