Facial Expression Analysis with AFFDEX and FACET: A Validation Study

Sabrina Stöckli[a]

Michael Schulte-Mecklenbeck[a, b]

Stefan Borer[a]

Andrea C. Samson[c,d]

a Institute of Marketing and Management, University of Bern, Bern, Switzerland

b Max Planck Institute for Human Development, Berlin, Germany

c Swiss Center for Affective Sciences, University of Geneva, Geneva, Switzerland

d Department of Psychiatry and Behavioral Science, Stanford University School of Medicine, Stanford, CA, USA

Author Note

We thank Adem Halimi and Elena von Wyttenbach for their help with the data collection.

Correspondence about this article should be addressed to: Sabrina Stöckli, University of Bern, Institute of Marketing and Management, Department of Consumer Behavior, Engehaldenstrasse 4, 3012 Bern, Switzerland, Email: Sabrina.Stoeckli@imu.unibe.ch

Abstract

The goal of this study was to validate AFFDEX and FACET, two algorithms classifying emotions from facial expressions, in iMotions's software suite. In Study 1, pictures of standardized emotional facial expressions from three databases, the *Warsaw Set of Emotional Facial Expression Pictures* (WSEFEP), the *Amsterdam Dynamic Facial Expression Set* (ADFES) and the *Radboud Faces Database* (RaFD), were classified with both modules. Accuracy (*Matching Scores)* was computed to assess and compare the classification quality. Results show a large variance in accuracy across emotions and databases, with a performance advantage for FACET over AFFDEX. In Study 2, 110 participants' facial expressions were measured while being exposed to emotionally evocative pictures from the *International Affective Picture System* (IAPS), the *Geneva Affective Picture Database* (GAPED) and the *Radboud Faces Database* (RaFD). Accuracy again differed for distinct emotions, and FACET performed better. Overall, iMotions can achieve acceptable accuracy for standardized pictures of prototypical (vs. natural) facial expressions, but performs worse for more natural facial expressions. We discuss potential sources for limited validity and suggest research directions in the broader context of emotion research.

Keywords: emotion classification; facial expression; FACS; AFFDEX; FACET

Facial Expression Analysis with AFFDEX and FACET: A Validation Study

The de facto standard for measuring emotional facial expressions is the *Facial Action Coding System* (FACS; Ekman & Friesen, 1976). This anatomy-based system allows human coders to evaluate emotions based on 46 observable Action Units (AU), facial movements that account for facial expressions and in turn for the expression of emotions (Ekman & Friesen, 1976). FACS coding requires certified coders who are trained for up to 100 hours (e.g., at workshops by the Paul Ekman Group LLC). On top of this time-intensive training, the coding process itself is also time- and labor-intensive. Video recordings of participants' faces are often recorded with a resolution of 24 frames per second, meaning that for each second of recording the coder has to produce 24 ratings of the 46 AUs. So for one participant with only one minute of video, 1440 individual ratings are necessary. Assuming that a coder could rate one picture per second, this would add up to approximately 24 minutes of work for one minute of video data (see Ekman & Oster, 1979).

Automated facial expression analysis has progressed significantly in the last three decades and developed into a promising tool that may overcome the limitations of human-based FACS coding. This progress is largely due to rapid developments in computer science which have made automated facial expression analysis more valid, reliable and accessible (e.g., Beumer, Tao, Bazen, & Veldhuis, 2006; Cootes, Edwards, & Taylor, 2001; Lewinski, den Uyl, & Butler, 2014; Swinton & El Kaliouby, 2012; Valstar, Jiang, Mehu, Pantic, & Scherer, 2011; Viola & Jones, 2001).

One commercial tool for automated facial expression analysis is part of a software suite by iMotions (www.imotions.com). iMotions's biometric research platform can be used for various types of academic and business-related research and offers automated facial expression analysis in combination with EEG, GSR, EMG, ECG, eye tracking and surveys. The automated facial expression analysis part allows the user to record videos with a laptop camera, mobile phone camera or standalone webcam. iMotions then detects changes in key face features (i.e., facial landmarks such as brows, eyes and lips) and generates data representing the basic emotions of the recorded face. Researchers can choose between two different modules to classify emotions of facial expressions: the FACET module, based on the FACET algorithm (formerly the Computer Expression Recognition Toolbox (CERT) algorithm; Littlewort et al., 2011) and the AFFDEX module, based on the AFFDEX algorithm by Affectiva Inc. (El Kaliouby & Robinson, 2005; McDuff, El Kaliouby, Kassam, & Picard, 2010). These algorithms detect facial landmarks and apply a set of rules based on psychological theories and statistical procedures to classify emotions. Different algorithms, like AFFDEX and FACET, use distinct statistical procedures, facial databases and facial landmarks to train the machine learning procedures and ultimately classify emotions (iMotions, 2016).

In contrast to the growing interest in applying automated facial expression analysis, there is only a surprisingly little number of peer-reviewed publications validating these algorithms (except for several conference presentations on this topic, e.g., Baltru, Robinson, Morency, & others, 2016; Littlewort et al., 2011; McDuff et al., 2010; Taggart, Dressler, Kumar, Khan, & Coppola, n.d.). It is notable that the lack of validations for automated emotion classification is more pronounced than the lack of validations for automated detection and description of distinct AUs. FaceReader, a software marketed by Noldus (www.noldus.com), is the only tool, we are aware of, with published validation work (den Uyl & van Kuilenburg, 2005; Lewinski, den Uyl, & Butler, 2014; van Kuilenburg, Wiering, & den Uyl, 2005). As there is no such validation for iMotions's AFFDEX and FACET modules, the present research fills this gap by validating and comparing their performance.

The Origins of Facial Expression Analysis

External facial expressions reveal much about our inner emotional states (Ekman & Friesen, 1982; Ekman, 1992a; Ekman & Oster, 1979). Early research on facial expressions is based on *discrete emotion theory* and has focused on analyzing basic emotions that are universally recognized (i.e., anger, disgust, fear, happiness, sadness, and surprise). *Discrete emotion theory* assumes these basic emotional facial expressions to reflect holistic emotion programs that cannot be broken down into smaller emotion units (e.g., Ekman, 1992a; Ekman et al., 1987). A crucial factor for the dominance of the distinct facial expressions of basic emotions was the introduction of FACS (Ekman & Friesen, 1976).

Within FACS, 46 facial AUs represent distinct movements displayed on the face, and emerge by activating one or a combination of facial muscles. FACS provides a coding schema for AU activity and intensity (Ekman et al., 1987; Ekman & Friesen, 1976). FACS coding, in turn, allows inferences about basic emotions, because research has demonstrated that the combination of certain AUs is associated with certain emotions. For instance, activating AU 4 (i.e., brow lowerer; *corrugator supercilii*) leads to a lowering of the eyebrows. This movement typically occurs when expressing emotions such as anger, disgust or sadness (Du, Tao, & Martinez, 2014; Ekman & Friesen, 2003). Given that there are numerous publications that address theoretical and practical aspects of FACS (e.g., Ekman & Friesen, 1976; Hwang & Matsumoto, 2016; Meiselman, 2016), we do not discuss these in more details.

Although FACS is widely acknowledged as being objective and reliable, there is an ongoing debate on FACS' legitimacy as basis of facial expression analysis. This debate originates from the two theoretical emotion perspectives: While discrete emotion theorists (i.e., *basic emotion perspective*) acknowledge only a small set of basic emotions and conceptualize these emotions as discrete and fundamentally different, other emotion theorists have urged for a paradigm shift from the basic emotion perspective to an *appraisal perspective* (Ellsworth & Scherer, 2003; Vallverdu, 2014). According to *appraisal theory*, there is a large set of (non-prototypical) emotions and a focus is set on the cognitive antecedents of emotions, namely that emotions are shaped by the evaluation of the context (e.g., Ellsworth & Scherer, 2003; Roseman & Smith, 2001).

For the legitimacy of the theoretical basis of iMotions's automated facial expression analysis, the debate between the basic and the appraisal perspective reveals three critical aspects: First, iMotions's automated facial expression analysis assumes that there is a direct link between emotion production and emotion recognition. Indeed, iMotions's algorithms recognize expressions but not inevitably emotions. Appraisal theorists argue that this one-to-one relationship between a facial expression and an experienced emotion can be incorrect and that a separate inference step is required (see Mortillaro, Meuleman, & Scherer, 2015). Second, iMotions's algorithms do not integrate contextual information into emotion recognition. Indeed, iMotions's algorithms categorize facial expressions without any information about environment, subject or other situational factors. Appraisal theorists suggest that the context influences emotions. When inferring appraisals from behavior, it is therefore necessary to not only rely on markers of emotions, but also to consider (contextual) information about what causes the emotion (see, e.g., Aviezer, Trope, & Todorov, 2012; Mortillaro, Meuleman, & Scherer, 2015). Third, iMotions's algorithms fail to detect non-prototypical emotions. While they are trained to recognize prototypical facial expressions identifying facial expressions of compound and/or subtle emotions is not within their ability. Many appraisal theorists argue that this is particularly problematic since facial expressions are rarely prototypical in everyday life (e.g., Du, Tao, & Martinez, 2014; Mortillaro, Meuleman, & Scherer, 2015; Scherer & Ellgring, 2007).

In the light of these aspects it has been suggested to adopt a dimensional framework. In fact, expanding the dimensional basis of emotion categories may facilitate to detect non-prototypical, i.e., subtle and more complex emotions. Yet, automated facial expression analysis adopting dimensional emotion models in inferring emotions is still underexplored in this regard (Mortillaro, Meuleman, & Scherer, 2015).

