The Visual Computer
https://doi.org/10.1007/s00371-020-01881-x

ORIGINAL ARTICLE

Benchmarking commercial emotion detection systems using realistic distortions of facial image datasets

Kangning Yang[1] · Chaofan Wang[1] · Zhanna Sarsenbayeva[1] · Benjamin Tag[1] · Tilman Dingler[1] · Greg Wadley[1] · Jorge Goncalves[1]

Abstract

Currently, there are several widely used commercial cloud-based services that attempt to recognize an individual's emotions based on their facial expressions. Most research into facial emotion recognition has used high-resolution, front-oriented, full-face images. However, when images are collected in naturalistic settings (e.g., using smartphone's frontal camera), these images are likely to be far from ideal due to camera positioning, lighting conditions, and camera shake. The impact these conditions have on the accuracy of commercial emotion recognition services has not been studied in full detail. To fill this gap, we selected five prominent commercial emotion recognition systems—Amazon Rekognition, Baidu Research, Face++, Microsoft Azure, and Affectiva—and evaluated their performance via two experiments. In Experiment 1, we compared the systems' accuracy at classifying images drawn from three standardized facial expression databases. In Experiment 2, we first identified several common scenarios (e.g., partially visible face) that can lead to poor-quality pictures during smartphone use, and manipulated the same set of images used in Experiment 1 to simulate these scenarios. We used the manipulated images to again compare the systems' classification performance, finding that the systems varied in how well they handled manipulated images that simulate realistic image distortion. Based on our findings, we offer recommendations for developers and researchers who would like to use commercial facial emotion recognition technologies in their applications.

Keywords Affective computing · Facial emotion recognition · Commercial emotion recognition systems · Non-ideal conditions · Validation analysis

1 Introduction

A basic goal of most affective computing systems is to automatically recognize and react to human affective states during interactions with human beings [18]. To this end, emotion recognition is one of the most important sub-tasks of affective computing [79,85] and has already been used in many real-world applications, including mental health systems [112], education systems [102,105], and music recommendation systems [21]. Furthermore, emotions are known to play a crucial role in human daily lives, as they can aid decision-making, learning, communication, and situation awareness in human-centric environments [85]. Thus, reliable emotion recognition is seen as crucial for several other research fields including, but not limited to, psychology, social sciences, and computer science.

Over the past few years, a variety of data sources including facial expressions, voice, body gestures, and linguistics has been leveraged for emotion recognition by researchers and industry [53]. Notwithstanding, given that facial expressions are the primary communication mode of non-verbal emotional expressions [31,68], this particular data source is still one of the preferred choices in the emotion recognition domain. Recently, with the increasing availability of 3D sensors and corresponding devices, researchers are able to acquire 3D dynamic facial expressions [11,12]. This methodology has the potential to provide more robust outcomes in practical applications [19,107].

In our work, we focus on emotion recognition systems that use 2D facial expressions as their data source. Such systems have been widely applied with the advent of commercial platforms like Amazon Rekognition [5], Baidu Research [7], Face++ [99], Google Vision [49], Microsoft Azure [78], and

✉ Kangning Yang
kangning.yang@student.unimelb.edu.au

[1] The University of Melbourne, Melbourne, Australia

Published online: 23 June 2020

Affectiva [1]. Most of these platforms offer cloud-based processing via an Application Programming Interface (API) or Software Development Kit (SDK) that enables developers and researchers to include emotion recognition capabilities in their applications [56].

However, to the best of our knowledge, only limited research has been conducted on validating the reliability and accuracy of different facial emotion recognition systems [10,17,30,96]. Furthermore, these studies have focused on model performance on selected databases rather than considering the imperfect conditions that might occur in real-world scenarios. For example, given the variation in individual habitual behaviors and the physical restrictions of smartphone cameras and surrounding environments, it is common that a user's face is only partially visible or that the light conditions are too dark or too bright in photographs taken using front-facing smartphone cameras [61]. In such situations, it is currently unclear which facial emotion recognition system will yield the best results.

To fill this gap, we conducted two experiments. First, we utilized three publicly available datasets containing emotion-labeled high-quality full-face photographs to evaluate five commercial automatic facial emotion recognition systems. Then, we determined possible causes of poor-quality face images in naturalistic settings (i.e., a person using their smartphone) and classified them into the following categories: *rotation*, *partially visible face*, *brightness*, *blur*, and *noise*. We then evaluated the five commercial automatic facial emotion recognition systems using manipulated images representing different examples of each category.

Thus, the contribution of this paper is threefold:

(1) We conducted an assessment of five commercial automatic facial expression recognition systems in a variety of conditions, such as high-quality full-face images and a variety of poor-quality images;
(2) We identified different scenarios that can lead to poor-quality phone camera images taken in real-world scenarios, and created a set of images with different poor-quality conditions (e.g., different degrees of blur or noise). This mechanism can be used in future work aimed at assessing the reliability of emotion recognition systems that rely on facial expressions;
(3) We provide recommendations to the research community regarding the limitations and emotion recognition capabilities of each system.

2 Related work

In this section, we describe research in topics related to our study, namely emotion recognition, facial emotional expression, and methods for facial emotion recognition systems.

2.1 Emotion recognition

Emotional expression refers to how people communicate internal affective states to others through both verbal and non-verbal behaviors [2,18,45,50,104,108]. During human communication, people can easily perceive the emotional states of others from multiple signals, for example, a happy emotional state from a smiley face and positive words, or a sad emotional state from a slight frown and a tearful tone. However, it is challenging to enable computers to deduce high-level affective states from low-level signal cues. This is mainly because of the gap between the extracted feature representation and human affective states, or the complexity of fusion patterns of different features [40,52,63].

Over the years, several approaches and algorithms have been proposed to detect people's emotional states. Some of them are based on the classification and analysis of facial expression including 2D and 3D representations [11,12,77,92,93], body language or posture [25,47,76], speech [69,94], written text [3,71] and physiological signals [4,54]. Among different available data sources, Ekman states that facial expression is "the most commanding" [31], and more than other behaviors can be "quite intense even when a person is alone" [31]. In other words, analyzing facial expressions is one of the most direct ways to study human emotion.

2.2 Facial emotional expression

In Darwin's seminal work, *The Expression of the Emotions in Man and Animals* [26], he posited that "facial expressions are the residual actions of more complete behavioral responses, and occur in combination with other bodily responses—vocalizations, postures, gestures, skeletal muscle movements, and physiological responses" [68,91]. More recently, researchers have theorized that all people express emotions via common facial expressions [35,37,39,58,74]. In other words, there exists a certain degree of universality in emotion representation through facial expressions, a position termed *the universality hypothesis*. Others question universality and propose that facial expressions of emotion vary with culture and language [8,59,80,90]. Some researchers argue that people interpret facial expressions in terms of the social situations in which they occur [20]. Despite such controversy, both proponents and critics agree that human beings can express and obtain emotional information through facial movements [80]. Most of the commercial emotion recognition systems today exactly rely on facial expressions to achieve emotion recognition function and provide related services.

Furthermore, studies have shown that facial expressions are recognizable in terms of different discrete emotion categories [9,18,32,36,41,57,75]. In particular, works on emotion recognition using facial expressions have classified six basic emotions that are typically recognizable: anger, disgust, fear, happiness, sadness, and surprise [14,33,73,101]. Facial emotion analysis has been an active topic within the research community, and recently several commercial facial emotion recognition systems have become available that focus mostly on these six basic emotions. Our goal in this study is to evaluate how five commonly used commercial emotion recognition systems perform when they are challenged with analyzing naturalistic facial photographs.

2.3 Methods for facial emotion recognition systems

A widely used approach for classifying facial expressions is based on the Facial Action Coding System (FACS) developed by Ekman and Friesen [38,42]. FACS describes facial behaviors in terms of a set of specific Action Units (AUs) each of which is associated with an individual face muscle or muscle group [38,42]. Using this approach, trained human coders can manually deconstruct and annotate nearly any possible facial activity as a combination of AUs. Follow-up work proposed a derivative of FACS called Emotion FACS (EMFACS), which focuses only on subsets of AUs that are likely to have universal emotional significance [43]. Performing analysis using FACS requires professional training and manual coding that is time- and labor-intensive, and until the 1980s most studies were completed by philosophers and psychologists [60,111].

With advances in computer technologies, such as machine learning, researchers began to design software tools for automated AU annotation and emotion recognition in order to overcome the limitations of FACS [13,67,97,103,110]. Automatic facial emotion recognition systems work in three stages: face detection, feature extraction, and emotion recognition. Given a still image, the system first determines the regions where faces are present. Then, by detecting specific AUs or facial landmarks such as eyes, eyebrow, and mouth, the system extracts and forms feature groups to represent the facial expressions. Finally, by applying machine learning algorithms such as naïve Bayes, decision tree, linear regression, or support vector machine, the system maps feature groups onto emotional states.

Recently, researchers have been further turning to 3D facial expressions and propose to improve facial feature tracking and emotion recognition based on 3D face scans. Compared with 2D facial expressions, 3D facial expressions are more robust with respect to the uncertainties and some external disturbances [11]. Most of these works focused on building generic 3D models based on manually extracted facial landmarks and corresponding features [48,72].

However, in terms of both 2D and 3D facial expressions, these extracted feature groups are low-level handcrafted features, and while having yielded promising classification accuracy, they are domain specific and do not perform well when generalized to varying real-world scenarios [86].

To address this issue, high-level visual features or representations were recently extracted using deep learning approaches such as convolutional neural networks (CNNs) [22,64,86,87] and recurrent neural networks (RNNs), including long short-term memories (LSTMs) and gated recurrent units (GRUs) [51,89]. Compared with low-level handcrafted features, deep neural network feature extraction can learn hierarchical features that are more generalizable and representative [51]. Not only that deep neural networks can also automate feature extraction and selection and achieve a much better emotion recognition performance than traditional machine learning models. For this reason, most of the current commercial facial emotion recognition systems rely on deep learning approaches to build their models and provide relevant recognition services.

3 Study overview

The aim of this study was to validate five popular commercial facial emotion recognition systems with respect to their performance in real-world usage scenarios, e.g., when people are using their smartphones. We designed two experiments: Experiment 1 investigated the five systems' performance on several facial emotion expression databases; Experiment 2 identified possible reasons for poor-quality photographs in real-world scenarios and examined how such photographs impacted the systems' performance. Before detailing both experiments, we describe the emotion recognition systems that we used in both experiments.

3.1 Facial emotion recognition system

There are numerous facial analysis products on the market today, which provide multiple services typically including face detection, face verification, and emotion recognition. However, previous works tend to select a given product without carefully considering performance before embedding it into their technologies. The selected system, therefore, may sometimes fail to provide adequate performance given the system's characteristics. In this study, in order to evaluate the potential performance and limitation of recognition systems, we chose five widely used commercial products (see Table 1) fulfilling the following criteria:

(1) has an existing API or SDK;
(2) recognizes at least six basic emotional categories [33–35] (highlighted in Table 1);
(3) returns a confidence or probability value for the detected emotion.

