

## EXPERIMENTAL EVALUATION OF CLOUD-BASED FACIAL EMOTION RECOGNITION SERVICES

Ján MAGYAR, Peter SINČÁK, Ivan ČÍK  
Andrinandrasana David RASAMOELINA

*Department of Cybernetics and Artificial Intelligence*

*Technical University of Košice*

*Vysokoškolská 4, 04200 Košice, Slovakia*

*e-mail: {jan.magyar, peter.sincak, ivan.cik, andrijdavid}@tuke.sk*

Cindy L. BETHEL

*Department of Computer Science and Engineering*

*Mississippi State University*

*665 George Perry St., Butler Hall, MS, USA*

*e-mail: cbethel@cse.msstate.edu*

Filippo CAVALLO

*Department of Industrial Engineering*

*University of Florence*

*Via Santa Martra 3, 50139 Firenze, Italy*

*e-mail: filippo.cavallo@unifi.it*

Ammar HAWBANI

*School of Computer Science and Technology*

*University of Science and Technology of China*

*JinZhai Road, Baohe District, Hefei, Anhui 230026, China*

*e-mail: anmande@ustc.edu.cn*

**Abstract.** The main goal of this paper is to perform an extensive analysis of the accuracy of six selected cloud-based facial emotion recognition services on three facial images datasets. The evaluation was performed on more than 33 000 images depicting eight different emotions. Results show that emotion recognition services show a varying level of accuracy over different types of datasets, having a lower accuracy for images of lower quality, but performing considerably better for images taken in ideal conditions. Based on these results we believe that cloud-based facial emotional recognition services do not have the expected accuracy for some use cases and therefore must be selected with care when developing a system that relies on emotion-based interactions.

**Keywords:** Affective computing, cloud computing, facial emotion recognition, Software as a Service

## 1 INTRODUCTION

Information about human emotions can be valuable in many aspects of our daily lives and is innate to human interaction. Any computer system that aims to assist or replace certain human activities and capabilities must therefore be equipped with the ability to recognize and emulate emotions. This makes the development of systems that operate in an affective loop with humans challenging. Knowledge about the emotional state of humans can be used in a number of applications, such as:

- detecting health problems related to the mental state of a human, such as depression, anxiety, etc. [1, 2, 3, 4];
- improving the quality of human–robot or human–computer interaction (e.g. conversation with a chatbot or robot) [5, 6, 7];
- assessing mood during the work day in correlation with work performance [8, 9, 10].

The need for systems that are able to recognize human emotions prompted extensive research in the field of facial emotion recognition. A number of ready-to-use solutions have been made available in recent years thanks to cloud computing, making the development of human-centered systems easier.

In this paper, we evaluate six selected cloud-based facial emotion recognition services (Amazon Rekognition, Face++, Google Vision AI, Microsoft Face API, Sightcorp F.A.C.E. API, Sighthound Cloud API) on more than 33 000 facial images with the aim to assess their readiness for use in real-life applications. Although the developers of these services provide some information with regard to the performance of their services, it can be insufficient and misleading. Here, we perform a more detailed analysis which can serve to researchers and developers as an example of

how to evaluate facial emotion recognition services, and how to choose the one most suited for a specific application.

Such detailed comparative analysis of facial emotion recognition services is rare in literature. Al-Omair and Huang [11] compared three services (Amazon Rekognition, Google Vision AI, and Microsoft Face API) using the Karolinska Directed Emotional Faces (KDEF) dataset. They evaluated the services only on emotions they support, and provided a confusion matrix for them, observing an accuracy of 64% for Amazon (for six emotions), 76% for Microsoft (for seven emotions), and 85% for Google (for four emotions). However, they did not explore how unsupported emotion types are handled by the selected services.

Microsoft's and Google's service were further evaluated on the KDEF dataset in [12]. The authors showed that while Microsoft FACE API outperformed Google Vision AI for most emotions, it had trouble detecting profile faces, while Google Vision AI was able to recognize some emotions like sorrow and surprise only from frontal images. However, neither of these papers analyzed the services' performance for each emotion in a detailed way.