Note that despite increasing concern of the basic emotion perspective defining facial expression analysis (see Scherer & Ellgring, 2007), to date, basic emotion theory (i.e., FACS) has considerably shaped all methods of measuring facial expressions. Given that iMotions's AFFDEX and FACET explicitly rely on FACS (Ekman & Friesen, 1976) and that this research aims to validate FACS-based iMotions's AFFDEX and FACET, we do not discuss the theoretical basis of iMotions's and the adequacy of other emotion theories in more detail here. A comprehensive and well-founded theoretical contextualization of automated facial expression analysis can be found elsewhere (see Mortillaro, Meuleman, & Scherer, 2015).

Measuring Facial Expressions

In addition to human observation and coding of facial expressions (e.g., by means of FACS), there are two automated methods of measuring emotions by means of facial expressions (see Cohn & Sayette, 2010; iMotions, 2016; Wolf, 2015): *facial electromyography activity* and computer-based video classification algorithms (e.g., AFFDEX, FACET, or FaceReader).

Facial electromyography activity (fEMG) directly measures electrical changes in facial muscles and thus can record even subtle facial muscle activities. fEMG requires special biosensors placed on the face, is sensitive to motion artifacts and can be intrusive (see Schulte-Mecklenbeck et al., 2017). Further, the direction of a specific muscle activity cannot be detected and crosstalk signals resulting from surrounding muscles can impede the analysis of specific muscles. It is therefore often not possible to clearly classify a distinct emotion with fEMG (Huang, Chen, & Chung, 2004; iMotions, 2016; Stets & Turner, 2014; Wolf, 2015). Automated facial expression analysis seems to be a promising alternative to fEMG for the measurement and classification of emotions by means of facial expressions.

Automated Facial Expression Analysis

In the last decade, most advancements in the area of automated facial expression analysis were on detecting distinct basic emotions and specific facial muscle activities (El Kaliouby & Robinson, 2005; Lewinski et al., 2014; Valstar et al., 2011; Zeng, Pantic, Roisman, & Huang, 2009; for a review see Calvo et al. 2014). CERT (precursor of FACET; Littlewort et al., 2011) and Noldus's FaceReader (den Uyl & van Kuilenburg, 2005) were the first software tools developed to automatically classify static (i.e., still pictures) and dynamic (i.e., videos) facial expressions. Since then, the market for automated facial expression analysis has changed rapidly. Currently, there are three major software tools for automated AU identification and emotion classification: Noldus's FaceReader (den Uyl & van Kuilenburg, 2005), iMotions's AFFDEX module (El Kaliouby & Robinson, 2005; McDuff, El Kaliouby, Cohn, & Picard, 2015; Zeng et al., 2009) and iMotions's FACET module (Littlewort et al., 2011)[1].

There is currently an ongoing lively debate on the paradigm shift from the basic emotion perspective to an appraisal perspective to find the appropriate theory integration in the area of automated facial emotion classification (see Vallverdu, 2014). In general, the criticism on the basic emotion perspective implies that, though automated facial expression analysis classifies basic emotional expression categories, it might not ultimately measure emotional states. The fact that automated facial expression analysis relies on the assumption of basic emotions and emotional coherence, that is, that there is coherence between emotion

[1] See Appendix A for more details on the emotion conceptualization of iMotions, how AFFDEX and FACET are specified by FACS and the assumption of a limited set of distinct basic emotions.

and facial expression (see Bonanno & Keltner, 2004; Reisenzein, Studtmann, & Horstmann, 2013) limits the interpretation of data generated by automated facial expression analysis and questions the generalizability of automated emotion classification (Wolf, 2015). Some researchers argue that inference based on data generated by automated facial expression analysis should build upon emotion theories that go beyond the basic emotion perspective, adopt an appraisal perspective and allow more flexibility to consider different contexts. An extended overview of the proposition of a paradigm shift from basic emotion recognition to an appraisal perspective can be found, e.g., in Vallverdu (2014).

Measuring Emotions with iMotions's Facial Expression Analysis

Initially, iMotions implemented automated facial expression analysis based on the FACET algorithm (see Littlewort et al., 2011) developed by the technology company Emotient. In 2016, iMotions announced a switch to AFFDEX from the technology company Affectiva. This switch was most likely connected to the acquisition of Emotient by Apple Inc. While new customers of iMotions are only able to purchase AFFDEX, existing customers are still able to apply FACET until 2020 (personal conversation with iMotions, 2016)[2].

Surprisingly, there is only little evidence that automated facial expression analysis is as reliable as human FACS coding and fEMG (Lewinski et al., 2014; Littlewort et al., 2011; Terzis, Moridis, & Economides, 2010). A validation study of FaceReader (Version 6; den Uyl & van Kuilenburg, 2005; Lewinski et al., 2014) resulted in a classification accuracy of 88% of the faces in the *Warsaw Set of Emotional Facial Expression Pictures* (WSEFEP) and of 89% of the faces in the *Amsterdam Dynamic Facial Expression Set* (ADFES), two publicly available datasets of validated facial expressions of emotions. In terms of basic emotions, FaceReader performs best for happiness (classification accuracy of 96% for WSEFEP and

[2] For a detailed description of the technical background, the data generation and analytics of iMotions's facial expression analysis, see Appendix A and Appendix B.

ADFES) and worst for anger (classification accuracy of 76% for WSEFEP and ADFES). Although Lewinski et al. (2014) provide a first estimation of the automated classification accuracy, we see room for further validation and improvement: (a) since it is not clear what criteria these authors applied to classify a picture as correctly recognized (see Lewinski et al., 2014)[3]; (b) there is currently no research available which validates and compares iMotions's AFFDEX and FACET modules. We aim to close that gap with this research.

Research Overview

We performed two studies to validate and compare the performance of iMotions's facial expression analysis modules AFFDEX and FACET (iMotions, 2016). In Study 1, we adapted a validation procedure based on Lewinski et al. (2014) by computing accuracy measures for recognizing facial expressions in images from three databases of normed facial expressions. In Study 2, we exposed participants to emotionally evocative pictures. We computed accuracy measures for the matching between the emotional content of the pictures and participants' facial expressions.

Study 1

Method

Design and Procedure. We measured the accuracy of emotion classification of iMotions's AFFDEX and FACET using three publicly available databases of facial expression pictures: WSEFEP (Olszanowski, Pochwatko, Kukliński, Ścibor-Rylski, Lewinski, & Ohme, 2008), ADFES (van der Schalk, Hawk, Fischer, & Doosje, 2011) and RaFD (Langner et al., 2010). All of these database pictures are validated to show FACS-consistent facial expressions of basic emotions. For both AFFDEX and FACET, a total of 600 pictures from the three databases were analyzed. The emotion classification was

[3] In addition, the FaceReader validation has not been conducted on the whole WSEFEP database (i.e., only on 207 instead of 210 pictures). Furthermore, the authors neither specify exclusion criteria in their paper nor did they provide such information upon request.

conducted in an automated manner using iMotions. Given that iMotions can only analyze video material, we generated a video (MP4 format) for all faces in all emotional states separately for WSEFEP, ADFES and RaFD pictures. In the video, every picture (i.e., facial expression) was shown for 5 seconds. For the analysis we cut the first and last second of data and analyzed the 'middle' 3 seconds. The first second (of the five second stimulus presentation window) was cut because iMotions's algorithms need ~1 second to converge toward a stable state (due to their neural network architecture). The last second was cut to ensure equal measurement periods. Analysis with and without the last second did not change our results.

Materials

The Amsterdam Dynamic Facial Expression Set (ADFES). This database consists of dynamic (video) and static (still picture) facial expressions of 22 white face models. Face models have been trained by FACS experts and pictures have been validated by 119 non-expert human judges (van der Schalk, Hawk, Fischer, & Doosje, 2011)[4]. In our analysis, we included the 153[5] static pictures (JPEG format, 1024 × 768 pixels) of the emotions anger, contempt, disgust, fear, happiness, sadness and surprise.

The Warsaw Set of Emotional Facial Expression Pictures (WSEFEP). This database consists of 210 pictures (JPEG format, 1725 × 1168 pixels) of 30 white face models. All pictures have been validated by a FACS coder and by a large sample (*N* = 1362) of non-expert human judges (Olszanowski et al., 2015)[6]. We included the pictures of the emotions anger, disgust, fear, happiness, sadness and surprise. For technical reasons, it was not

[4] The ADFES is freely accessible for non-commercial use at http://aice.uva.nl/research-tools/adfes-stimulus-set/adfes-stimulus-set.html
[5] ADFES does not provide a picture of face model F10 expressing surprise.
[6] The WSEFEP is freely accessible for non-commercial use at http://www.emotional-face.org./

possible to generate a video for face model MK. Thus, we used 174 WSEFEP pictures for this study.

The Radboud Faces Database (RaFD). This database consists of 536 pictures of 67 face models expressing basic emotions. All face models have been trained by FACS experts to express prototypical basic emotions. In addition to this, all pictures have been validated by FACS coders as well as by a large sample (*N* = 238) of non-expert human judges (Langner et al., 2010).[7] For the present study, we included 273 pictures of 39 white adults that express the emotions anger, contempt, disgust, fear, happiness, sadness and surprise. We limited ourselves to pictures of white adults. We selected these stimuli because facial expression analysis algorithms seem to be most accurate for Caucasian faces. Further, only using white faces allows a more accurate comparability with previous validations of methods to categorize emotional facial expressions (see Lewinski et al., 2014) and across different facial databases (see, e.g., O'Toole et al., 2008). Note that neither the here validated facial expression databases nor other databases comprise faces of all ethnicities.