Table 1 Commercial facial emotion recognition services

Software platform	Emotions detected	Response	Approach
Amazon Rekognition	**Angry, disgusted, fear, happy, sad, surprised** calm, confused, unknown	Confidence range: [0,100]	Deep neural network
Baidu Research	**Anger, disgust, fear, happy, sad, surprise** neutral	Probability range: [0,1]	Deep neural network
Face++	**Anger, disgust, fear, happiness, sadness, surprise** neutral	Confidence range: [0,100]	Deep neural network
Microsoft Azure	**Anger, disgust, fear, happiness, sadness, surprise** neutral, contempt	Intensity range: [0,1]	Deep neural network
Affectiva	**Anger, disgust, fear, joy, sadness, surprise** contempt	Probability range: [0,100]	AFFDEX algorithm

Table 2 Test set (static pictures)

	Angry	Disgust	Fear	Happy	Sad	Surprise	Neutral	Total
ADFES	25	22	23	23	23	21	22	159
RaFD	67	67	67	67	67	67	67	469
WSEFEP	30	30	30	30	30	30	30	210
Total	122	119	120	120	120	118	119	838

4 Experiment 1: performance evaluation using ideal images

4.1 Databases

First, we introduce the three facial expression databases that we used in our experiments: the *Amsterdam Dynamic Facial Expressions Set* (ADFES[1]) [106], the *Radboud Faces Database* (RaFD[2]) [65], and the *Warsaw Set of Emotional Facial Expression Pictures* (WSEFEP[3]) [81]. These contain pictures of various emotional expressions, validated by FACS coders and non-expert human judges [96]. In this study, we evaluated seven emotions that were present in all databases including the six basic emotions (angry, disgust, fear, happy, sad, and surprise) and a neutral facial expression.

ADFES This database contains both dynamic and static emotional expressions displayed by 22 models. The set includes 10 female and 12 male actors (who are either Northern European or Mediterranean) showing ten emotions (anger, disgust, fear, joy, sadness, surprise, neutral, contempt, pride, and embarrassment). The used set consists of 159 static pictures.

RaFD This database contains a picture set of emotional expressions displayed by 67 models. This facial database includes 25 females and 42 males (who are either Caucasian or Moroccan) expressing eight emotions (anger, disgust, fear, happiness, sadness, surprise, neutral, and contempt). The used set consists of 469 static pictures.

WSEFEP This database contains a set of 210 emotional pictures captured from 30 individuals. The full set includes 16 female and 14 male models covering seven emotions (anger, disgust, fear, enjoyment, sadness, surprise, neutral). The used set consists of 210 static pictures.

Similar to previous work [96], we combined the selected subsets from ADFES, RaFD, and WSEFEP to form a whole set as our validation database, which consists of 122 *angry*, 119 *disgust*, 120 *fear*, 120 *happy*, 120 *sad*, 118 *surprise*, and 119 *neutral* images (838 pictures in total, see Table 2).

4.2 Method

4.2.1 Overall procedure

Due to the source codes of the commercial tools being encapsulated, i.e., the exact details of the inner workings of these tools are unknown and developers can only access the provided APIs or SDKs, we utilized a black box testing approach [84].

The overall procedure to determine an emotion expressed by a person within an image is shown in Fig. 1. The overall procedure can be divided into five stages as follows: (1) image selection: we chose every image from the aforementioned validation database and inputted them to each of the five commercial systems; (2) image analysis: after each system completes the analysis, we received and saved the corresponding result responses; (3) attribute extraction: we extracted the emotional attributes along with their confidence

[1] https://aice.uva.nl/research-tools/adfes-stimulus-set/adfes-stimulus-set.html.

[2] http://www.rafd.nl/.

[3] http://www.emotional-face.org/.

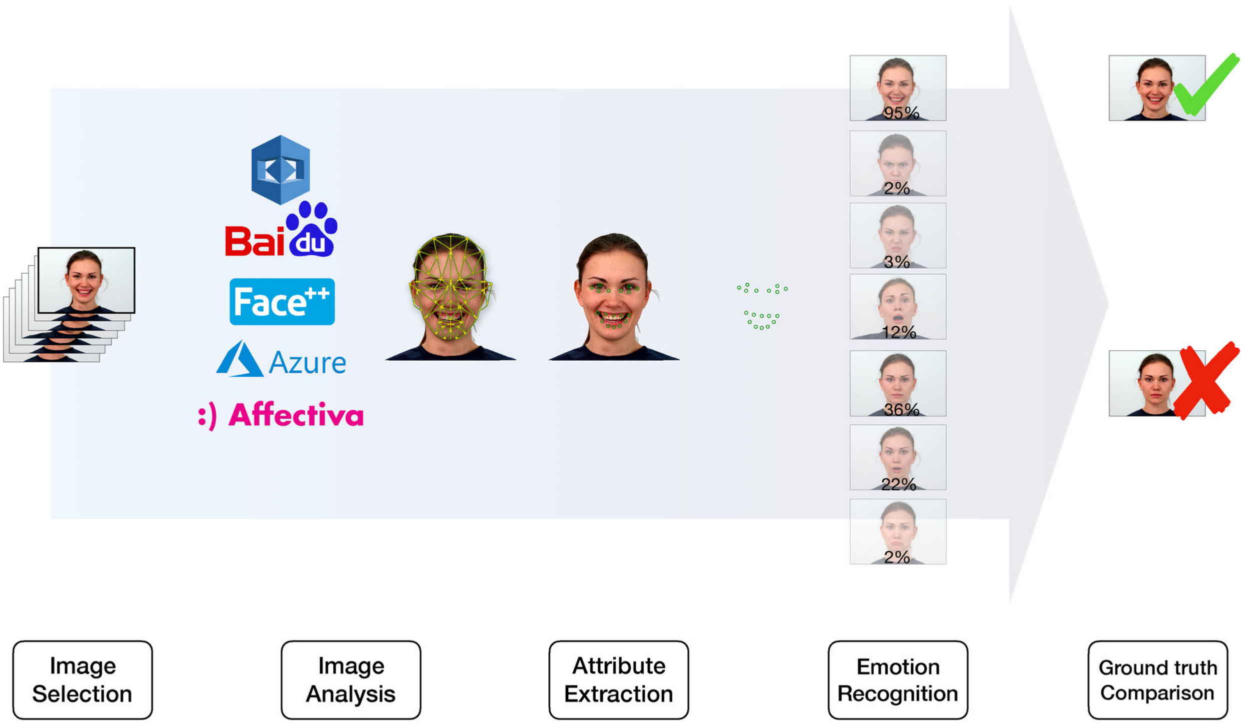


Fig. 1 Overall procedure

or probability values from the results; (4) emotion recognition: we selected the highest confidence or probability value of all detected emotions and marked this emotion as the detection label; (5) ground truth comparison: we compared the detection label with the ground truth label from the validation database to assess the recognition correctness.

If the detection label of a picture matched with the ground truth label for a given picture, that classification was labeled as "correct"; otherwise, the classification was labeled as "incorrect." If the system failed to detect the face in the picture, e.g., because it was only partially visible, there were no emotion recognition and an empty response: in such cases, the classification was labeled as "missed."

4.2.2 Comparison criteria

We used criteria similar to previous works to assess the emotion recognition systems [17,67,96], including **matching score**, **precision**, and **F1 score**, which we describe next.

Matching score (MS) measures the system's ability to correctly recognize an emotion category from a particular class (e.g., happiness, sadness, etc.) of emotional expression pictures, which is also called **true positive rate** or **recall**. A higher matching score means that the system has a higher accuracy on this type of emotional expression pictures. Specifically,

$$\text{MS} = \frac{\Sigma\text{true positive}}{\Sigma\text{true positive} + \Sigma\text{false negative} + \Sigma\text{missed}} \tag{1}$$

where "true positive" means the detection label matches the ground truth label; "false negative" means the detection label contradicts the ground truth label; "missed" means the system fails to detect an emotion (no detection label), among all images with the same annotated label.

Precision denotes the system's reliability of recognizing results, which is also called **positive predictive value**. A higher precision value indicates a higher confidence level, when classifying an expression picture as a certain emotional category. It can be defined as:

$$\text{Precision} = \frac{\Sigma\text{true positive}}{\Sigma\text{true positive} + \Sigma\text{false positive}} \tag{2}$$

F1 Score is a balance between MS and precision. We calculated an $F1$ score in which MS and precision contribute equally by:

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{MS}}{\text{Precision} + \text{MS}} \tag{3}$$

where the $F1$ score reaches its best value at 1 and worst value at 0.

Table 3 Experiment 1 results. Matching score (%), precision (%), $F1$ score, overall accuracy (%), and detection rate (%)

		Amazon	Baidu	Face++	Microsoft	Affectiva
Surprise	MS	89.8	97.5	99.2	**100.0**	97.5
	Precision	73.1	**92.0**	68.8	74.7	49.8
	$F1$	0.806	**0.947**	0.812	0.855	0.659
Fear	MS	68.3	**90.0**	49.2	62.5	7.5
	Precision	82.8	97.3	95.2	**100.0**	50.0
	$F1$	0.749	**0.935**	0.649	0.769	0.130
Disgust	MS	83.2	91.6	**95.0**	93.3	89.9
	Precision	96.1	97.3	86.9	**100.0**	75.4
	$F1$	0.892	0.944	0.908	**0.965**	0.820
Happy	MS	**100.0**	**100.0**	**100.0**	**100.0**	94.2
	Precision	**100.0**	**100.0**	99.2	**100.0**	96.6
	$F1$	**1.000**	**1.000**	0.996	**1.000**	0.954
Sad	MS	90.8	**98.3**	82.5	90.0	63.3
	Precision	90.1	**92.2**	81.2	86.4	62.3
	$F1$	0.904	**0.952**	0.818	0.882	0.628
Angry	MS	77.1	**96.7**	48.4	48.4	50.8
	Precision	82.5	**93.7**	86.8	92.2	88.6
	$F1$	0.797	**0.952**	0.621	0.635	0.646
Neutral	MS	98.3	96.6	97.5	**100.0**	
	Precision	86.0	**99.1**	70.3	64.3	
	$F1$	0.917	**0.978**	0.817	0.783	
Overall accuracy		86.8	**95.8**	81.5	84.7	67.0
Detection rate		**100.0**	**100.0**	**100.0**	**100.0**	97.4

Bold values indicate the highest value (horizontal direction) among all systems under different conditions

4.3 Results

Table 3 depicts all detailed values of Experiment 1. Figure 2 shows the confusion matrix for each emotion recognition system.