The rest of the paper is structured as follows: Section 2 introduces the field of affective computing, and a number of emotion models used in affective computing; Section 3 lists the datasets used in our study; in Section 4 we describe the services we evaluated; Section 5 presents our methodology and the metrics used to evaluate the accuracy of the tested emotion recognition services; Section 6 contains an overview of the results; Section 7 concludes the paper and outlines future work.

## 2 AFFECTIVE COMPUTING

The importance of human emotion in engineering was described in [13] in 1995, where feelings were described as an important factor for human-centered technology, and the term *affective computing* was coined. The application potential of this new field was enormous, especially for building artificial intelligence into human-centered systems. A comprehensive review of machine-empathy problems can be found in [14].

By nature, humans can express various types of emotion, which can be presented to the outside world in a multi-modal way. Multi-modality is the representation of human emotions across multiple channels and can represent a human's internal emotional state. However, computer systems usually consider only one channel for emotion recognition, most often the visual one. An overview of multi-modality research in affective computing is presented in [15].

Incorporating emotion recognition and affective interaction in human-computer systems is a significant challenge since dealing with emotions comes naturally to most people but is hard to define in a machine-friendly way. Multiple emotion models exist [16] that formally describe human emotions and their perception. These models can be divided into three main groups: discrete, dimensional, and hybrid emotion models.

Discrete emotion models define a set of emotions that are thought to be universal across cultures. Discrete models consider emotions to be easily discernible from one another based on an individual’s facial expression, voice pitch, word choice, and further modalities. This also supposes that a person’s emotion can be determined from his or her brain activity and physiology alone [17].

One of the best-known examples of a discrete emotion model is the Ekman emotional model [18], which is the result of a cross-cultural study. It defines six basic emotions (anger, disgust, fear, happiness, sadness, surprise), all with particular characteristics and a unique set of expressions. The original model can be extended with further categories, and relationships between emotions can be defined [19]. Commercial emotion recognition services typically use emotion models based on Ekman’s model.

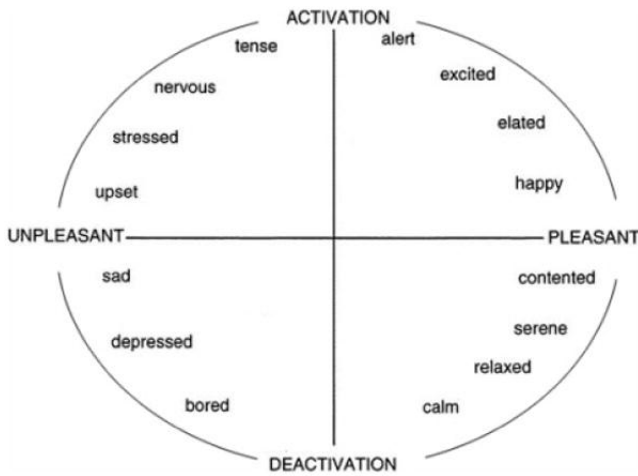


Figure 1. Circumplex model of affect [20]

Dimensional emotion models were defined because discrete models were not sufficient to describe the high variability of human emotions. Dimensional models instead represent emotions on a number of dimensions and describe them as a combination of independent systems, e.g. valence and arousal as in [20], where valence represents the emotion’s connotation (positive or negative) and arousal refers to its intensity (high or low), as shown in Figure 1.

By combining discrete and dimensional models, we arrive at hybrid emotion models that describe emotions using both approaches. The model presented in [21] first used valence to classify the emotion into one of two categories, and then used discrete categories. Another hybrid model is Plutchik’s model of emotion, which arranges emotions into concentric circles (see Figure 2) where inner circles comprise basic emotions and outer layers contain more complex emotions [23]. In this case,

emotions of the same valence are placed along the same axis and their intensity increases the closer they are to the circle's center.