Setting and Apparatus. iMotions's AFFDEX and FACET modules (Version 6.2) were used to classify the pictures from the three databases. We ran iMotions on a Lenovo T450s with Windows 8.1. Standard settings as described in the iMotions manual were used. iMotions provides probability-like values for all basic emotions anger, contempt, disgust, fear, happiness, sadness and surprise (see iMotions, 2016). In FACET these values are referred to as 'evidence values'; in AFFDEX as 'probabilities'. For detailed information about AFFDEX's and FACET's metrics see Appendix B.

---

[7] The RaFD is freely accessible for non-commercial use at http://www.socsci.ru.nl:8180/RaFD2/RaFD?p=main

Results

Matching Scores for Basic Emotions. Replicating the analysis technique of Lewinski et al., (2014) we computed a *Matching Score* (MS), which represents an estimate of iMotions's accuracy at recognizing facial expressions of basic emotions. MS is defined as the percentage of pictures that iMotions classified correctly (see Lewinski et al., 2014; Nelson & Russell, 2013). A classification was recorded as 'correct' when the highest value (out of all generated values for all basic emotions) matched with the database's emotion label. Thus, a higher MS indicates a greater likelihood of correctly classifying the target emotion. We computed MS for AFFDEX and FACET separately for each emotion. Figure 1 depicts the results of Study 1. For an overview of detailed accuracy values see Table C1 in Appendix C[8]. Note that for the values of all emotions, we considered the maximal value of all frames of the "middle" 3 seconds of the stimulus presentation window. This approach of considering the "strongest indication" for a certain emotion follows iMotions guidelines (https://imotions.com/guides/) and should provide the clearest results.

Overall, AFFDEX correctly recognized 73% of the emotions across the three databases. AFFDEX recognized 73% of the emotions in ADFES, 66% of the emotions in WSEFEP and 77% of the emotions in RaFD. In contrast, FACET correctly recognized 97% of the emotions across all the database pictures. FACET recognized 99% of the emotions in ADFES, 92% of the emotions in WSEFEP and 99% of the emotions in RaFD. While AFFDEX failed to detect a face at all in 1% of the pictures, FACET's analysis did not result in any detection failures.

[8] Data and analysis code from both studies is available at: https://github.com/michaelschulte/FacialExpressionAnalysis

As Figure 1 reveals, the algorithms performed differently for different emotions. Both modules performed particularly well for *happy* expressions. AFFDEX showed relatively poor accuracy with the emotions *fear* and *anger.*

Distinctness Index for Emotion Classification. In order to provide evidence on how distinct the matching for emotions (i.e., the MS) is we additionally constructed a *Distinctness Index* (DI). The DI describes how confident the classification is by comparing how close the probability(-like) value of the first predicted emotion is to the probability(-like) value of the second predicted emotion. The DI is defined as the distance from the value of the classified emotion to the value of the next-highest-scoring emotion. Thus, higher DIs indicate a more distinct performance of iMotions's classification and differentiation abilities. We computed average DI separately for all correctly recognized pictures for all emotions for AFFDEX and FACET. We z-transformed the DI, creating a standardized version (sDI) to allow a direct comparison of AFFDEX and FACET.

Table C1 (see Appendix C) summarizes the sDI for iMotions's AFFDEX and FACET for all basic emotions and picture databases. Whereas AFFDEX had an overall sDI of 0.10, FACET had an overall sDI of 0.03. Relatively low sDI (across all databases) for AFFDEX were found for the emotions *anger* and *fear*[9]. Relatively low sDI for FACET were found for the emotions *sadness* and *fear*.

Appendix C provides a confusion matrix of the classification with a detailed overview of true (false) positives and true (false) negatives as well as further performance indices commonly used to assess algorithms in the field of machine learning. Overall, both AFFDEX and FACET relatively infrequently confused happiness, disgust, contempt and surprise. For the other emotions (i.e., anger, fear and sadness), however, AFFDEX and FACET showed a

[9] Note that for fear, we could only compute the DI for ADFES because we had an MS of 0% for WSEFEP and RaFD.

higher confusion prevalence and more pronounced differences between AFFDEX and FACET. Noteworthy, AFFDEX (but not FACET) usually confused fear with surprise (underprediction of fear and overprediction of surprise). Another peculiarity is that AFFDEX often confused anger with sadness (underprediction of anger and overprediction of sadness).

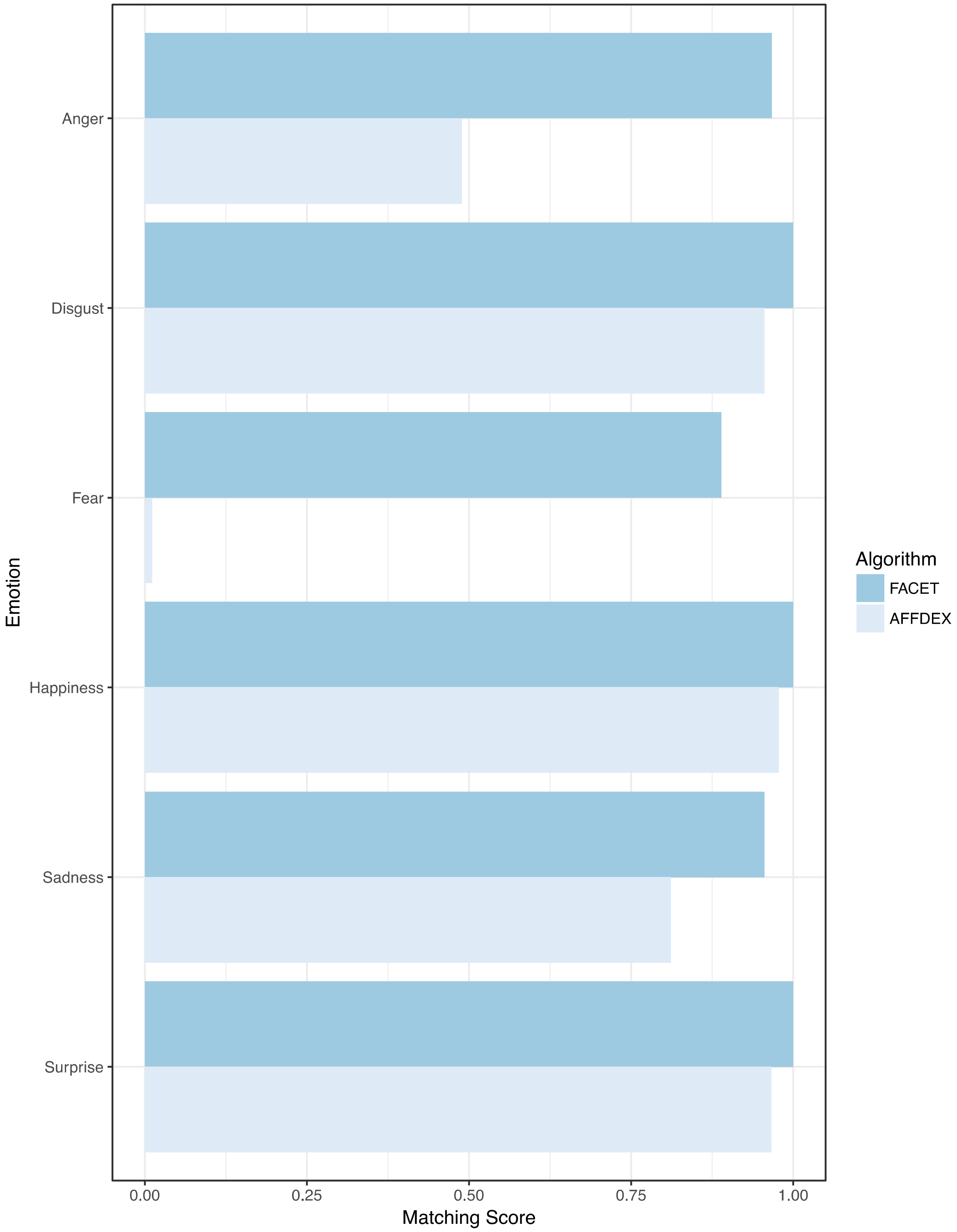

Figure 1. Overview of the non-baseline corrected classification accuracy for basic emotions separately for the iMotions modules AFFDEX and FACET across ADFES, WSEFEP and RaFD. Contempt is not depicted here, since WSEFEP does not provide facial expression pictures for contempt (cf. Appendix C). Note that figures depicting non-baseline corrected data have a blue color code, while figures depicting baseline corrected data have a red color code (cf. Figure C1 in Appendix C).

In addition, we run the previous analysis on baseline corrected data. According to iMotions, baseline corrected data allow more accurate emotion classification than raw (i.e., non-baseline corrected) data. For more details on the rational for baseline correction, see Appendix B. For computational details and results see Appendix C (Figure C1 and Table C4 - Table C6). There are only minor differences between the non-baseline corrected results and the baseline corrected results. For instance, overall accuracy for AFFDEX changed from 73% (non-baseline corrected data) to 72% (baseline corrected data) and for FACET from 97% (non-baseline corrected data) to 95% (baseline corrected data).

Study 1 provides the first evidence regarding iMotions's accuracy in classifying emotions of prototypical facial expressions from a standardized facial expression database. FACET generally outperforms AFFDEX with differences for the employed picture databases and distinct emotions. Given these results, we cannot make any inferences about iMotions's accuracy for natural (vs. prototypical) and dynamic (vs. static) emotional facial expressions.

In order to validate iMotions in a more natural setting with more subtle, dynamic facial expressions, Study 2 employed a validation procedure using human participants, with natural facial expressions. Specifically, first iMotions's accuracy was examined when identifying participants' emotional facial expressions in response to emotional pictures. Second, iMotions's accuracy was examined when identifying emotional facial expressions in participants who were instructed to imitate pictures of facial expressions.