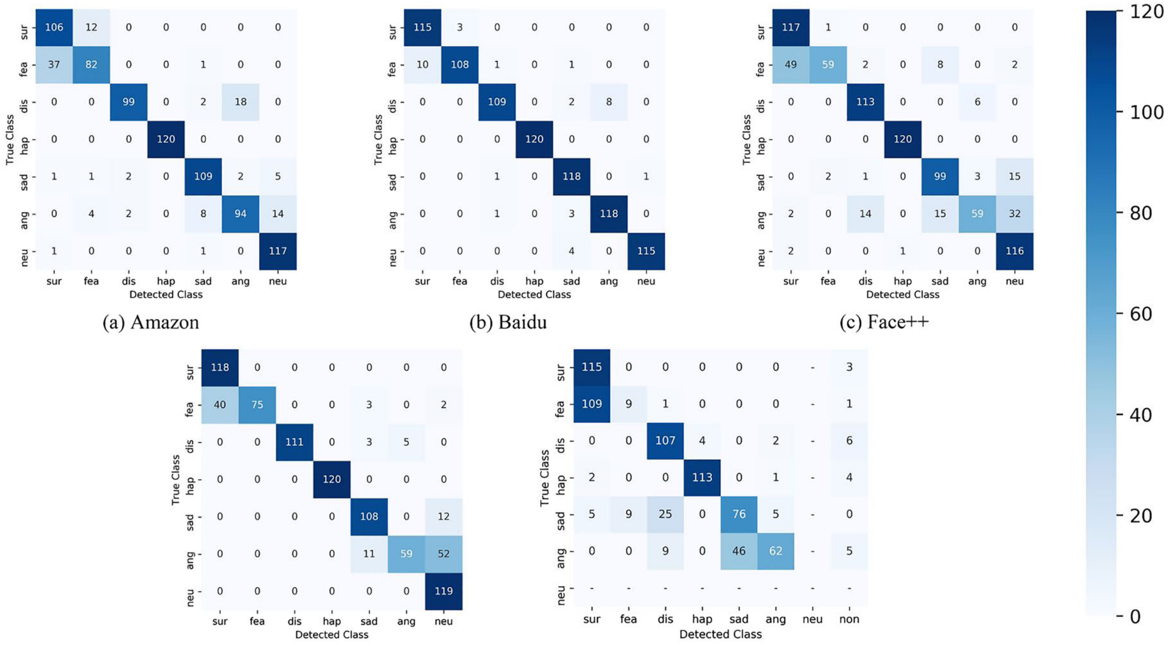
We found that the emotion *happy* was the easiest to identify and classify. All emotion recognition systems achieved ≈ 100.0% MS, precision and $F1$ score, except for Affectiva (94.2% MS, 96.6% precision, 0.954 $F1$ score). For *surprise* and *neutral*, Microsoft Azure had the highest MS (100.0% and 100.0%), but Baidu achieved the highest precision value (92.0% and 99.1%). In other words, given a *surprise* (or *neutral*) annotated facial expression image, Microsoft Azure was more likely to recognize the emotion correctly; however, if given a facial expression image without annotated ground truth, the *surprise* (or *neutral*) recognition result from Baidu had more credibility. In contrast, for *fear* Baidu had the highest MS (90.0%), but Microsoft Azure had the highest Precision value (100.0%). For *disgust*, Face++ achieved the highest MS (95.0%), yet Microsoft Azure performed the best on the Precision value (100.0%). Finally, for *sad* and *angry*, Baidu had the best performance on both MS (98.3% and 96.7%) and precision value (92.2% and 93.7%). In terms of $F1$ score from each emotion category, Baidu performed the best on *surprise*, *fear*, *sad*, *angry*, and *neutral* (0.947, 0.935, 0.952, 0.952, and 0.978); Microsoft Azure performed the best on *disgust* (0.965), while Amazon, Baidu, Face++, and Microsoft Azure performances were equally comparable to each other on *happy* (≈ 1.000).

We measured the **overall accuracy** of each emotion recognition system by globally counting the total number of true positives. Baidu had the highest overall accuracy value (95.8%), with Amazon second (86.8%), followed by Microsoft Azure (84.7%), Face++ (81.5%), and finally Affectiva (67.0%). Furthermore, while Affectiva failed to detect a face in 2.6% of the pictures, other systems performed particularly well and did not produce any detection failures.

Figure 2 shows that *surprise*, *disgust*, *happy*, and *neutral* expressions were rarely confused with other emotions by the recognition systems. For other emotion categories, we found the following:

(1) Amazon, Face++, Microsoft Azure, and Affectiva usually underpredict the *fear* emotion but overpredict the *surprise* emotion;

(2) Affectiva usually underpredicts the *sad* emotion but overpredicts the *disgust* emotion;



(a) Amazon

(b) Baidu

(c) Face++

(d) Microsoft

(e) Affectiva

Fig. 2 Confusion Matrix for Experiment 1. (sur = *surprise*; fea = *fear*; dis = *disgust*; hap = *happy*; sad = *sad*; ang = *angry*; neu = *neutral*). Affectiva cannot detect *neutral* expression

(3) Face++ and Microsoft Azure usually underpredict the *angry* emotion but overpredict the *neutral* emotion, while Affectiva usually underpredicts the *angry* emotion but overpredicts the *sad* emotion.

Generally, *fear* and *angry* expressions were the most challenging to identify and classify.

5 Experiment 2: performance evaluation using reduced quality images

Facial emotion recognition systems typically perform best when inputting high-quality images that show a person's full face, as they are trained and tested on pristine image databases. However, in practice, such requirements are unlikely to be satisfied, as distortions are quite likely to be encountered during image acquisition, transmission, or storage [27]. For example, in recent work, where researchers continuously collected photographs from the front-facing camera of 11 smartphone users for 2 weeks, only 29% of the time was the user's full face visible, while most of the time the face was only partially visible [61]. Thus, during smartphone use it is not realistic to expect that images collected will always contain the users' full faces. This is of particular importance as several previous studies have shown that the way people hold their smartphones can influence the performance of facial analysis algorithms [16,61,66].

This problem attracts much attention from researchers since it is impossible to avoid when implementing facial emotion recognition algorithms in real-life products. Especially for applications that need real-time facial emotion recognition capability (e.g., E-learning monitoring or mental health management systems), a more robust recognition framework that can (nearly) continually detect human emotion over time, regardless of image or video quality, is essential. Researchers have begun attempting to apply off-the-shelf recognition systems in different environments [95]. Thus, a comprehensive validation study of current commercial facial emotion recognition systems on images of varying quality is warranted.

After a comprehensive consideration of in-the-wild user behaviors and image quality distortions, we identified five potential reasons for poor-quality natural photographs: *rotation*, *partial face*, *brightness*, *blur*, and *noise*. These are the most common distortions known to impact the performance of deep neural networks [24,28,29,62,88,98]. We augmented the validation database of Experiment 1 using these five distortions (see Table 4) and then quantitatively investigated changes in performance for each of the facial recognition systems. To the best of our knowledge, we are the first to take into account both user behaviors and image quality distortions in a large-scale evaluation of commercial facial emotion recognition systems on multi-type and multi-granularity level.

Table 4 Augmented test set (static pictures)

	Angry	Disgust	Fear	Happy	Sad	Surprise	Neutral	Total
Rotation	732	714	720	720	720	708	714	5028
Partial Face	976	952	960	960	960	944	952	6704
Brightness	732	714	720	720	720	708	714	5028
Blur	366	357	360	360	360	354	357	2514
Noise	366	357	360	360	360	354	357	2514
Total	3172	3094	3120	3120	3120	3068	3094	21,788

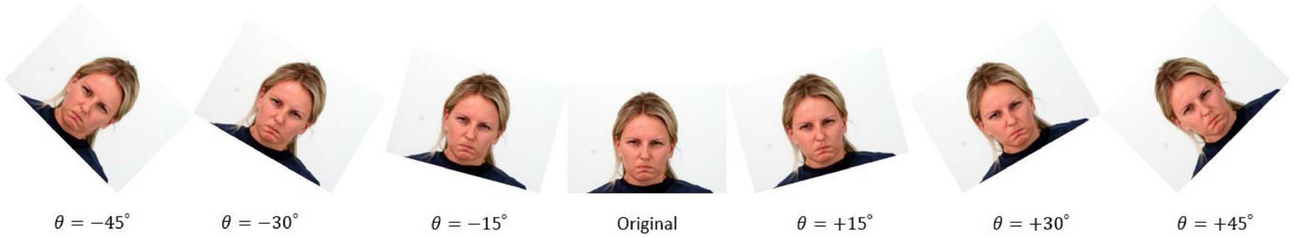


Fig. 3 Example of rotation manipulations

5.1 Rotation

Rotation occurs when photographs are taken from diverse angles by users. The rotation operation is achieved through a geometric transformation by the specified angle θ. We used OpenCV and Python to rotate each photograph around its center from -45° to 45° in steps of 15°, without scaling or cropping. We set 45° as the boundary value because pictures will be auto-rotated on some smartphone devices if they are tilted too far. For example, if taking photographs in portrait mode on some devices, the resulting photographs will be rotated by 90°. Figure 3 shows an example of the image rotation. Here, a negative θ means clockwise rotation, and vice versa.

Table 5 and Fig. 4 show the emotion recognition results for the rotated images. We found that Amazon, Baidu, and Face++ performed well and were stable to different θ. Their overall accuracies for the rotated images were almost as good as for the original images. However, Microsoft Azure and Affectiva were vulnerable to rotation manipulation. Their detection accuracies decreased significantly when the absolute value of θ increased. Overall, Baidu had the highest accuracy value, followed by Amazon and Face++.

There are several specific results worth mentioning: (1) Microsoft Azure's performance was resilient when $|\theta| = 15^{\circ}, 30^{\circ}$; however, performance began to decrease when $|\theta| > 30^{\circ}$ and deteriorated further at $|\theta| = 45^{\circ}$ (OA $= 4.2\%, \theta = -45^{\circ}$; OA $= 3.1\%, \theta = 45^{\circ}$); (2) there was little change to Affectiva's performance at $|\theta| = 15^{\circ}$; Affectiva presented a substantial performance decrease when $|\theta| > 15^{\circ}$, and it stopped functioning at $|\theta| = 30^{\circ}, 45^{\circ}$ (OA $= 0.0\%$); (3) compared with Microsoft Azure, Affectiva was more sensitive to rotation manipulation; (4) when $|\theta| = 15^{\circ}$, Microsoft Azure performed better than Face++.

In terms of detection rates, Amazon, Baidu, and Face++ had the best results with 100.0%. However, Microsoft Azure only detected a face in $\approx 5\%$ of the photographs when $|\theta| = 45^{\circ}$; Affectiva performed worse as it failed to detect a face in all photographs when $|\theta| = 30^{\circ}, 45^{\circ}$, further explaining why the overall accuracy of these two systems diminished with the change in rotation degrees.