Figure 2. Plutchik's wheel of emotions [22]

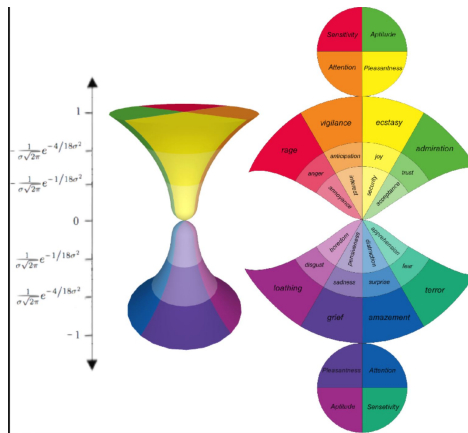


Figure 3. The Hourglass of Emotions model with the strength of the emotion on the vertical axis [24]

A more recent three-dimensional emotion model (presented in Figure 3) organizes emotions into the shape of an hourglass [24]. This model was inspired by Plutchik's wheel of emotions and describes emotions in four categories that define

the overall emotional state (pleasantness, attention, sensitivity, aptitude). The vertical axis in this three-dimensional space represents the strength of each emotion from strongly positive to strongly negative. The model was defined with an emphasis on its use in the context of human–computer interaction by making key indicators of the interaction easily measurable.

### 3 DATASETS USED FOR THE STUDY

For the purpose of testing facial emotion recognition services, we used images from three datasets that are freely available for academic research: the Karolinska Directed Emotional Faces (KDEF) [25], the Radboud Faces Database (RaFD) [26], and AffectNet [27]. The datasets contain images with the emotions anger, contempt, disgust, fear, happiness, neutral, sadness, and surprise present. The photos in KDEF and RaFD were taken in conditions that can be considered ideal for facial emotion recognition: the faces are well-lit and in focus, and the emotions were expressed clearly by models instructed to do so. AffectNet contains more varied images collected from multiple sources and annotated later, and represents a bigger challenge for emotion recognition algorithms, as shown in [27].

**KDEF** was developed in 1998 by Daniel Lundqvist, Anders Flykt, and Arne Öhman at Karolinska Institutet [25]. It contains 4 900 pictures taken of 70 individuals (35 male and 35 female) from the age range 20–30. The subjects who participated in the creation of the dataset received instructions describing the emotions to be expressed in advance so they could practice before the photo session. Models were expected to express emotions naturally and clearly. KDEF comprises images with seven emotions: anger, disgust, fear, happiness, neutral, sadness, and surprise. For each expression, photos were taken in two sessions from five different angles (two full profiles, two half profiles, and one frontal). For our study, we selected only frontal images for each emotion and model, resulting in 973 images (139 for each emotion), with one set of images of a male not available in the downloadable dataset.

**RaFD** dataset was created by the Behavioural Science Institute of the Radboud University Nijmegen [26]. It is a collection of pictures of 67 individuals of different ethnicities. The models displayed anger, contempt, disgust, fear, happiness, neutrality, sadness, and surprise; expressions were trained using the Facial Action Coding System [28]. Pictures were taken from five angles (two full profiles, two partial profiles, and one frontal) and with three gaze directions (left, front, and right). For our evaluation, we used only frontal images for each emotion and model with all three gaze directions. To be able to better compare results with the KDEF dataset, we did not select images with contempt, resulting in a set of 1 407 images (201 per emotion). Images selected from KDEF and RaFD together can be used to determine how well the services work under ideal conditions.

**AffectNet** [27] contains facial images collected from the Internet, from which 420 299 were manually labeled by 12 expert annotators at the University of

Denver. Discrepancies between annotators’ opinions suggest that facial emotion classification was not straightforward in some images; for a more detailed discussion please refer to [27]. Due to the large number of images selected for testing, however, we do not believe that such disagreements between annotators have a significant effect on the results presented in this paper.

Labels are available in both discrete and dimensional emotion models. Based on the discrete model, images were classified into ones depicting anger, contempt, disgust, fear, happiness, neutral, sadness, and surprise. In addition to these categories, annotators could label the images as not containing a face, none, or uncertain. The dimensional model describes the emotions present in the images on the dimensions valence and arousal, both from the interval  $\langle -1, 1 \rangle$ . For testing, we selected only images containing one emotion expression, 3 910 for each emotion (with the addition of the KDEF and RaFD it results in 4 250 images per emotion except for contempt), 31 280 images from AffectNet in total. For each emotion, the first 3 910 images were selected from the training and evaluation sets provided by the authors of the dataset.