Study 2

Method

Participants. A total of 119 students of a Swiss University participated in this study. Only Caucasian participants without facial artifacts (e.g., disruptive glasses, beards or scarves) were included. Data from 9 participants were excluded from the sample because of missing data, i.e., the software was not able to detect their face (due to technical problems, head movements and/or insufficient video quality). Specifically, participants were excluded from the sample when iMotions failed to generate data for more than 10 percent of all displayed pictures. We considered iMotions to have failed in generating data for a certain picture when it was not possible to detect a participants' face in more than 50 percent of all measurements. For every picture, data from the first 177 frames of the 6 second, 30 Hz video recording was used (because iMotions did not record 180 frames for all pictures). The final sample consisted of 110 participants (63 female; M_{Age} = 21.20, SD_{Age} = 5.20). Three Amazon vouchers, worth CHF 500.-, were raffled among participants.

Design and Procedure. Participants signed a consent form declaring that they agreed to being filmed with a webcam. The study was part of a set of multiple, unrelated studies and always ran first in the session. To ensure good data quality, the laboratory was evenly and clearly lit. Participants were seated in a chair in front of a screen and instructed to remain in a stable and straight position without their hands near their face. Subsequently, the experimenter asked participants to read the description of the study procedure and instructions on the screen. Participants were informed that we were interested in how people respond to pictures that represent various events occurring in daily life. Finally, the written instruction on the screen repeated our oral instruction to remain in a stable position with a straight view on the screen and to avoid to bring their hands close to the face.

In the first part of the study, participants were exposed to two blocks of emotionally evocative pictures (constant block order: *International Affective Picture System* (IAPS) pictures, *Geneva Affective Picture Database* (GAPED) pictures) and their facial expressions were recorded. Within these blocks, pictures were shown in random order. Each picture was presented for 6 seconds and was preceded by a neutral black slide with a white, centrally displayed fixation cross (3 s). Participants were asked to fixate on the cross for the duration of its display. The neutral slides provided baseline measurements for the classification. In the second part of the study, participants were asked to imitate facial expressions for all pictures in the RaFD database for 6 seconds (i.e., as long as every picture was displayed). The RaFD pictures were separately displayed in a random order. Finally, participants were asked for demographics, were thanked and debriefed.

Materials

Emotional Facial Responses to Emotionally Evocative Pictures. In order to capture iMotions's accuracy at detecting participants' emotional facial expressions in response to emotional pictures, we exposed participants to a subset of emotionally evocative pictures from the IAPS[10] and GAPED[11] database. Here, we rely on the assumption that there is coherence between the displayed pictures, participants' emotions, and their facial expressions. IAPS and GAPED pictures are standard stimuli with "positive" and "negative" emotional content used to elicit emotions, or more specifically, pleasure and arousal in experimental research (see Coan & Allen, 2007; Dan-Glauser & Scherer, 2011).

The IAPS database consists of pictures (JPEG format, varying resolution) showing a wide range of emotional content, confirmed to be emotionally evocative (Lang, Bradley, &

[10] The IAPS is freely accessible for non-commercial use at http://csea.phhp.ufl.edu/media/iapsmessage.html
[11] The GAPED is freely accessible for non-commercial use at http://www.affective-sciences.org/en/home/research/materials-and-online-research/research-material/

Cuthbert, 1999). Based on a valence assessment (ranging from *unpleasant* to *pleasant*), we chose four pictures. We chose the pictures with the most distinct (i.e., highest and lowest) valence. The specific picture numbers are: 1710, 1750 (highest valence showing puppies and bunnies); 9940, 9570 (lowest valence showing a hurt dog and an explosion[12]).

The GAPED database consists of pictures (JPEG format, 640 × 480 pixels) that include negative, neutral and positive emotional content (Dan-Glauser & Scherer, 2011). Based on a valence assessment (ranging from *very negative* to *very positive*), we chose two pictures, one with positive content (P067 showing a landscape; highest valence) and one with negative content (A075 showing a cow bleeding to death; lowest valence).

Note that the IAPS and GAPED pictures are appropriate to evoke “positive” or “negative” emotional states but they are not necessarily appropriate to evoke specific emotions such as anger or fear (Bradley, Codispoti, Sabatinelli, & Lang, 2001; Coan & Allen, 2007). Importantly, emotionally evocative stimuli such as IAPS pictures prompt emotional facial muscle activity that relates to evaluative pleasure judgment. For instance, pictures that are perceived as increasingly unpleasant come along with increasing corrugator activity (frown; above eye brow). In contrast, pictures that are perceived as increasingly pleasant come along with decreasing corrugator activity (see Greenwald, Cook, & Lang, 1989; Lang, Greenwald, Bradley, & Hamm, 1993; Larsen, Norris, & Cacioppo, 2003). Overall, the idea that IAPS and GAPED pictures can evoke “positive” or “negative” facial responses, asks for an evaluation to what extend the valence (and not a certain emotion) of participants’ facial responses complies with the pictures’ valence.

Imitation of Facial Expressions. As in Study 1, we used pictures from the RaFD database (Langner et al., 2010). We chose one female face model (female model number 01)

[12] There were concerns about showing the lowest valence pictures (e.g., burn victims). Thus, less disturbing pictures with low valence were chosen.

looking frontal into the camera and showing the basic emotions anger, contempt, disgust, fear, happiness, sadness and surprise. Participants were exposed to the six RaFD pictures and instructed to imitate the currently displayed facial expression.

Setting and Apparatus. We closely followed iMotions's recommendations for experimental setups. For details see the "Definitive guide for facial expression analysis" (https://imotions.com/guides/). The iMotions software (Version 6.2) ran on a Lenovo T450s with Windows 8.1 and an attached 24" (60 cm) BenQ XL2411Z screen to display the pictures. A Logitech C920 webcam (full HD video recording up to 1920 × 1080 pixels and automatic low-light correction) recorded participants' faces. Following iMotions's recommendations we recorded participants with a camera resolution of 640 × 480 pixels. With this apparatus, data (i.e., values for basic emotions) was generated approximately every 32 ms for a total of 177 measurements (frames) for every picture.

Results

Emotional Facial Responses to Emotionally Evocative Pictures. We computed a MS (see Study 1 for details) to estimate the accuracy of classifying the valence of participants' responses to pictures with negative and positive emotionally evocative content. Higher MS values indicate a greater likelihood of correct valence classification. We computed MS separately for AFFDEX and FACET for the positive and negative picture set (IAPS, GAPED pictures).

Prior to computing the MS, we baseline corrected the values generated by iMotions[13]. We did this for the facial responses to all used pictures individually for all participants and separately for all basic emotions. For every participant, we subtracted for every basic emotion the median of the baseline slides' values from all 177 frames of the pictures' values. Based

[13] Non-Baseline corrected results can be found in Appendix D.

on this, we identified the maximal value for all emotions within the 177 measurements for every picture. Finally, these maximal values were used to classify the valence of participants' facial responses as positive or negative. If a maximal value was recorded for happiness, we labeled the facial response as positive. If a maximal value was recorded for anger, contempt, disgust, fear or sadness, we labeled the facial response as negative (in accordance with the valence classification iMotions uses; iMotions, 2016). Surprise was not included in building the valence measures, as iMotions does not consider surprise in their positive/negative aggregate measure[14]. To compute the MS, we identified the number of detected participant faces and the number of correctly labeled facial responses for every picture. We coded participants' facial responses for a certain picture as "correctly labeled" when the assigned valence label for the facial response matched the database's valence label.

Table 1 reveals that AFFDEX classified 57% of all facial responses with the correct valence; it correctly classified 17% of facial responses to positive pictures and 97% of facial responses to negative ones. FACET classified 67% of all facial responses with the correct valence; it correctly classified 63% of facial responses to positive pictures and 71% of facial responses to the negative pictures.

Table 1

Baseline Corrected Classification Accuracy of Valence for iMotions Modules AFFDEX and FACET

<table>
<tr><th rowspan="2">Valence</th><th rowspan="2">Picture</th><th colspan="4">AFFDEX | FACET</th></tr>
<tr><th>Matched</th><th>picturewise MS</th><th>valencewise MS</th><th>Overall MS</th></tr>
<tr><td rowspan="3">Positive</td><td>IAPS 1710</td><td>29 | 77</td><td>0.26 | 0.70</td><td rowspan="3">0.17 | 0.63</td><td rowspan="4">0.57 | 0.67</td></tr>
<tr><td>IAPS 1750</td><td>20 | 65</td><td>0.18 | 0.59</td></tr>
<tr><td>GAPED P067</td><td>8 | 65</td><td>0.07 | 0.59</td></tr>
<tr><td>Negative</td><td>IAPS 9940</td><td>106 | 79</td><td>0.97 | 0.72</td><td></td></tr>
</table>

[14] As surprise has ambiguous valence both positive and negative classifications can be found in the literature (see, e.g., Kim et al., 2004; Neta, Davis, & Whalen, 2011).

IAPS 9570	106 | 74	0.96 | 0.67	0.97 | 0.71
GAPED A075	105 | 80	0.96 | 0.73	

Note. Matched = number of participant faces that match the picture's valence (true positives). MS = Matching Score. While the left side of the vertical bar shows the numbers for AFFDEX, the right side shows the numbers for FACET (AFFDEX | FACET).