5.2 Partial face

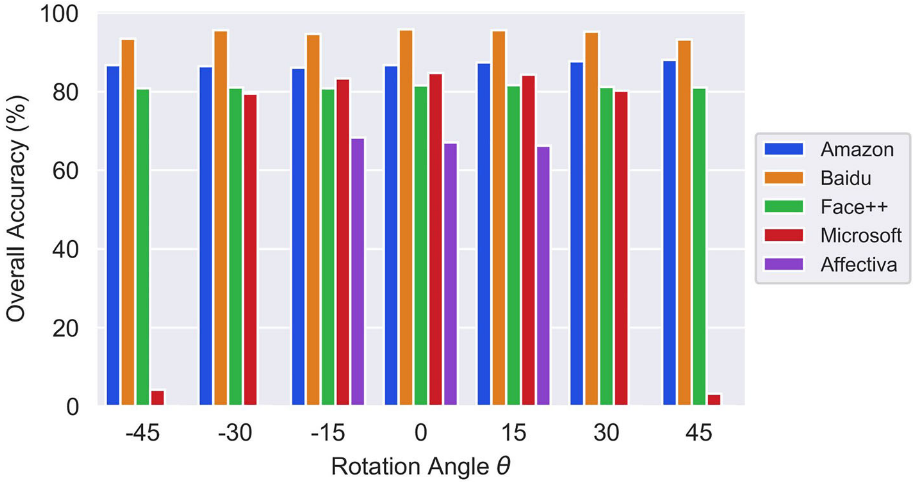
Partially visible face images are common when devices are naturally held by users [61]. We considered both portrait and landscape orientations and further classified partial face into eight sub-categories: left quarter (LQ), left half (LH), right quarter (RQ), right half (RH), up quarter (UQ), up half (UH), down quarter (DQ), and down half (DH). We implemented the CascadeClassifier in OpenCV [82] and then manipulated photographs for the partial face experiment using two steps: first, we determined the rectangular region of the face using the CascadeClassifier; second, we shifted this region on the x-axis or y-axis through coordinate translation. Figure 5 shows an example of partial face manipulations in both x-axis and y-axis. We did not consider three quarters scale face translations (e.g., left three quarters or up three quarters), since the vast majority of facial landmarks (or AUs) are concentrated in brow, lid, eyes, nose, cheek, and lip [38]. When shifting the face in a photograph by three quarters, it will leave only ears (left, right three quarters), chin (up three quarters), or forehead (down three quarters) that contribute little to facial emotion recognition.

Table 5 Detailed results of rotation manipulations

		Amazon	Baidu	Face++	Microsoft	Affectiva
$\theta = -45°$	OA	86.8	**93.4**	80.8	4.2	0.0
	DR	100.0	100.0	100.0	5.1	0.0
$\theta = -30°$	OA	86.4	**95.6**	81.0	79.5	0.0
	DR	100.0	100.0	100.0	99.5	0.0
$\theta = -15°$	OA	86.0	**94.6**	80.8	83.3	68.3
	DR	100.0	100.0	100.0	100.0	96.9
Original	OA	86.8	**95.8**	81.5	84.7	67.0
	DR	100.0	100.0	100.0	100.0	97.4
$\theta = 15°$	OA	87.4	**95.6**	81.6	84.3	66.2
	DR	100.0	100.0	100.0	100.0	94.0
$\theta = 30°$	OA	87.7	**95.2**	81.2	80.2	0.0
	DR	100.0	100.0	100.0	99.3	0.0
$\theta = 45°$	OA	88.1	**93.2**	81.0	3.1	0.0
	DR	100.0	100.0	100.0	5.0	0.0

Bold values indicate the highest value (horizontal direction) among all systems under different conditions
OA overall accuracy (%), *DR* detection rate (%)

Fig. 4 Overall accuracy under different rotation angles



An overview of the accuracy values and detection rates is shown in Table 6. For partial face manipulation on the x-axis, Fig. 6a reveals that Amazon, Baidu, and Face++ were relatively resilient to partial face manipulation. Their performance began to worsen only when facing a high-level visible face change (LH or RH). Overall, Amazon showed the greatest resiliency and had the highest overall accuracy values at high-level visible face changes (LH and RH); Baidu performed best at low-level visible face changes (LQ and RQ). As for Microsoft Azure, it performed well at low-level visible face changes with the second-highest overall accuracy values (89.6% and 87.0%) for LQ and RQ. However, it was quite sensitive to high-level visible face changes and struggled to detect any face for LH and RH. Affectiva had the lowest accuracy values and detection rates at all levels of partial face manipulation. It is also noteworthy that the overall accuracy of all systems was approximately symmetrical to the original image, except for Affectiva. This is probably due to similar characteristics (e.g., approximate number of facial landmarks) on both sides of the face.

For partial face manipulation on the y-axis, Fig. 6b shows that all systems were highly sensitive to visible face changes in portrait orientation. Even for low-level visible face changes, their performance decreased significantly. In general, Baidu produced the best overall accuracy results; Amazon and Face++ performed just slightly worse. Microsoft Azure had a similar performance to Amazon and Face++ at low-level visible face changes, but failed completely to detect a face at high-level visible face changes (0.0% and 0.0% at both DH and UH). Affectiva had a large number of detection failures from low-level visible changes (12.9% detection rate at DQ and 11.8% detection rate at UQ). Furthermore, when examining the trend of the systems' performance, our results show that all systems, except

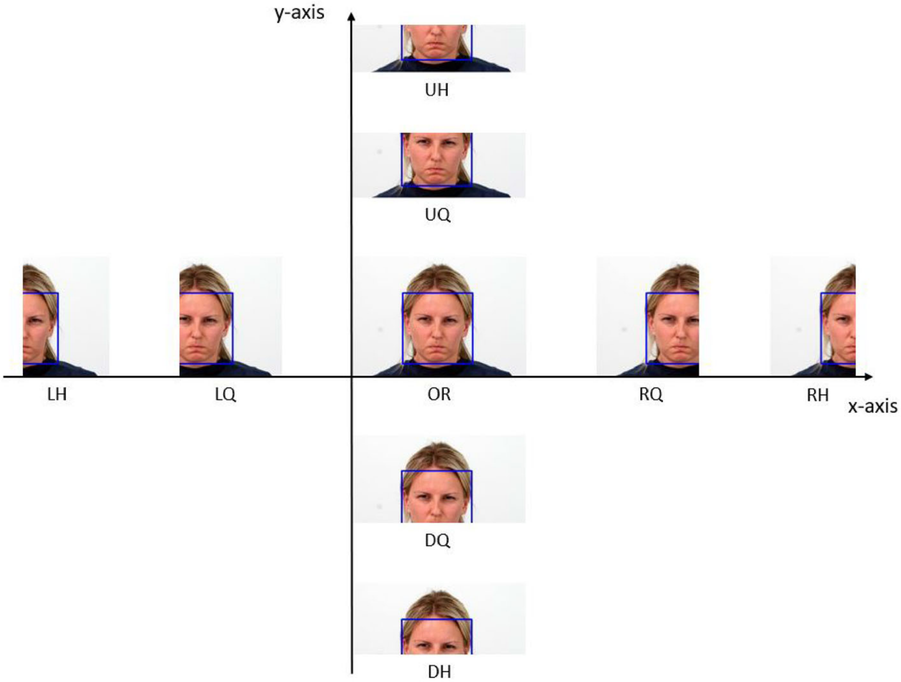


Fig. 5 Example of partial face manipulations

Table 6 Detailed results of partial face manipulations

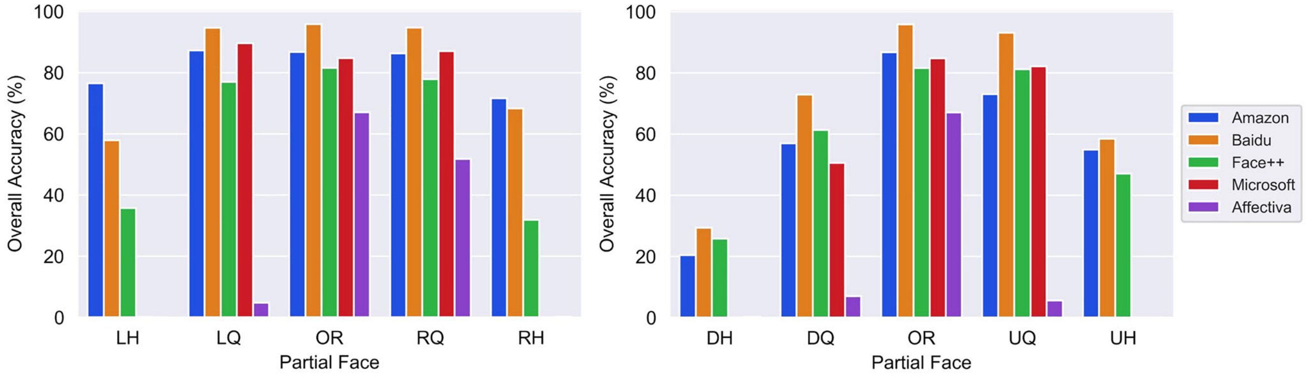
		Amazon	Baidu	Face++	Microsoft	Affectiva
LH	OA	**76.5**	57.9	35.7	0.1	0.0
	DR	100.0	88.8	69.5	0.5	0.0
LQ	OA	87.2	**94.6**	77.0	89.6	4.7
	DR	100.0	100.0	100.0	100.0	6.8
RQ	OA	86.3	**94.8**	77.8	87.0	51.7
	DR	100.0	100.0	100.0	99.6	78.4
RH	OA	**71.6**	68.3	31.9	0.0	0.1
	DR	100.0	96.2	66.2	0.0	0.1
Original	OA	86.8	**95.8**	81.5	84.7	67.0
	DR	100.0	100.0	100.0	100.0	97.4
DH	OA	20.4	**29.4**	25.8	0.0	0.1
	DR	99.8	58.2	99.5	0.0	0.1
DQ	OA	56.9	**72.9**	61.3	50.5	7.0
	DR	100.0	100.0	100.0	96.8	12.9
UQ	OA	73.0	**93.1**	81.2	82.1	5.6
	DR	100.0	100.0	100.0	100.0	11.8
UH	OA	54.9	**58.5**	47.0	0.0	0.0
	DR	100.0	90.3	95.2	0.0	0.3

Bold values indicate the highest value (horizontal direction) among all systems under different conditions
OA overall accuracy (%), *DR* detection rate (%)

for Affectiva, were always more sensitive to the downward direction manipulation (e.g., DQ and DH) than the upward direction manipulation (e.g., UQ and UH). This is likely due to these systems having a bias toward features present in the lower part of a face.

5.3 Brightness

Brightness is a key factor that impacts the quality of images taken in real-world scenarios. A universally accepted approach is to divide brightness problems into two categories: overexposure and underexposure [55]. When taking a photograph with strong lighting, the image sensor will cap-



(a) x-axis

(b) y-axis

Fig. 6 Overall accuracy of partial face manipulations. *OR* Original

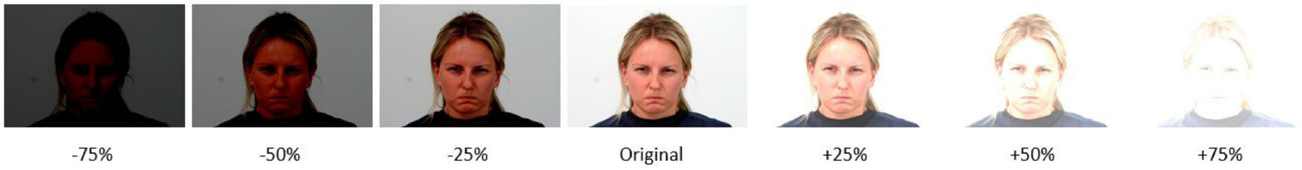


Fig. 7 Example of brightness manipulations

ture too much light and make the photographs lose some highlight details (called overexposure). On the contrary, when taking a photograph with poor lighting, the image sensor will capture too little light and make the photograph lose some shadow details (called underexposure). In this experiment, we varied the brightness of the images from 75% reduction (−75%) to 75% enhancement (+75%) in steps of 15% by decreasing/increasing the value of every pixel in each image. We note that an image will become completely black or white if the brightness is reduced or increased by 100%. Figure 7 shows an example of the brightness levels we tested.