Figure 4 shows some examples of the images used for testing for each emotion (all images are from AffectNet). The size of images used ranges between  $133 \times 133$  pixels and  $3 881 \times 3 881$  pixels. In later parts of the paper we refer to selected images by the name of the dataset they were drawn from.



Figure 4. Examples of the facial images in AffectNet selected for the comparison study

#### 4 FACIAL EMOTION RECOGNITION SERVICES

For this study, we selected commercial cloud-based computer vision services that support facial emotion recognition. We focused on services that categorized emotions in a way corresponding to the categories present in the datasets used for this research, namely anger, contempt, disgust, fear, happiness, neutral, sadness, and surprise.

Cloud-based emotion recognition services are available through an application programming interface (API), through which users can send images and get a result in the form of a JSON (JavaScript Object Notation) object; this transaction

constitutes a call. Apart from the results of emotion analysis, the services typically return information regarding the position of the face on the image, as well as the position of key facial landmarks.

The studied facial emotion recognition services use machine learning algorithms (deep learning especially) to understand images and are based on a discrete emotional model (see Section 2). The numerical output describes the confidence scores in a given emotion in the face but does not describe the intensity. The available documentation states that the services first identify the individual landmarks in the face and identify the presence of emotions based on the relationships between them. These APIs determine a given emotion based only on the presence of a human face in the image and do not determine the internal state of the human and should therefore not be used for such tasks (e.g. a person who is depressed may smile).

In this section we describe the six services selected for evaluation. For an overview of the emotions recognized by these services, please refer to Table 1.

Service	Recognized Emotions								Output Type
	A	C	D	F	H	N	Sa	Su	
Amazon Rekognition	✓		✓		✓	✓	✓	✓	$\langle 0.0, 100.0 \rangle$
Face++	✓		✓	✓	✓	✓	✓	✓	$\langle 0.0, 100.0 \rangle$
Google Vision AI	✓				✓		✓	✓	string values
Microsoft Face API	✓	✓	✓	✓	✓	✓	✓	✓	$\langle 0.0, 1.0 \rangle$
Sightcorp F.A.C.E. API	✓		✓	✓	✓		✓	✓	$\langle 0, 100 \rangle$
Sighthound Cloud API	✓		✓	✓	✓	✓	✓	✓	$\langle 0.0, 1.0 \rangle$

Table 1. List of detected emotions with the selected emotion recognition services (A – anger, C – contempt, D – disgust, F – fear, H – happiness, N – neutral, Sa – sadness, Su – surprise)

**Amazon Rekognition** [29] is offered as part of Amazon Web Services (AWS) and provides image and video analysis solutions. The facial emotion recognition module supports the recognition of anger, calmness (considered as neutral emotion for the purposes of comparison), confusion, disgust, happiness, sadness, and surprise. If the service fails to detect any of the listed emotions, the category Unknown is returned. According to the service’s official documentation, fear is also recognized but was never returned in our evaluation, therefore we did not consider it. Amazon Rekognition provides the confidence values for each emotion category as a real number between 0.0 and 100.0. The sum of the confidence values for all emotion categories equals 100.0.

**Face++** offers its emotion recognition service as part of their Cognitive Services [30]. It can detect seven emotions: anger, disgust, fear, happiness, neutral, sadness, and surprise. Confidence values are real numbers between 0.0 and 100.0, with a total sum of 100.0.



**Vision AI.** Google provides emotion recognition through their Vision AI service [31]. It detects four emotions: anger, joy, sorrow, and surprise. For the purpose of this paper, we considered joy to be equal to happiness, and sorrow to sadness. The confidence values for this system are discrete values (unknown, very unlikely, unlikely, possible, likely, very likely) rather than from a numeric interval. During our evaluation, we converted these discrete categories to numeric values with very unlikely being 0 and very likely 1. The interval was split into even parts, and the categories were given values accordingly (unlikely – 0.25, possible – 0.5, likely – 0.75), with unknown being represented as NULL. In cases where the service returned the same confidence value for more than one emotion, we considered that image to be incorrectly classified, since no single emotion was labeled as present or more likely than others.