Overall, results show that AFFDEX and FACET differ in their accuracy of classifying negatively and positively valenced video recordings of participants displaying emotional expressions. These differences might be considered small when the aggregated, overall measures are compared (57% versus 67%) but drilling down to the picture-wise results these differences grow considerably (e.g., for GAPED P067, AFFDEX: 7% versus FACET: 59%)

We re-ran the present analysis on non-baseline corrected data.[15] As one would expect, due to having real participants generating facial expression, unlike in Study 1 where we used rated pictures, there are more distinct differences between the non-baseline corrected results and the baseline corrected results for FACET. Without baseline correction (vs. with baseline correction), FACET's accuracy is worse for positive valence but better for negative valence: for positive valence FACET's accuracy (MS) changes from 22% (non-baseline corrected data) to 63% (baseline corrected data). For negative valence, FACET's accuracy changes from 92% (non-baseline corrected data) to 71% (baseline corrected data). The overall accuracy (i.e., overall MS) of FACET is worse for non-baselined data (57%) compared to baseline corrected data (67%). Similar to Study 1, AFFDEX showed only marginal differences between the non-baseline corrected results and the baseline corrected results for valence measures - with an overall accuracy of 55% for baseline corrected data and an overall accuracy of 57% for non-baseline corrected data.

[15] We thank two anonymous reviewers for motivating this analysis. Detailed results can be found in Appendix D.

Imitation of Facial Expressions. We computed the MS for estimating iMotions's accuracy when classifying emotions displayed on participants' faces when they imitate the basic emotions displayed in the RaFD pictures. MS is defined as the percentage of participants' imitations that iMotions matched with the correct emotion. We computed MS separately for AFFDEX and FACET for each RaFD picture (see Figure 2). For an overview of detailed accuracy values see Appendix D. We applied the same baseline correction procedure as described above for the valence task.

Table D2 reveals (see Appendix D) that AFFDEX classified 55% and FACET 63% of all facial imitations with the correct emotion. MS differed considerably across emotions (see Figure 2). While both modules were relatively accurate in recognizing posed facial expressions of *happiness* (AFFDEX: 91%; FACET: 98%), they performed poorly for posed facial expressions of *fear* (AFFDEX: 1%; FACET: 10%).

To provide evidence on how distinct these MSs are, we computed standardized DI (sDI) following the procedure described in Study 1. Table D2 (see Appendix D) provides sDIs for all RaFD pictures and both AFFDEX and FACET. Whereas AFFDEX had an overall sDI of 0.02, FACET had an overall sDI of 0.35. For AFFDEX, the lowest sDI related to *fear* and the largest sDI to *contempt*. For FACET, the lowest sDI also related to *fear* and the largest sDI to *happiness*.

Appendix D (Table D3 – Table D4) provides the confusion matrix of the classification as well as further performance indices to assess AFFDEX and FACET. Overall, AFFDEX and FACET differed in their confusion prevalence across different emotions. Noteworthy, AFFDEX's and FACET's highest (lowest) confusion prevalence was found for fear (happiness): While AFFDEX usually confused fear with surprise or contempt (underprediction of fear and overprediction of surprise and contempt), AFFDEX rarely confused happiness. Similarly, FACET usually confused fear with surprise (underprediction of fear and overprediction of surprise), but rarely confused happiness.

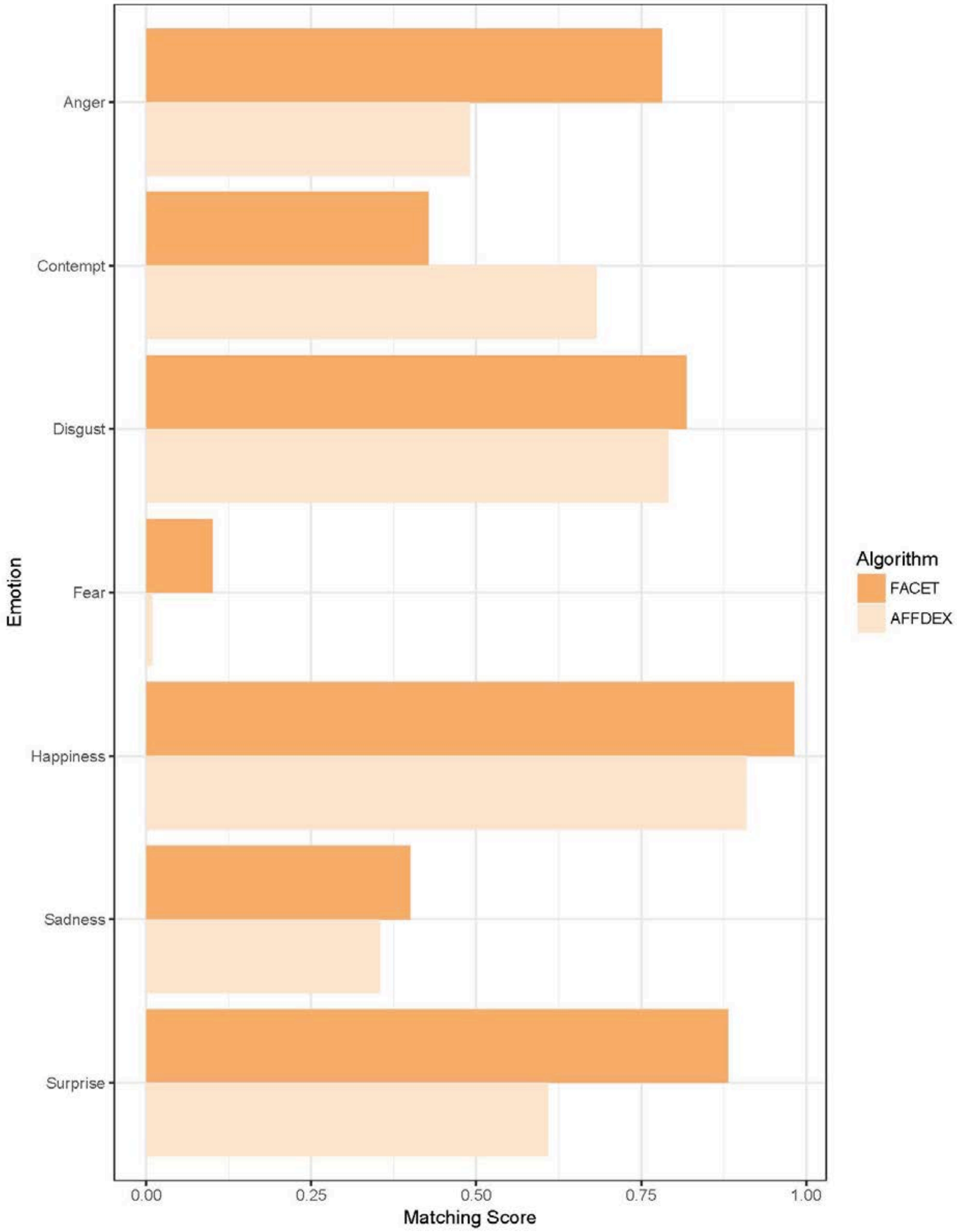

Figure 2. Overview of the baseline corrected classification accuracy for basic emotions separately for the iMotions modules AFFDEX and FACET. Note that figures depicting

baseline corrected data have a red color code while figures depicting non-baseline corrected data have a blue color code (cf. Figure D1 in Appendix D).

Additionally, we re-ran the present analysis on non-baseline corrected data. Detailed results can be found in Appendix D (Figure D1 and Table D5 – Table D7). Similar to Study 1, there are only minor overall differences between the non-baseline corrected results and the baseline corrected results. The overall accuracy (i.e., overall MS) for AFFDEX differs only little (51% non-baseline corrected data versus 55% baseline corrected data) and differs not at all for FACET (63% accuracy for non-baseline corrected and baseline corrected data). Yet, a closer look at the results reveals that differences between non-baseline corrected results and baseline corrected results for AFFDEX and FACET are more pronounced for some emotions (e.g., sadness, fear) than for others (e.g., happiness, disgust).

Study 2 provides the first evidence regarding iMotions's accuracy in classifying emotions in natural and dynamic emotional facial expressions within a laboratory setting. Compared to iMotions's accuracy for classifying standardized, prototypical facial expression pictures (Study 1), Study 2 reveals reduced accuracy for people's natural facial responses to diverse emotionally evocative pictures. The accuracy of iMotions differs for distinct emotions (and valence), and is generally higher for FACET than for AFFDEX.

General Discussion

This research validates iMotions's facial expression analysis modules AFFDEX and FACET as software-based tools for emotion classification. When identifying prototypical facial expressions from three picture databases (Study 1), we find overall accuracy of 73% for AFFDEX and 97% for FACET. When using participants instead of prototypical pictures, accuracy drops for the valence of people's facial responses to diverse emotionally evocative pictures (55% for AFFDEX, 57% for FACET; Study 2). Taken together, iMotions's performance is better for recognizing prototypical static versus more natural dynamic facial expressions, and shows different results for distinct emotions (and valence). Overall, FACET outperforms AFFDEX on nearly all measures.

Validation and Comparison of iMotions (AFFDEX and FACET)

This research contributes by independently measuring and comparing the performance of iMotions's AFFDEX and FACET modules, and making the results publicly available for a broad audience. In general, there is support for the idea that automated facial expression analysis is technically feasible (e.g., Baltrusaitis et al., 2016; Bartlett, Hager, Ekman, & Sejnowski, 1999; Lien, Kanade, Cohn, & Li, 1998; Littlewort, Bartlett, Fasel, Susskind, & Movellan, 2006; Meiselman, 2016; Vallverdu, 2014). Moreover, it is evident that automated facial expression analysis (e.g., Noldus's FaceReader) can produce valid data for prototypical facial expressions that are recorded under standardized conditions (Lewinski et al., 2014; Littlewort et al., 2006; Valstar et al., 2011).

The present findings support the skepticism that current automated facial expression analysis is not yet mature enough for operational use (Meiselman, 2016) by revealing that, while iMotions's automated facial expression analysis can produce data with an acceptable degree of accuracy for prototypical facial expressions, it is less accurate for subtle, more natural facial expressions.