Results for all facial emotion recognition systems are shown in Table 7. Figure 8 depicts the overall accuracy of all systems under different brightness conditions. Our results show that the ranking of system in terms of overall accuracy remained consistent at low and medium degrees of brightness changes (25% and 50% reduction/enhancement). Baidu was the top performing system; Amazon and Microsoft Azure ranked similarly (but Amazon performed better at a medium degree of enhancement; Microsoft Azure performed better at medium degree of reduction), followed by Face++ and Affectiva. However, for the highest degree of brightness changes (75% reduction/enhancement), the performance of all systems experienced a sharp decline: Baidu remained the relatively most accurate with dark images (36.5%), whereas Amazon achieved better performance on the overall accuracy for bright images (45.2%) with a high detection rate of 95.0%. Another interesting observation is that the detection rates of Amazon were relatively stable ($\geq$ 94%) across all degrees of brightness changes, which contributed toward relatively good performance in terms of overall accuracy of this system (Table 7).

5.4 Blur

Blur is a common factor in poor-quality images and can be caused by long exposure time, movement of the camera, or inappropriate focus when using a smartphone or other handheld cameras [27,109]. In this experiment, we used Gaussian Blur with different granularities implemented by the Python Imaging Library (PIL) to model and simulate image blur that can occur in a real-world scenario. As recommended by Gedraite and Hadad [44], we set the blur radius (the standard deviation of Gaussian) σ to three values: 1, 3, and 10, to represent low, medium, and high image blurring, respectively. An example of this manipulation is shown in Fig. 9.

We show the overall results in Table 8. Figure 10 reveals that all systems showed stable performance when it comes to low ($\sigma = 1$) and medium ($\sigma = 3$) levels of image blur. Specifically, Baidu achieved the best performance with 96.1% ($\sigma = 1$) and 95.0% ($\sigma = 3$) overall accuracy values; Amazon had the second-best performance with 86.8% and 86.0% accuracy values (on $\sigma = 1, 3$); Microsoft Azure (85.0%) performed better than Face++ (81.5%) at low-level blurring, but Face++ (81.9%) yielded a higher accuracy result than Microsoft Azure (80.3%) at medium-level blurring. Another finding was that Baidu, Microsoft Azure, and Affectiva showed a significant decrease at high-level blurring, in

Table 7 Detailed results of brightness manipulations

		Amazon	Baidu	Face++	Microsoft	Affectiva
−75%	OA	30.6	**36.5**	24.2	9.7	12.8
	DR	94.9	69.2	77.0	14.8	29.5
−50%	OA	67.5	**87.2**	65.8	80.9	64.0
	DR	100.0	99.8	100.0	99.5	96.4
−25%	OA	85.1	**95.0**	77.1	84.8	70.0
	DR	100.0	100.0	100.0	100.0	98.2
Original	OA	86.8	**95.8**	81.5	84.7	67.0
	DR	100.0	100.0	100.0	100.0	97.4
25%	OA	86.8	**95.2**	80.6	85.4	66.6
	DR	100.0	100.0	100.0	100.0	96.8
50%	OA	78.0	**82.6**	71.8	71.1	58.1
	DR	100.0	90.7	99.6	92.4	90.8
75%	OA	**45.2**	3.8	22.4	0.5	7.0
	DR	95.0	5.6	53.0	0.8	12.2

Bold values indicate the highest value (horizontal direction) among all systems under different conditions

OA overall accuracy (%), *DR* detection rate (%)

Fig. 8 Overall accuracy under different brightness conditions

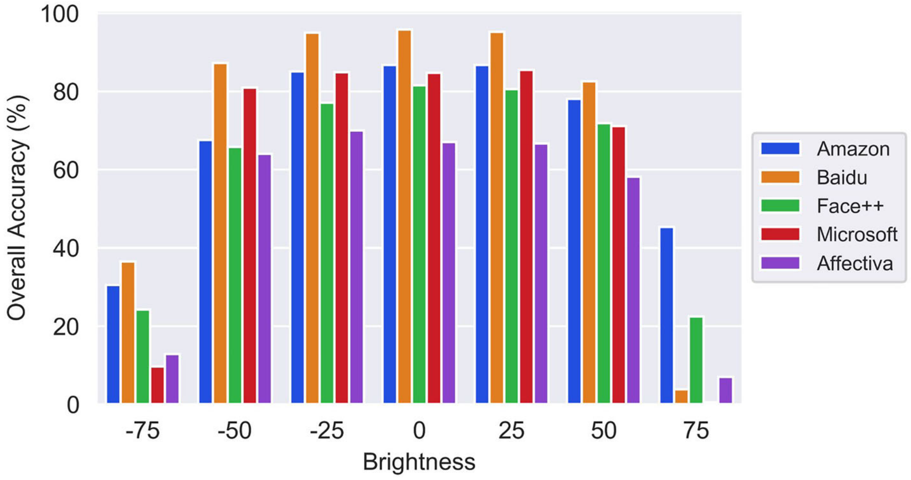
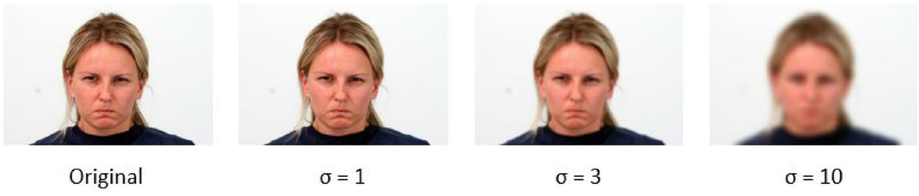


Fig. 9 Example of blur manipulations



which their overall accuracies fell by 45.5% $\left(\frac{95.8-52.2}{95.8}\right)$, 50.6%, and 50.0%, respectively. However, there was little effect on Amazon and Face++. Therefore, at high-level blurring, Amazon worked best; Face++ was the second best; followed by Baidu, Microsoft Azure, and Affectiva. In terms of stability, our results show that Amazon and Face++ performed better than the other systems.

5.5 Noise

Image noise is usually a form of electronic noise [15], and it is typically caused by the internal sensors of cameras [15,27,109]. In practice, noise can obscure image details, degrade its quality, or even make the image completely unusable. Previous works have commonly modeled image noise using Additive White Gaussian Noise (AWGN) [70,83], since it can effectively mimic the noise that occurs in real world. Therefore, in this experiment, we also used AWGN to

Table 8 Detailed results of blur manipulations

		Amazon	Baidu	Face++	Microsoft	Affectiva
Original	OA	86.8	**95.8**	81.5	84.7	67.0
	DR	100.0	100.0	100.0	100.0	97.4
$\sigma = 1$	OA	86.8	**96.1**	81.5	85.0	68.6
	DR	100.0	100.0	100.0	100.0	96.9
$\sigma = 3$	OA	86.0	**95.0**	81.9	80.3	60.5
	DR	100.0	100.0	100.0	100.0	95.3
$\sigma = 10$	OA	**83.2**	52.2	75.1	41.8	33.5
	DR	100.0	100.0	100.0	100.0	82.8

Bold values indicate the highest value (horizontal direction) among all systems under different conditions
OA overall accuracy (%), *DR* detection rate (%)

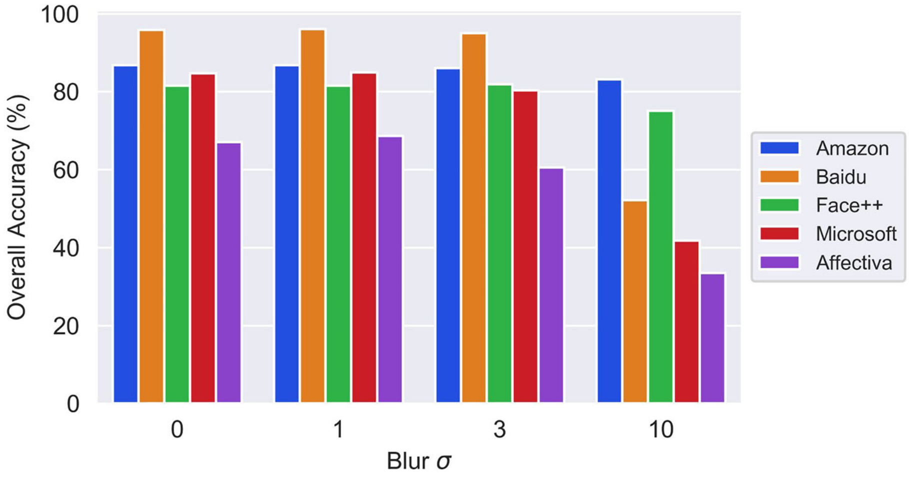


Fig. 10 Overall accuracy under different blur conditions

simulate natural noise in digital images. We set the standard deviation of the noise to be: 10, 40, and 70, to represent low, medium, and high noise levels, respectively. Figure 11 shows an example of noise manipulations.

Table 9 summarizes the overall results of the noise manipulations. Figure 12 shows that (1) at low noise level ($\sigma = 10$), all systems' performance remained basically unchanged and kept the same ranking positions: Baidu, Amazon, Microsoft Azure, Face++, and Affectiva; (2) at medium noise level ($\sigma = 40$), all systems showed a decrease in accuracy values, but the performance of Microsoft Azure and Face++ deteriorated at a slower rate (by 4.1% and 4.5%, respectively), whereas the performance of Baidu, Amazon, and Affectiva deteriorated quicker (by 18.1%, 12.8%, and 59.3%, respectively); (3) at high noise level ($\sigma = 70$), the performance of Baidu, Amazon, and Affectiva still fell much quicker compared to that of Microsoft Azure and Face++. In general, Microsoft Azure and Face++ showed greater reliability in terms of processing images with medium- and high-level noise strength.

6 Discussion

In this study, we evaluated the performance of five commercial facial emotion recognition systems both with high-quality front facing images and images whose quality had been reduced by common real-world distortions (rotation, partial face, brightness, blur, noise).

6.1 Facial emotion recognition under ideal conditions

When identifying emotions from full-face high-quality photographs (Experiment 1), the Baidu system had the highest overall accuracy value (95.8%), followed by Amazon (86.8%), Microsoft Azure (84.7%), Face++ (81.5%), and Affectiva (67.0%). Likewise, a recent validation analysis using similar facial expression databases [96] found that Affectiva had an overall accuracy value of 73%, which is comparable to our result of 67.0%. That analysis also reported that Affectiva usually confused *fear* with *surprise* and *anger* with *sadness*, which is in line with our findings. Furthermore, previous works using human judges have shown that participants achieved an overall accuracy between 82% and 85% [65,67,81,106]. Our results highlight that most of the popular commercial systems, particularly those using a deep neural network approach, can achieve accuracy rates comparable to human judges and sometimes even outperform them.

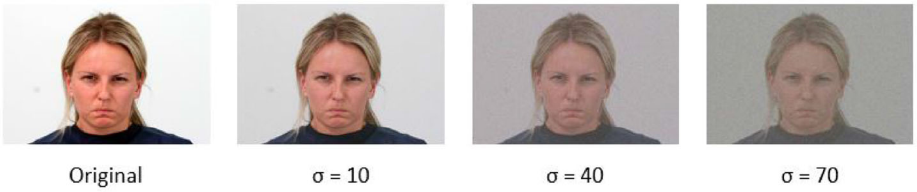


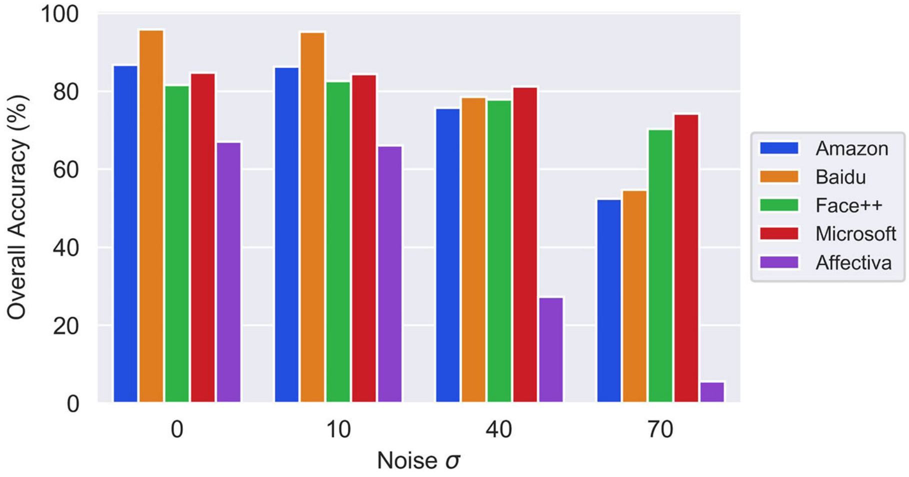
Fig. 11 Example of noise manipulations

Table 9 Detailed results of noise manipulations

		Amazon	Baidu	Face++	Microsoft	Affectiva
Original	OA	86.8	**95.8**	81.5	84.7	67.0
	DR	100.0	100.0	100.0	100.0	97.4
$\sigma = 10$	OA	86.3	**95.2**	82.6	84.4	66.1
	DR	100.0	100.0	100.0	100.0	98.6
$\sigma = 40$	OA	75.7	78.5	77.8	**81.2**	27.3
	DR	100.0	100.0	100.0	100.0	93.9
$\sigma = 70$	OA	52.4	54.7	70.3	**74.2**	5.6
	DR	100.0	95.9	99.2	99.8	30.5

Bold values indicate the highest value (horizontal direction) among all systems under different conditions
OA overall accuracy (%), *DR* detection rate (%)

Fig. 12 Overall accuracy under different noise conditions



Another noteworthy observation is that the tested systems were most accurate when identifying and classifying the *happy* emotion, and least accurate when processing the *fear* emotion. Similarly, previous works have shown that the *happy* emotion is also the easiest emotion to distinguish for human observers, and the *fear* emotion is the hardest [46,65]. This similarity is not surprising given that automated facial emotion recognition systems rely on human-generated labels to improve their performance.

6.2 Facial emotion recognition with reduced quality images

From the results of Experiment 2, we find that the tested systems have different advantages and limitations depending on the particular reason behind the poor quality of the images. This is most likely because different systems rely on their own training datasets, with their own set of biases. This finding highlights that a *one system approach* for facial emotion recognition in real-world scenarios is far from ideal. Instead, a robust system that uses facial expressions for emotion recognition should first determine the quality of the image and the level of distortion and then route the image to the facial emotion recognition system that is most likely to produce a reliable result (or discard the image if the quality is particularly poor).

Our recommendations are the following:

– For rotation, we recommend Baidu as the best system and Amazon as the second best;

- For partial face, on the x-axis, we recommend the use of Baidu for low-level visible face changes and Amazon for high-level visible face changes; on the y-axis, we recommend Baidu as first choice;
- For brightness, we recommend using Baidu when encountering underexposure or low- and medium-level overexposure, and Amazon when encountering high-level overexposure. We do not recommend Baidu at high-level overexposure;
- For blur, we recommend using Baidu when processing low- and medium-level image blurring, and Amazon or Face++ system when processing high-level image blurring;
- For noise, we recommend the use of Baidu at low-level noise strength and the use of Microsoft Azure or Face++ at medium- and high-level noise strength.

The reason why facial emotion recognition systems have poor overall accuracy rates in certain conditions is also noteworthy. There are two kinds of detection failures: face detection failure (“missed”) and emotion detection failure (“incorrect”). In detail,

- For Amazon Rekognition, emotion detection failure is the main issue and present itself as an opportunity for further improvement;
- For Baidu and Face++, not only emotion detection but also face detection needs to be improved for high-level partial face and underexposure/overexposure manipulations;
- Microsoft Azure suffers from both emotion detection failure and face detection failure at high-level changes of rotation, partial face, and brightness manipulations;
- Affectiva is very sensitive to most of the low- and high-level manipulations.

Finally, we note that the evaluated systems are likely to be improved in the future, and that other systems may emerge with a whole new set of advantages and limitations. Also, there are studies that show human beings have the capability to recognize very distorted images [6,23,28,29,100], which means existing recognition systems still have potential to overcome their limitations and achieve stronger robustness. Thus, another contribution of our work is the categorization and operationalization of the distortions that can lead to poor-quality images and which are likely to occur in real-world scenarios, e.g., smartphone usage. These manipulations can be used as standard methods for evaluating other facial emotion recognition systems, or for further improving the five systems validated in this study. Not only that researchers can also use our approach to build their own datasets for validation analysis according to different needs.

6.3 Limitations

There are limitations in our study. First, in Experiment 2, we analyzed and assessed the impacts of image manipulations using a relatively high-level of granularity (e.g., increments of 15° for the rotation condition). Thus, a finer-grained manipulation of the images may yield a more detailed validation study in the future. Second, we evaluated the impact of different distortion factors independently. Future work could investigate the potential for combination effects and even the prioritization of the individual factors. Third, it is likely that most of these systems will update their frameworks and corresponding network features and weights regularly, which may influence their performance under the different presented conditions. To tackle this issue, we recommend re-using or even automating our method to evaluate new commercial and non-commercial systems that may appear in the future.

7 Conclusion

In this study, we utilized three commonly used facial expression databases and proposed a series of manipulations to simulate potential image quality reduction problems that are likely to occur during real-world smartphone usage. We conducted two experiments in order to assess the facial emotion recognition capabilities of five popular commercial emotion recognition systems: Amazon, Baidu, Face++, Microsoft Azure, and Affectiva. Through our experiments, we found that each system has its own strengths and limitations under different distortion conditions (rotation, partial face, brightness, blur, noise). Based on our findings, we offer recommendations on how to achieve reliable facial emotion recognition results for applications in the wild, by selecting different systems depending on the nature of the captured image. Finally, we recommend the use of our image manipulation methods for future testing of facial emotion recognition system performance.

Funding This work is supported by the Australian Research Council (DP190102627).

Compliance with ethical standards

Conflict of interest Authors, Kangning Yang, Chaofan Wang, Zhanna Sarsenbayeva, Benjamin Tag, Tilman Dingler, Greg Wadley, and Jorge Goncalves, declare that they have no conflict of interest.