Accuracy measures for AFFDEX and FACET show that iMotions can provide data as valid as that produced by human judges. Human performance in recognizing emotions in prototypical facial expressions in database pictures is often situated between 60% and 80% and normally does not attain 90% accuracy (Nelson & Russell, 2013). Human judges are usually better at selecting the correct emotion label for *happy* than for other emotional facial expressions. When discriminating between non-happy expressions (i.e., anger, disgust, fear, sadness, surprise), judges' accuracy in recognizing emotions is particularly weak for fearful faces (Calvo et al., 2014; Nelson & Russell, 2013). Testing iMotions's accuracy (on similar pictures of prototypical emotions; Study 1) reveals comparable performance to human judges. One can also compare the performance of human judges and iMotions for identical sets of facial expressions. For the WSEFEP and ADFES databases, human judges have a performance of 85% (see Lewinski et al., 2014; Olszanowski et al., 2015; van der Schalk et al., 2011). The performances of the AFFDEX and FACET modules are 70% and 96% respectively (Study 1)[16]. While AFFDEX's accuracy is in the middle of the range of the accuracy of human judges (i.e., 60–80%), FACET's accuracy seems to outperform human judges. Moreover, results show that, like human judges, iMotions's accuracy differs for distinct emotions and performs particularly well (poorly) for happy (fearful) faces.

A comparison of iMotions's automated facial expression analysis modules with Noldus's FaceReader leads to similar inferences. Lewinski et al., (2014) found FaceReader to correctly classify 88% of emotions in the WSEFEP and ADFES pictures. According to the results of Study 1, iMotions's AFFDEX shows lower performance (70%) than Noldus's FaceReader; however, FACET outperforms Noldus's FaceReader (96% vs. 88%). These results can be due to various characteristics of the two algorithms such as the different number of facial landmarks: 6 (FACET) vs. 34 (AFFDEX). It is important to consider that the present comparison of the performance of Noldus's FaceReader and iMotions's AFFDEX and FACET could also be biased because producers do not use the databases in the algorithm's training set. If one facial expression analysis engine, but not the others, includes WSEFEP or ADFES in the machine learning process, then this will result in an overestimated relative accuracy. More comprehensive specifications of the different training sets would help to solve this issue.

[16] As per Lewinski et al. 2014, we computed unweighted MS for the AFFDEX and FACET module based on the non-baselined MS for the ADFES and WSEFEP.

Regarding a direct comparison of the validity of automated facial expression analysis with human FACS coders, two problems arise. First, automated facial expression analysis is based on FACS and uses FACS classified pictures as training database. Second, FACS coders primarily describe AUs (i.e., anatomically independent facial muscle movements) and do not directly measure emotions. Looking into the literature reveals that many studies on FACS coder accuracy focus on performance on certain AUs rather than on emotion classification (cf. Lewinski et al., 2014). Clearly, certain AU configurations are associated with certain basic emotions. Such predictions of emotions, however, involve comprehensive definitions of AU configurations and consistent decisions on which (variants of prototypical) AU configurations account for a certain basic emotion. This makes direct comparisons unreliable.

A secondary contribution of this validation study is that it provides a comprehensive comparison of baseline correction approaches. Overall, there are only marginal differences between the non-baseline corrected results and the baseline corrected results, yet, these differences varied for AFFDEX and FACET and were more pronounced for certain emotions (e.g., contempt, disgust) than for others (e.g., happiness).

Limitations of the Present Research

The standardized and controlled setting may impede generalizability of our results. Study 1 classifies prototypical, static facial expressions that are uncommon in real-life situations. Accuracy measures are thus likely to be inflated. Study 2 partially addresses this limitation by using more natural, dynamic facial expressions within a controlled laboratory setting. Still, real-life settings differ from laboratory settings in motion and uneven light and color.

We also build on the assumption that positive (negative) pictures elicit positive (negative) facial responses. This assumption, however, is controversial. Facial expressions occur for various reasons: they can be generated internally (e.g., by thoughts or memories), produced by external stimuli (e.g., photographs or films; Reiman et al., 1997) and be determined by social interaction and display rules (Vallverdu, 2014). Further, positive (negative) stimuli do not only produce positive (negative) facial expressions but also expressions that are reserved for negative (positive) emotions or a mix of diverse emotions (Aragón, Clark, Dyer, & Bargh, 2015; Fredrickson & Levenson, 1998). We thus cannot be sure whether our positive (negative) pictures were actually effective in eliciting the intended valence in participants' faces. This ties into the finding that iMotions's performance is better at recognizing negative versus positive facial expressions. It is important to refer to a bias that is introduced by iMotions valence classification: According to this classification, positive valence is recorded for happiness and negative valence for anger, contempt, disgust, fear and sadness (iMotions, 2016). Hence, simple probability (i.e., positive valence is only recorded for one emotion while negative valence is recorded for five emotions) calls into question the conclusion that iMotions's performance is better for negative versus positive facial expressions.

Regarding our choice of emotionally evocative pictures, it is also worth mentioning that emotion researchers increasingly use dynamic film stimuli (vs. static picture stimuli). Indeed, dynamic stimuli showed to be more powerful in evoking emotional responses. This is because dynamic stimuli are more realistic and complex (see, e.g., Manera, Samson, Pehrs, Lee, & Gross, 2014; Schlochtermeier, Pehrs, Kuchinke, Kappelhoff, & Jacobs, 2015). Given that we exclusively used static stimuli to evoke emotional responses in Study 2, we cannot rule out that our results are biased due to inadequate emotion induction.

A second limitation of Study 2 is that we rely on the assumption that participants can imitate pictures of emotional facial expressions. In fact, we do not know how accurately participants imitated the displayed facial expressions. Results of Study 2 could therefore be confounded by limitations in participants' ability to imitate emotions accurately; we cannot rule out that iMotions would actually perform better.

A third limitation of Study 2 arises from evidence that people more likely respond to negative stimuli compared to positive stimuli (e.g., IAPS pictures or pictures of faces with different emotional expressions). Different studies found shorter latencies as well as higher amplitudes in response to negative pictures than to positive ones (e.g., Carretié, Mercado, Tapia, & Hinojosa, 2001; Gotlib, Krasnoperova, Yue, & Joormann, 2004; Huang & Luo, 2006; Öhman, Lundqvist, & Esteves, 2001). Results of Study 2 could thus be biased by participant's general sensitivity to emotionally negative (vs. positive) stimuli.

Overall, these limitations substantiate the need to improve the application of automated facial expression analysis in real life settings. It is thus not surprising that affective computing researchers are currently addressing issues such as varying camera angles and changing head poses. Improvements are also needed in analyzing non-posed faces, the sensitivity of measuring subtle changes in facial expressions and the discrimination of more difficult expressions (i.e., compound emotions) and expression intensity (see, e.g., Facial Expression Recognition and Analysis challenge 2015 (www.sspnet.eu/fera2015/) and 2017 (www.sspnet.eu/fera2017/); McDuff, et al., 2010). In view of the steady improvements of the validity of automated facial expression analysis in real-world settings, it will be a useful exercise to continually validate iMotions as well as other providers, particularly in real-world settings.

From a theoretical viewpoint, limitations become apparent when interpreting the present results under consideration of the ongoing debate about an appropriate theory for automated facial expression analysis. Automated facial expression analysis tools typically generate probability(-like) measures for distinct basic emotions and are trained with databases of prototypical facial expressions. Not surprisingly, these tools are often successful with prototypical facial expressions (Lewinski et al., 2014; Vallverdu, 2014). This prototypical perspective, however, is problematic as it limits the generalizability of automated facial expression analysis. There are many types of facial expressions that vary in their distinctness and intensity, ranging from subtle to very intense (Ekman, Friesen & Ancoli, 1980; Hess, Banse, & Kappas, 1995). In the present research, we did not distinguish between measuring prototypical versus natural facial expressions; i.e., Study 1 and Study 2 were not designed for direct comparison of iMotions's accuracy. Nevertheless, it seems unsurprising that the present research found higher accuracy when classifying posed, intense facial expressions (Study 1) rather than subtle, more natural facial expressions (Study 2). Future validation of iMotions is needed to systematically test its accuracy for prototypical facial expressions versus more natural facial expressions. One possibility to address this is to use existing face databases of more natural facial expressions (see, e.g., face database by McEwan et al., 2014).

Due to the current basic emotion perspective of automated facial expression analysis, it is often ignored that cultural and contextual aspects can be essential for the classification of expressed emotions (see, e.g., Aviezer, Trope, & Todorov, 2012; Barrett, Mesquita, & Gendron, 2011; Elfenbein & Ambady, 2002). Further, real life facial expressions are rarely prototypical and rather reflect compound (vs. distinct) emotions, i.e., combinations of single components of basic emotions (e.g., Du, Tao, & Martinez, 2014; Naab & Russel, 2007; Scherer & Ellgring, 2007). People often experience and express emotional states that cannot be assigned to only one basic emotion (Scherer, Wranik, Sagsue, Tran & Scherer, 2004). There is considerable evidence showing that there are different degrees of dissimilarity between facial expressions (of different basic emotions). As previous research (e.g., Wegrzyn, Vogt, Kireclioglu, Schneider, & Kissler, 2017) and our confusion matrices suggest, happiness seems to belong to the most distinctively expressed, i.e. least confused emotions. In contrast, emotions such as fear and surprise seem to be more similar, i.e. more frequently confused. Clearly, these confusions occur because facial expressions (of different basic emotions) vary in the extent with which they overlap in their AU patterns. For instance, fear as well as surprise are characterized by raised eyebrows and eyelids (Hager, Ekman, & Friesen, 2002; Wegrzyn, Vogt, Kireclioglu, Schneider, & Kissler, 2017).

Another aspect that questions the basic emotion perspective is that people can use facial expressions to regulate their emotional feeling states by altering outward facial expressions. Sometimes it can be useful for people to hide or suppress facial expressions in order to portray external facial expressions that don’t reflect internal feeling states (Gross, 2002).