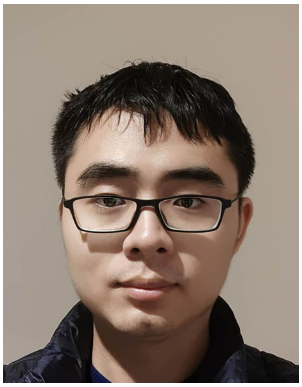
References

1. Affectiva: Home—Affectiva : Affectiva. https://www.affectiva.com/ (2019)

2. Albohn, D.N., Adams Jr., R.B.: Social vision: at the intersection of vision and person perception. In: Cloutier, J., Absher, J.R. (eds.) Neuroimaging Personality, Social Cognition, and Character, pp. 159–186. Elsevier, Amsterdam (2016)
3. Alm, C.O., Roth, D., Sproat, R.: Emotions from text: machine learning for text-based emotion prediction. In: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, pp. 579–586. Association for Computational Linguistics (2005)
4. AlZoubi, O., Calvo, R.A., Stevens, R.H.: Classification of EEG for affect recognition: an adaptive approach. In: Australasian Joint Conference on Artificial Intelligence, pp. 52–61. Springer, Berlin (2009)
5. Amazon: Amazon Rekognition-Video and Image-aws. https://aws.amazon.com/rekognition/?nc1=h_ls (2019)
6. Bachmann, T.: Identification of spatially quantised tachistoscopic images of faces: how many pixels does it take to carry identity? Eur. J. Cogn. Psychol. **3**(1), 87–103 (1991)
7. Baidu: Baidu ai. https://ai.baidu.com/docs#/Face-Detect-V3/top (2019)
8. Barrett, L.F., Adolphs, R., Marsella, S., Martinez, A.M., Pollak, S.D.: Emotional expressions reconsidered: challenges to inferring emotion from human facial movements. Psychol. Sci. Public Interest **20**(1), 1–68 (2019)
9. Bartlett, M.S., Littlewort, G., Fasel, I., Movellan, J.R.: Real time face detection and facial expression recognition: Development and applications to human computer interaction. In: 2003 Conference on Computer Vision and Pattern Recognition Workshop, vol. 5, pp. 53–53. IEEE (2003)
10. Bernin, A., Müller, L., Ghose, S., von Luck, K., Grecos, C., Wang, Q., Vogt, F.: Towards more robust automatic facial expression recognition in smart environments. In: Proceedings of the 10th International Conference on PErvasive Technologies Related to Assistive Environments, pp. 37–44. ACM (2017)
11. Berretti, S., Amor, B.B., Daoudi, M., Del Bimbo, A.: 3d facial expression recognition using sift descriptors of automatically detected keypoints. Vis. Comput. **27**(11), 1021 (2011)
12. Berretti, S., Del Bimbo, A., Pala, P.: Automatic facial expression recognition in real-time from dynamic sequences of 3d face scans. Vis. Comput. **29**(12), 1333–1350 (2013)
13. Bettadapura, V.: Face expression recognition and analysis: the state of the art. arXiv preprint arXiv:1203.6722 (2012)
14. Bourel, F., Chibelushi, C.C., Low, A.A.: Robust facial expression recognition using a state-based model of spatially-localised facial dynamics. In: Proceedings of Fifth IEEE International Conference on Automatic Face Gesture Recognition, pp. 113–118. IEEE (2002)
15. Boyat, A.K., Joshi, B.K.: A review paper: noise models in digital image processing. arXiv preprint arXiv:1505.03489 (2015)
16. Bröhl, C., Mertens, A., Ziefle, M.: How do users interact with mobile devices? an analysis of handheld positions for different technology generations. In: International Conference on Human Aspects of IT for the Aged Population, pp. 3–16. Springer, Berlin (2017)
17. Bryant, D., Howard, A.: A comparative analysis of emotion-detecting AI systems with respect to algorithm performance and dataset diversity. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, pp. 377–382. ACM (2019)
18. Calvo, R.A., D'Mello, S.: Affect detection: an interdisciplinary review of models, methods, and their applications. IEEE Trans. Affect. Comput. **1**(1), 18–37 (2010)
19. Carlotta Olivetti, E., Violante, M.G., Vezzetti, E., Marcolin, F., Eynard, B.: Engagement evaluation in a virtual learning environment via facial expression recognition and self-reports: a preliminary approach. Appl. Sci. **10**(1), 314 (2020)
20. Carroll, J.M., Russell, J.A.: Do facial expressions signal specific emotions? judging emotion from the face in context. J. Personal. Soc. Psychol. **70**(2), 205 (1996)
21. Celma, O.: Music recommendation. In: Gerstner, R. (ed.) Music Recommendation and Discovery, pp. 43–85. Springer, Berlin (2010)
22. Chang, W.Y., Hsu, S.H., Chien, J.H.: Fatauva-net: an integrated deep learning framework for facial attribute recognition, action unit detection, and valence-arousal estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 17–25 (2017)
23. Chen, Y., McBain, R., Norton, D.: Specific vulnerability of face perception to noise: a similar effect in schizophrenia patients and healthy individuals. Psychiatry Res. **225**(3), 619–624 (2015)
24. Cheng, Y., Jiang, B., Jia, K.: A deep structure for facial expression recognition under partial occlusion. In: 2014 Tenth International Conference on Intelligent Information Hiding and Multimedia Signal Processing, pp. 211–214. IEEE (2014)
25. Coulson, M.: Attributing emotion to static body postures: recognition accuracy, confusions, and viewpoint dependence. J. Nonverbal Behav. **28**(2), 117–139 (2004)
26. Darwin, C.: The Expression of the Emotions in Man and Animals, Anniversary edn. Harper Perennial, London (1872). (P. Ekman, ed)
27. Dodge, S., Karam, L.: Understanding how image quality affects deep neural networks. In: 2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX), pp. 1–6. IEEE (2016)
28. Dodge, S., Karam, L.: Can the early human visual system compete with deep neural networks? In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2798–2804 (2017)
29. Dodge, S., Karam, L.: A study and comparison of human and deep learning recognition performance under visual distortions. In: 2017 26th International Conference on Computer Communication and Networks (ICCCN), pp. 1–7. IEEE (2017)
30. Dupré, D., Andelic, N., Morrison, G., McKeown, G.: Accuracy of three commercial automatic emotion recognition systems across different individuals and their facial expressions. In: 2018 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops), pp. 627–632. IEEE (2018)
31. Ekman, P.: Methods for measuring facial action. In: Scherer, K.R., Ekman, P. (eds.) Handbook of Methods in Nonverbal Behavior Research, pp. 45–90. Cambridge University Press, Cambridge (1982)
32. Ekman, P.: Expression and the nature of emotion. Approaches Emot. **3**, 19–344 (1984)
33. Ekman, P.: An argument for basic emotions. Cogn. Emot. **6**(3–4), 169–200 (1992)
34. Ekman, P.: Basic emotions. In: Dalgleish, T., Power, M.J. (eds.) Handbook of Cognition and Emotion, vol. 98, pp. 45–60. Wiley, New York (1999)
35. Ekman, P., Friesen, W.V.: Constants across cultures in the face and emotion. J. Personal. Soc. Psychol. **17**(2), 124 (1971)
36. Ekman, P., Friesen, W.V.: Unmasking the Face: A Guide to Recognizing Emotions from Facial Clues. ISHK, Los Altos (2003)
37. Ekman, P., Friesen, W.V., Ellsworth, P.: Emotion in the Human Face: Guide-Lines for Research and an Integration of Findings: Guidelines for Research and an Integration of Findings. Pergamon, Oxford (1972)
38. Ekman, P., Friesen, W.V., Hager, J.C.: Facial Action Coding System: The Manual on CD ROM, pp. 77–254. A Human Face, Salt Lake City (2002)
39. Ekman, P., Friesen, W.V., O'sullivan, M., Chan, A., Diacoyanni-Tarlatzis, I., Heider, K., Krause, R., LeCompte, W.A., Pitcairn, T., Ricci-Bitti, P.E., et al.: Universals and cultural differences in

the judgments of facial expressions of emotion. J. Personal. Soc. Psychol. **53**(4), 712 (1987)
40. El Ayadi, M., Kamel, M.S., Karray, F.: Survey on speech emotion recognition: features, classification schemes, and databases. Pattern Recognit. **44**(3), 572–587 (2011)
41. Fridlund, A.J.: Evolution and facial action in reflex, social motive, and paralanguage. Biol. Psychol. **32**(1), 3–100 (1991)
42. Friesen, E., Ekman, P.: Facial Action Coding System: A Technique for the Measurement of Facial Movement, vol. 3. Consulting Psychologists Press, Palo Alto (1978)
43. Friesen, W.V., Ekman, P., et al.: Emfacs-7: emotional facial action coding system, vol. 2, no. 36, p. 1. Unpublished manuscript, University of California at San Francisco (1983)
44. Gedraite, E.S., Hadad, M.: Investigation on the effect of a Gaussian blur in image filtering and segmentation. In: Proceedings ELMAR-2011, pp. 393–396. IEEE (2011)
45. Gellman, M.D.: Behavioral Medicine. Springer, Berlin (2013)
46. Goeleven, E., De Raedt, R., Leyman, L., Verschuere, B.: The Karolinska directed emotional faces: a validation study. Cogn. Emot. **22**(6), 1094–1118 (2008)
47. Goncalves, J., Pandab, P., Ferreira, D., Ghahramani, M., Zhao, G., Kostakos, V.: Projective testing of diurnal collective emotion. In: Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing, UbiComp' 14, pp. 487–497. New York, NY, USA (2014)
48. Gong, B., Wang, Y., Liu, J., Tang, X.: Automatic facial expression recognition on a single 3d face by exploring shape deformation. In: Proceedings of the 17th ACM International Conference on Multimedia, pp. 569–572 (2009)
49. Google: Vision AI-derive image insights via ml-cloud vision api-google cloud. https://cloud.google.com/vision/ (2019)
50. Gross, J.J.: Emotion regulation: past, present, future. Cogn. Emot. **13**(5), 551–573 (1999)
51. Gu, Y., Li, X., Huang, K., Fu, S., Yang, K., Chen, S., Zhou, M., Marsic, I.: Human conversation analysis using attentive multimodal networks with hierarchical encoder–decoder. In: 2018 ACM Multimedia Conference on Multimedia Conference, pp. 537–545. ACM (2018)
52. Gu, Y., Yang, K., Fu, S., Chen, S., Li, X., Marsic, I.: Hybrid attention based multimodal network for spoken language classification. In: Proceedings of the Conference. Association for Computational Linguistics. Meeting, vol. 2018, pp. 2379–2390. NIH Public Access (2018)
53. Gu, Y., Yang, K., Fu, S., Chen, S., Li, X., Marsic, I.: Multimodal affective analysis using hierarchical attention strategy with word-level alignment. arXiv preprint arXiv:1805.08660 (2018)
54. Heraz, A., Frasson, C.: Predicting the three major dimensions of the learner's emotions from brainwaves. Int. J. Comput. Sci. **2**(3), 187–193 (2007)
55. Hou, L., Ji, H., Shen, Z.: Recovering over-/underexposed regions in photographs. SIAM J. Imaging Sci. **6**(4), 2213–2235 (2013)
56. Howard, A., Zhang, C., Horvitz, E.: Addressing bias in machine learning algorithms: a pilot study on emotion recognition for intelligent systems. In: 2017 IEEE Workshop on Advanced Robotics and Its Social Impacts (ARSO), pp. 1–7. IEEE (2017)
57. Huang, D., De la Torre, F.: Bilinear kernel reduced rank regression for facial expression synthesis. In: European Conference on Computer Vision, pp. 364–377. Springer, Berlin (2010)
58. Izard, C.E.: The Face of Emotion. Appleton-Century Crofts, New York (1971)
59. Jack, R.E., Garrod, O.G., Yu, H., Caldara, R., Schyns, P.G.: Facial expressions of emotion are not culturally universal. Proc. Natl. Acad. Sci. **109**(19), 7241–7244 (2012)
60. Keltner, D., Ekman, P., Gonzaga, G., Beer, J.: Facial Expression of Emotion. Guilford Publications, New York (2000)
61. Khamis, M., Baier, A., Henze, N., Alt, F., Bulling, A.: Understanding face and eye visibility in front-facing cameras of smartphones used in the wild. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, p. 280. ACM (2018)
62. Kheradpisheh, S.R., Ghodrati, M., Ganjtabesh, M., Masquelier, T.: Deep networks can resemble human feed-forward vision in invariant object recognition. Sci. Rep. **6**, 32672 (2016)
63. Kim, Y., Lee, H., Provost, E.M.: Deep learning for robust feature generation in audiovisual emotion recognition. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 3687–3691. IEEE (2013)
64. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
65. Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D.H., Hawk, S.T., Van Knippenberg, A.: Presentation and validation of the radboud faces database. Cogn. Emot. **24**(8), 1377–1388 (2010)
66. Le, H.V., Mayer, S., Wolf, K., Henze, N.: Finger placement and hand grasp during smartphone interaction. In: Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems, pp. 2576–2584. ACM (2016)
67. Lewinski, P., den Uyl, T.M., Butler, C.: Automated facial coding: validation of basic emotions and FACS AUs in facereader. J. Neurosci. Psychol. Econ. **7**(4), 227 (2014)
68. Lewis, M., Haviland-Jones, J.M., Barrett, L.F.: Handbook of Emotions. Guilford Press, New York (2010)
69. Litman, D.J., Forbes-Riley, K.: Predicting student emotions in computer-human tutoring dialogues. In: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, p. 351. Association for Computational Linguistics (2004)
70. Liu, C., Freeman, W.T., Szeliski, R., Kang, S.B.: Noise estimation from a single image. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), vol. 1, pp. 901–908. IEEE (2006)
71. Ma, C., Osherenko, A., Prendinger, H., Ishizuka, M.: A chat system based on emotion estimation from text and embodied conversational messengers. In: Proceedings of the 2005 International Conference on Active Media Technology, 2005. (AMT 2005), pp. 546–548. IEEE (2005)
72. Maalej, A., Amor, B.B., Daoudi, M., Srivastava, A., Berretti, S.: Local 3d shape analysis for facial expression recognition. In: 2010 20th International Conference on Pattern Recognition, pp. 4129–4132. IEEE (2010)
73. Mao, X., Xue, Y., Li, Z., Huang, K., Lv, S.: Robust facial expression recognition based on RPCA and AdaBoost. In: 2009 10th Workshop on Image Analysis for Multimedia Interactive Services, pp. 113–116. IEEE (2009)
74. Matsumoto, D., Ekman, P.: Facial expression analysis. Scholarpedia **3**(5), 4237 (2008)
75. Matsumoto, D., Keltner, D., Shiota, M.N., O'Sullivan, M., Frank, M.: Facial expressions of emotion. In: Lewis, M., Haviland-Jones, J.M., Barrett, L.F. (eds.) Handbook of Emotions, vol. 3, pp. 211–234. Guilford Press, New York (2008)
76. Matthews, O., Sarsenbayeva, Z., Jiang, W., Newn, J., Velloso, E., Clinch, S., Goncalves, J.: Inferring the mood of a community from their walking speed: a preliminary study. In: Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing, UbiComp' 18, pp. 1144–1149 (2018)
77. McDaniel, B., D'Mello, S., King, B., Chipman, P., Tapp, K., Graesser, A.: Facial features for affective state detection in learning environments. In: Proceedings of the Annual Meeting of the Cognitive Science Society, vol. 29 (2007)
78. Microsoft: Face API—facial recognition software–microsoft azure. https://azure.microsoft.com/en-au/services/cognitive-services/face/ (2019)

79. Narwekar, A.A., Girju, R.: Uiuc at semeval-2018 task 1: recognizing affect with ensemble models. In: Proceedings of The 12th International Workshop on Semantic Evaluation, pp. 377–384 (2018)
80. Nelson, N.L., Russell, J.A.: Universality revisited. Emot. Rev. **5**(1), 8–15 (2013)
81. Olszanowski, M., Pochwatko, G., Kuklinski, K., Scibor-Rylski, M., Lewinski, P., Ohme, R.K.: Warsaw set of emotional facial expression pictures: a validation study of facial display photographs. Front. Psychol. **5**, 1516 (2015)
82. Opencv: Cascade Classifier Training. https://docs.opencv.org/3.1.0/dc/d88/tutorial_traincascade.html#gsc.tab=0
83. Panigrahi, S.K., Gupta, S., Sahu, P.K.: Phases under Gaussian additive noise. In: 2016 International Conference on Communication and Signal Processing (ICCSP), pp. 1771–1776. IEEE (2016)
84. Patton, R.: Software Testing. Pearson Education India, New Delhi (2006)
85. Poria, S., Cambria, E., Bajpai, R., Hussain, A.: A review of affective computing: from unimodal analysis to multimodal fusion. Inf. Fusion **37**, 98–125 (2017)
86. Poria, S., Cambria, E., Gelbukh, A.: Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 2539–2544 (2015)
87. Poria, S., Chaturvedi, I., Cambria, E., Hussain, A.: Convolutional MKL based multimodal emotion recognition and sentiment analysis. In: 2016 IEEE 16th International Conference on Data Mining (ICDM), pp. 439–448. IEEE (2016)
88. Rodner, E., Simon, M., Fisher, R.B., Denzler, J.: Fine-grained recognition in the noisy wild: sensitivity analysis of convolutional neural networks approaches. arXiv preprint arXiv:1610.06756 (2016)
89. Rodriguez, P., Cucurull, G., Gonzàlez, J., Gonfaus, J.M., Nasrollahi, K., Moeslund, T.B., Roca, F.X.: Deep pain: Exploiting long short-term memory networks for facial expression classification. IEEE Trans. Cybern. 2017. https://doi.org/10.1109/TCYB.2017.2662199
90. Russell, J.A.: Is there universal recognition of emotion from facial expression? a review of the cross-cultural studies. Psychol. Bull. **115**(1), 102 (1994)
91. Sander, D., Scherer, K.: Oxford Companion to Emotion and the Affective Sciences. Oxford University Press, Oxford (2014)
92. Sarsenbayeva, Z., Ferreira, D., van Berkel, N., Luo, C., Vaisanen, M., Kostakos, V., Goncalves, J.: Vision-based happiness inference: a feasibility case-study. In: Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing, UbiComp '17, pp. 494–499. ACM, New York, NY, USA (2017)
93. Sarsenbayeva, Z., Marini, G., van Berkel, N., Luo, C., Jiang, W., Yang, K., Wadley, G., Dingler, T., Kostakos, V., Goncalves, J.: Does smartphone use drive our emotions or vice versa? a causal analysis. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, CHI' 20, pp. 1–15. New York, NY, USA (2020)
94. Schuller, B., Stadermann, J., Rigoll, G.: Affect-robust speech recognition by dynamic emotional adaptation. In: Proceedings of Speech Prosody 2006, Dresden (2006)
95. Sharma, P., Esengönül, M., Khanal, S.R., Khanal, T.T., Filipe, V., Reis, M.J.: Student concentration evaluation index in an e-learning context using facial emotion analysis. In: International Conference on Technology and Innovation in Learning, Teaching and Education, pp. 529–538. Springer, Berlin (2018)
96. Stöckli, S., Schulte-Mecklenbeck, M., Borer, S., Samson, A.C.: Facial expression analysis with affdex and facet: a validation study. Behav. Res. Methods **50**(4), 1446–1460 (2018)
97. Swinton, R., El Kaliouby, R.: Measuring emotions through a mobile device across borders, ages, genders and more. In: Proceedings of the ESOMAR Congress, Atlanta, pp. 1–12 (2012)
98. Tan, X., Triggs, W.: Enhanced local texture feature sets for face recognition under difficult lighting conditions. IEEE Trans. Image Process. **19**(6), 1635–1650 (2010)
99. Technology, M.: Emotion recognition—face++ AI open platform. https://www.faceplusplus.com/emotion-recognition/ (2019)
100. Torralba, A., Fergus, R., Freeman, W.T.: 80 million tiny images: a large data set for nonparametric object and scene recognition. IEEE Trans. Pattern Anal. Mach. Intell. **30**(11), 1958–1970 (2008)
101. Towner, H., Slater, M.: Reconstruction and recognition of occluded facial expressions using PCA. In: International Conference on Affective Computing and Intelligent Interaction, pp. 36–47. Springer, Berlin (2007)
102. Useche, O., El-Sheikh, E.: An intelligent system framework for measuring attention levels of students in online course environments. In: Proceedings on the International Conference on Artificial Intelligence (ICAI), p. 452. The Steering Committee of The World Congress in Computer Science, Computer... (2015)
103. Valstar, M.F., Jiang, B., Mehu, M., Pantic, M., Scherer, K.: The first facial expression recognition and analysis challenge. In: Face and Gesture 2011, pp. 921–926. IEEE (2011)
104. van Berkel, N., Goncalves, J., Koval, P., Hosio, S., Dingler, T., Ferreira, D., Kostakos, V.: Context-informed scheduling and analysis: improving accuracy of mobile self-reports. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI' 19 (2019)
105. van der Haar, D.T.: Student emotion recognition in computer science education: a blessing or curse? In: International Conference on Human–Computer Interaction, pp. 301–311. Springer, Berlin (2019)
106. Van Der Schalk, J., Hawk, S.T., Fischer, A.H., Doosje, B.: Moving faces, looking places: validation of the Amsterdam dynamic facial expression set (ADFES). Emotion **11**(4), 907 (2011)
107. Violante, M.G., Marcolin, F., Vezzetti, E., Ulrich, L., Billia, G., Di Grazia, L.: 3d facial expression recognition for defining users' inner requirements-an emotional design case study. Appl. Sci. **9**(11), 2218 (2019)
108. Visuri, A., Sarsenbayeva, Z., Goncalves, J., Karapanos, E., Jones, S.: Impact of mood changes on application selection. In: Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct, pp. 535–540. ACM (2016)
109. Yuan, L., Sun, J., Quan, L., Shum, H.Y.: Image deblurring with blurred/noisy image pairs. ACM Trans. Graph. (TOG) **26**(3), 1 (2007)
110. Zeng, Z., Pantic, M., Roisman, G.I., Huang, T.S.: A survey of affect recognition methods: audio, visual, and spontaneous expressions. IEEE Trans. Pattern Anal. Mach. Intell. **31**(1), 39–58 (2008)
111. Zhang, L., Verma, B., Tjondronegoro, D., Chandran, V.: Facial expression analysis under partial occlusion: a survey. ACM Comput. Surv. (CSUR) **51**(2), 25 (2018)
112. Zhang, Y., Chen, M., Huang, D., Wu, D., Li, Y.: idoctor: personalized and professionalized medical recommendations based on hybrid matrix factorization. Future Gener. Comput. Syst. **66**, 30–35 (2017)



Kangning Yang is currently a Ph.D. student at the School of Computing and Information Systems, University of Melbourne. His research interests are emotion recognition, human–computer interaction, and deep learning.

Chaofan Wang is currently a Ph.D. student at the School of Computing and Information Systems, University of Melbourne. His research interests are human–computer interaction, ubiquitous computing, and wearable sensors.

Zhanna Sarsenbayeva is a Postdoctoral Research Fellow in the School of Computing and Information Systems at the University of Melbourne. Her research interests include accessibility, ubiquitous computing, human-computer interaction, and affective computing. Zhanna holds a PhD in Engineering from the University of Melbourne, and during her PhD, she investigated the effects of different situational impairments on mobile interaction performance.

Benjamin Tag is an early career researcher who completed his Ph.D. at the Graduate School of Media Design at KEIO University in Japan, in March 2019. He is currently a Research Fellow at the School of Computing and Information Systems, University of Melbourne. His research interest is located in the fields of ubiquitous computing and cognition-aware systems. He is investigating ways to understand human cognition by combining methods from the fields of cognitive psychology and pervasive computing. His recent research focuses on digital emotion regulation, cognitive biases, and the application of digital nudges to improve media literacy among technology users.

Tilman Dingler is a research fellow in the School of Computing and Information Systems at the University of Melbourne. He studied media computer science in Munich, Web Science in San Francisco, and received a Ph.D. in computer science from the University of Stuttgart, Germany, in 2016. Tilman is an associate editor for the PACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT) and serves as associate chair for CHI among others. He is co-founder of the SIGCHI Melbourne Local Chapter. Tilman's research focuses on cognition-aware systems and technologies that support users' information processing capabilities.

Greg Wadley is a senior lecturer in the School of Computing and Information Systems at the University of Melbourne. He holds degrees in computer science, cognitive science, and human–computer interaction. His research involves designing and evaluating technology interventions as well as studying the user experience and social impact of digital technologies. He is a part of a team at Melbourne, Stanford, and University College London studying how people use technologies such as smartphones to shape their emotions in daily life. More details can be found at http://people.eng.unimelb.edu.au/gwadley/.

Jorge Goncalves is a senior lecturer at the School of Computing and Information Systems at the University of Melbourne, Australia. His research interests include ubiquitous computing, human–computer interaction, crowdsourcing, affective computing, and social computing. He is a member of the Association for Computing Machinery. Contact him at jorge.goncalves@unimelb.edu.au.