Taken together, various cultural and contextual aspects add to the complexity of analyzing facial expressions. In order to more realistically relate facial expressions to underlying emotional processes, automated facial expression analysis could adopt an appraisal perspective, i.e., consider cultural and contextual aspects (Barrett & Wager, 2006; Ekman, 1992b; Ortony & Turner, 1990; Russell, 2003; Scherer, 2005).

Implications for Researchers and Practitioners

There are various approaches to measuring emotions, from verbal ratings to nonverbal indicators. The advantages of automated facial expression analysis are low time and labor costs, simplicity and the potential for less intrusive measurements (see iMotions, 2016; Meiselman, 2016). Thus, valid automated facial expression analysis offers opportunities in diverse fields of emotion research, not only for academics but also for practitioners such as marketers or IT providers. In the future, academics could use such tools to efficiently validate new databases of prototypical basic emotional expressions. The commercial application of such tools, for example in smartphones, media and advertisement testing, or even the design of avatars, has recently become pronounced (see iMotions, 2016; Lee, Sang Choi, Lee, & Park, 2012).

In view of this need for valid facial expression analysis tools, it would be advantageous if providers of automated facial expression analysis would not only improve the validity of their products further, but also provide transparent and complete product information that complies with scientific requirements. For instance, development and algorithmic details should be clear and sufficiently documented; the databases on which the algorithms are trained should be specified; and details on the generation and interpretation of data, as well as on the validity of this data, should be available.

We encourage researchers to define and apply standard methods to validate and compare automated facial expression analysis tools. The present accuracy measures, for instance, could be used to (re-)validate (updated) automated facial expression analysis tools in a standardized manner. To a certain extent, these accuracy measures could also serve to compare automated facial expression analysis with other measurement methods.

Note that comprehensive validation of facial expression analysis tools also provides fundamental information for computer scientists to improve facial expression analysis algorithms. Thus, we encourage the developers of AFFDEX and FACET to use the present performance indices and confusion matrices to improve their algorithms. For instance, the present confusion matrices imply that one future contribution of the developers of AFFDEX and FACET should be to improve the discrimination of the facial expressions of fear and surprise. However, as mentioned earlier, the increased confusion of certain emotions (e.g., fear and surprise) might be inherent in nature of emotions that share more or less common (AU) patterns.

Conclusion

Two validation studies reveal that iMotions has the potential to measure basic emotions expressed by faces. iMotions performs better for prototypical versus natural facial expressions, and shows different results depending on the studied emotion. iMotions's FACET module outperforms the AFFDEX module.

References

Aragón, O. R., Clark, M. S., Dyer, R. L., & Bargh, J. A. (2015). Dimorphous expressions of positive emotion: Displays of both care and aggression in response to cute stimuli. *Psychological Science*, *26*(3) 1–15.

Aviezer, H., Trope, Y., & Todorov, A. (2012). Holistic person processing: Faces with bodies tell the whole story. *Journal of Personality and Social Psychology, 103*(1), 20–37.

Baltrusaitis, T., Robinson, P., & Morency, L.-P. (2016). OpenFace: An open source facial behavior analysis toolkit. Proceedings from *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)* (pp. 1–10). IEEE.

Barrett, L. F., Mesquita, B., & Gendron, M. (2011). Context in emotion perception. *Current Directions in Psychological Science*, *20*(5), 286–290.

Barrett L. F., & Wager T. D. (2006). The structure of emotion: evidence from neuroimaging studies. *Current Directions in Psychological Science, 15*, 79–83. doi: 10.1111/j.0963-7214.2006.00411

Bartlett, M. S., Hager, J. C., Ekman, P., & Sejnowski, T. J. (1999). Measuring facial expressions by computer image analysis. *Psychophysiology*, *36*(02), 253–263.

Beumer, G. M., Tao, Q., Bazen, A. M., & Veldhuis, R. N. (2006). A landmark paper in face recognition. In *7th International Conference on Automatic Face and Gesture Recognition*, (pp. 73–78). IEEE.

Bonanno, G., & Keltner, D. (2004). Brief Report: The coherence of emotion systems: Comparing "on-line" measures of appraisal and facial expressions, and self-report. *Cognition and Emotion*, *18*(3), 431–444.

Calvo, M. G., Gutiérrez-García, A., Fernández-Martín, A., & Nummenmaa, L. (2014). Recognition of facial expressions of emotion is related to their frequency in everyday life. *Journal of Nonverbal Behavior*, *38*(4), 549–567.

Carretié, L., Mercado, F., Tapia, M., & Hinojosa, J. A. (2001). Emotion, attention, and the ‘negativity bias’, studied through event-related potentials. *International Journal of Psychophysiology*, *41*(1), 75–85.

Coan, J. A., & Allen, J. J. (Eds.). (2007). *Handbook of emotion elicitation and assessment*. Oxford, UK: Oxford University Press.

Cohn, J. F., & Sayette, M. A. (2010). Spontaneous facial expression in a small group can be automatically measured: An initial demonstration. *Behavior Research Methods*, *42*(4), 1079–1086.

Cootes, T. F., Edwards, G. J., & Taylor, C. J. (2001). Active appearance models. *IEEE Transactions on pattern analysis and machine intelligence*, *23*(6), 681–685.

Dan-Glauser, E. S., & Scherer, K. R. (2011). The Geneva affective picture database (GAPED): a new 730-picture database focusing on valence and normative significance. *Behavior Research Methods*, *43*(2), 468–477.

den Uyl, M. J., & van Kuilenburg, H. (2005). The FaceReader: Online facial expression recognition. In *Proceedings of Measuring Behavior 2005* (pp. 589–590).

Du, S., Tao, Y., & Martinez, A. M. (2014). Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences*, *111*(15), 1454–1462.

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

Ekman, P. (1992b). Are there basic emotions? *Psychological Review*, *99*(3), 550-553.

Ekman, P. (1992c). Facial expressions of emotion: New findings, new questions. *Psychological Science*, *3*(1), 34–38.

Ekman, P., & Friesen, W. V. (1976). Measuring facial movement. *Environmental Psychology and Nonverbal Behavior*, *1*(1), 56–75.

Ekman, P., & Friesen, W. V. (1982). Felt, false, and miserable smiles. *Journal of Nonverbal Behavior*, *6*(4), 238–252.

Ekman, P., Friesen, W. V., & Ancoli, S. (1980). Facial signs of emotional experience. *Journal of Personality and Social Psychology*, *39*(6), 1125–1134.

Ekman, P., & Friesen, W. V. (2003). *Unmasking the face: A guide to recognizing emotions from facial clues*. Cambridge, MA: Malor Books.

Ekman, P., Friesen, W. V., O'Sullivan, M., Chan, A., Diacoyanni-Tarlatzis, I., Heider, K. … Tzavaras, A. (1987). Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of Personality and Social Psychology*, *53*(4), 712–717.

Ekman, P., & Oster, H. (1979). Facial expressions of emotion. *Annual Review of Psychology*, *30*(1), 527–554.

Elfenbein, H. A., & Ambady, N. (2002). On the universality and cultural specificity of emotion recognition: a meta-analysis. *Psychological Bulletin*, *128*(2), 203-235.

El Kaliouby, R., & Robinson, P. (2005). Real-time inference of complex mental states from facial expressions and head gestures. In B. Kisačanin, V. Pavlović & T. S. Huang (Eds), *Real-time vision for human-computer interaction* (pp. 181–200). New York: Springer.

Ellsworth, P. C., & Scherer, K. R. (2003). Appraisal processes in emotion. In R. J. Davidson, K. R. Scherer, & H. H. Goldsmith (Eds.), *Handbook of affective sciences* (pp. 572–595). New York: Oxford University Press.

Fredrickson, B. L., & Levenson, R. W. (1998). Positive emotions speed recovery from the cardiovascular sequelae of negative emotions. *Cognition & Emotion*, *12*(2), 191–220.

Gotlib, I. H., Krasnoperova, E., Yue, D. N., & Joormann, J. (2004). Attentional biases for negative interpersonal stimuli in clinical depression. *Journal of Abnormal Psychology*, *113*(1), 127–135.

Greenwald, M. K., Cook, E. W., & Lang, P. J. (1989). Affective judgment and psychophysiological response: Dimensional covariation in the evaluation of pictorial stimuli. *Journal of Psychophysiology*, *3*(1), 51–64.

Gross, J. J. (2002). Emotion regulation: Affective, cognitive, and social consequences. *Psychophysiology*, *39*(3), 281–291.

Hager J. C., P. Ekman, & W. V. Friesen (2002). Facial action coding system. Salt Lake City, UT: A Human Face.

Hess, U., Banse, R., & Kappas, A. (1995). The intensity of facial expression is determined by underlying affective state and social situation. *Journal of Personality and Social Psychology*, *69*(2), 280–288.

Huang, C.-N., Chen, C.-H., & Chung, H.-Y. (2004). The review of applications and measurements in facial electromyography. *Journal of Medical and Biological Engineering*, *25*(1), 15–20.

Huang, Y.-X., & Luo, Y.-J. (2006). Temporal course of emotional negativity bias: An ERP study. *Neuroscience Letters*, *398*(1), 91–96. https://doi.org/10.1016/j.neulet.2005.12.074

Hwang, H. C., & Matsumoto, D. (2016). Facial expressions. In D. Matsumoto, H. C. Hwang, & M. G. Frank (Eds.), *APA handbook of nonverbal communication* (pp. 257–287). Washington, DC, US: American Psychological Association. https://doi.org/10.1037/14669-010

iMotions (2016). *Facial Expression Analysis: The definitive guide*. Retrieved from https://imotions.com/facialexpression-guide-ebook/

Kim, H., Somerville, L. H., Johnstone, T., Polis, S., Alexander, A. L., Shin, L. M., & Whalen, P. J. (2004). Contextual modulation of amygdala responsivity to surprised faces. *Journal of Cognitive Neuroscience*, *16*(10), 1730–1745.

Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D. H., Hawk, S. T., & van Knippenberg, A. (2010). Presentation and validation of the Radboud Faces Database. *Cognition and Emotion*, *24*(8), 1377–1388.

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (1999). *International affective picture system (IAPS): Technical manual and affective ratings*. Gainesville, FL: The Center for Research in Psychophysiology, University of Florida. Retrieved from http://www2.unifesp.br/dpsicobio/Nova_versao_pagina_psicobio/adap/instructions.pdf

Lang, P. J., Greenwald, M. K., Bradley, M. M., & Hamm, A. O. (1993). Looking at pictures: Affective, facial, visceral, and behavioral reactions. *Psychophysiology*, *30*(3), 261–273. https://doi.org/10.1111/j.1469-8986.1993.tb03352.x

Larsen, J. T., Norris, C. J., & Cacioppo, J. T. (2003). Effects of positive and negative affect on electromyographic activity over zygomaticus major and corrugator supercilii. *Psychophysiology*, *40*(5), 776–785.

Lee, H., Choi, Y. S., Lee, S., & Park, I. P. (2012). Towards unobtrusive emotion recognition for affective social communication. In *Consumer Communications and Networking Conference (CCNC), 2012 IEEE* (pp. 260-264). IEEE. doi: 10.1109/CCNC.2012.6181098

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

Lien, J. J., Kanade, T., Cohn, J. F., & Li, C.-C. (1998). Automated facial expression recognition based on FACS action units. Proceedings from *Third IEEE International Conference on Automatic Face and Gesture Recognition* (pp. 390–395). IEEE. Retrieved from http://ieeexplore.ieee.org/abstract/document/670980/

Littlewort, G., Bartlett, M. S., Fasel, I., Susskind, J., & Movellan, J. (2006). Dynamics of facial expression extracted automatically from video. *Image and Vision Computing, 24*(6), 615–625. doi: 10.1016/j.imavis.2005.09.011

Littlewort, G. C., Bartlett, M. S., & Lee, K. (2009). Automatic coding of facial expressions displayed during posed and genuine pain. *Image and Vision Computing*, *27*(12), 1797–1803.

Littlewort, G., Whitehill, J., Wu, T., Fasel, I., Frank, M., Movellan, J., & Bartlett, M. (2011). The computer expression recognition toolbox (CERT). Proceedings from *2011 IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG 2011),* (pp. 298–305). IEEE. https://doi.org/10.1109/FG.2011.5771414

Manera, V., Samson, A. C., Pehrs, C., Lee, I. A., & Gross, J. J. (2014). The eyes have it: The role of attention in cognitive reappraisal of social stimuli. *Emotion*, *14*(5), 833–900.

McDuff, D., El Kaliouby, R., Kassam, K., & Picard, R. (2010). Affect valence inference from facial action unit spectrograms. Proceedings from *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition — Workshops* (pp. 17–24). IEEE. https://doi.org/10.1109/CVPRW.2010.5543833

McDuff, D., El Kaliouby, R., Cohn, J. F., & Picard, R. W. (2015). Predicting ad liking and purchase intent: Large-scale analysis of facial responses to ads. *Affective Computing, IEEE Transactions*, *6*(3), 223–235.

McDuff, D. (2016). Discovering Facial Expressions for States of Amused, Persuaded, Informed, Sentimental and Inspired. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction* (pp. 71–75). New York, NY, USA: ACM. https://doi.org/10.1145/2993148.2993192

McEwan, K., Gilbert, P., Dandeneau, S., Lipka, S., Maratos, F., Paterson, K. B., & Baldwin, M. (2014). Facial expressions depicting compassionate and critical emotions: The development and validation of a new emotional face stimulus set. *PloS one, 9*(2), 1–8.

Meiselman, H. L. (2016). *Emotion measurement*. Cambridge, UK: Woodhead.

Mortillaro, M., Meuleman, B., & Scherer, K. R. (2015). Automated Recognition of Emotion Appraisals. In J. Vallverdu (Ed.), *Handbook of Research on Synthesizing Human Emotion in Intelligent Systems and Robotics* (pp. 338-351). Hershey, PA: IGI Global.

Naab, P. J., & Russell, J. A. (2007). Judgments of emotion from spontaneous facial expressions of New Guineans. *Emotion*, 7, 736–744.

Nelson, N. L., & Russell, J. A. (2013). Universality revisited. *Emotion Review*, *5*(1), 8–15.

Neta, M., Davis, F. C., & Whalen, P. J. (2011). Valence resolution of ambiguous facial expressions using an emotional oddball task. *Emotion, 11*(6), 1425–1433.

Öhman, A., Lundqvist, D., & Esteves, F. (2001). The face in the crowd revisited: a threat advantage with schematic stimuli. *Journal of Personality and Social Psychology, 80*(3), 381–396.

Olszanowski, M., Pochwatko, G., Kuklinski, K., Scibor-Rylski, M., Lewinski, P., & Ohme, R. K. (2015). Warsaw set of emotional facial expression pictures: A validation study of facial display photographs. *Frontiers in Psychology*, *5,* 1–8.

Ortony, A., & Turner, T. J. (1990). What's basic about basic emotions?. *Psychological Review*, *97*(3), 315–331.

O'Toole, A. Phillips, P. J. Narvekar, A. Jiang, F. Ayyad, J. (2008). Face recognition algorithms and the “other-race” effect. *Journal of Vision, 8*(6). https://journalofvision.org/8/6/256/

Reiman, E. M., Lane, R. D., Ahern, G. L., Schwartz, G. E., Davidson, R. J., Friston, K. J., … Chen, K. (1997). Neuroanatomical correlates of externally and internally generated human emotion. *American Journal of Psychiatry*, *154*(7), 918–925.

Reisenzein, R., Studtmann, M., & Horstmann, G. (2013). Coherence between emotion and facial expression: Evidence from laboratory experiments. *Emotion Review*, *5*(1), 16–23. https://doi.org/10.1177/1754073912457228

Roseman, I. J., & Smith, C. A. (2001). Appraisal theory. Appraisal theory: Overview, assumptions, varieties, controversies. In K. R. Scherer, A. Schorr & T. Johnstone (Eds.), *Appraisal processes in emotion* (pp. 3–19). Oxford, UK: Oxford University Press.

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological Review, 110*(1), 145–172.

Scherer, K. R., & Ellgring, H. (2007). Are facial expressions of emotion produced by categorical affect programs or dynamically driven by appraisal? *Emotion*, 7, 113–130.

Scherer, K. R., Wranik, T., Sangsue, J., Tran, V., & Scherer, U. (2004). Emotions in everyday life: Probability of occurrence, risk factors, appraisal and reaction patterns. *Social Science Information, 43*, 499–570.

Scherer, K. R. (2005). What are emotions? And how can they be measured?. *Social Science Information, 44*(4), 695–729.

Schlochtermeier, L. H., Pehrs, C., Kuchinke, L., Kappelhoff, H., & Jacobs, A. M. (2015). Emotion processing in different media types: realism, complexity and immersion. *Journal of Systems and Integrative Neuroscience, 1*, 41-47.

Schulte-Mecklenbeck, M., Johnson, J.G., Böckenholt, U., Goldstein, D., Russo, J., Sullivan, N., & Willemsen, M. (in press). Process tracing methods in decision making: On growing up in the 70ties. *Current Directions in Psychological Science*.

Stets, J. E., & Turner, J. H. (2014). *Handbook of the Sociology of Emotions* (Vol. 2). Heidelberg, Germany: Springer.

Swinton, R., & El Kaliouby, R. (2012). *Measuring emotions through a mobile device across borders, ages, genders and more*. ESOMAR. Retrieved from http://www.affectiva.com/wp-content/uploads/2014/09/Measuring_Emotion_Through_Mobile_Esomar.pdf

Taggart, R. W., Dressler, M., Kumar, P., Khan, S., & Coppola, J. F. (n.d.). Determining emotions via facial expression analysis software. Retrieved from http://csis.pace.edu/~ctappert/srd2016/2016PDF/c2.pdf

Terzis, V., Moridis, C. N., & Economides, A. A. (2010). Measuring instant emotions during a self-assessment test: the use of FaceReader. In *Proceedings of the 7th International Conference on Methods and Techniques in Behavioral Research* (p. 18). ACM. Retrieved from http://dl.acm.org/citation.cfm?id=1931362

Vallverdu, J. (2014). *Handbook of research on synthesizing human emotion in intelligent systems and robotics*. Hershey PA, USA: IGI Global.

Valstar, M. F., Jiang, B., Mehu, M., Pantic, M., & Scherer, K. (2011). The first facial expression recognition and analysis challenge. Proceedings from *2011 IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG 2011),* (pp. 921–926). IEEE. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5771374

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.

van Kuilenburg, H., Wiering, M., & den Uyl, M. (2005). A model based method for automatic facial expression recognition. In *European Conference on Machine Learning* (pp. 194–205). Springer. DOI: 10.1007/11564096_22

Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, *1*(1), 1–9.

Wegrzyn M, Vogt M, Kireclioglu B, Schneider J, & J. Kissler (2017). Mapping the emotional face. How individual face parts contribute to successful emotion recognition. *PLoS ONE*, *12*(5). DOI: 10.1371/journal.pone.0177239

Wolf, K. (2015). Measuring facial expression of emotion. *Dialogues in Clinical Neuroscience*, *17*(4), 457–462.

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