**Microsoft Face API** is part of Microsoft’s Cognitive Services on the Azure cloud platform, and it can be used for face detection, face verification, and emotion recognition [32]. It supports all eight emotions present in the used datasets: anger, contempt, disgust, fear, happiness, neutral, sadness, and surprise. It provides confidence values as real numbers between 0.0 and 1.0 with a sum of 1.0.

**SightCorp F.A.C.E. API** [33] is a solution offering face detection, emotion analysis, attention analysis, and crowd demographics data. It can recognize six emotions, namely anger, disgust, fear, happiness, sadness, and surprise. Confidence values are integers between 0 and 100 and are independent of one another, thus their sum might not be equal to 100.

**Sighthound Cloud API** [34] is a multi-purpose solution that can be used for face detection and recognition, and vehicle recognition applications. The service recognizes anger, disgust, fear, happiness, neutral, sadness, and surprise. Confidence values are real numbers between 0.0 and 1.0, and their sum is 1.0.

## 5 METHODOLOGY

In this study, a comparison of commercial cloud-based facial emotion recognition services was performed on images depicting the following eight emotions: anger, contempt, disgust, fear, happiness, neutral, sadness, and surprise. These emotions are the most frequently supported by emotion recognition services and are also the most likely to be present in everyday human–machine interaction. This section describes the methodology and metrics used to compare the performance of each service.

From the point of view of machine learning, emotion recognition is a classification problem, and we used metrics associated with classification problems for the evaluation of the six emotion recognition services. First, we tested the services by sending to them the images selected from three source datasets as described in Section 3. Then, we looked at the emotion identified with the highest confidence value by the services, and we generated confusion matrices for each service. A confusion

matrix (CM) organizes the results of classification into a two-dimensional table with actual classes in columns and predicted classes in rows with identical sets of categories. Such a representation enables the detection of common misclassifications and systematic errors in classifiers.

If we consider a two-class classification of facial images that either contain a happy or non-happy face, we can create a confusion matrix similar to the one shown in Table 2. Within the table, each cell contains the number of instances from a given actual category classified as a given predicted category.

	Actual Happy	Actual Non-Happy
Predicted Happy	TP	FP
Predicted Non-Happy	FN	TN

Table 2. An example of a confusion matrix for two-class classification of happy/non-happy facial expressions

Table 2 contains the following cells:

- **true positive (TP)** – number of correctly classified images showing happiness,
- **false positive (FP)** – number of images showing a non-happy face incorrectly classified as ones showing happiness,
- **false negative (FN)** – number of images showing a happy face incorrectly classified as ones showing a non-happy face,
- **true negative (TN)** – number of correctly classified images showing a non-happy face.

A confusion matrix can be adjusted to show the results of a multi-class classification, with the correctly classified instances placed on the main diagonal (from top-left to bottom-right). Values under the diagonal are false negatives, while values above the diagonal are false positives. A perfect classifier's confusion matrix would have all values along the main diagonal with all other values being 0. A number of standard measures can be derived from a confusion matrix for evaluating classification.

A classifier's performance is most often expressed with its accuracy, but a single number cannot describe performance with the necessary depth. In this study we used accuracy ( $ACC$ ) to evaluate the services' overall performance, precision (P), recall (R), and  $F_1$  score to get more emotion-specific metrics.

**Precision** (or positive predicted value) is the fraction of relevant instances from the retrieved instances, and so describes the classification's reliability for a given class. It is computed using Formula (1):

$$P = \frac{TP}{TP + FP} \cdot 100\%. \quad (1)$$

**Recall** (or sensitivity) is the fraction of correctly retrieved relevant instances over the total amount of relevant instances, representing how well the classifier is able to recognize and correctly classify a single class. It can be computed using Formula (2):

$$R = \frac{TP}{TP + FN} \cdot 100\%. \quad (2)$$

**Accuracy** is the proportion of true results among the total set and can be computed using Formula (3), and so expresses the proportion of correctly classified instances. If  $ACC = 100\%$ , the classifier never makes an error.

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \cdot 100\%. \quad (3)$$

The  $F_1$  score (also F-score or F-measure) is the harmonic average of precision and recall, computed according to Formula (4). Its worst possible value is 0, while the best value 1 expresses perfect precision and recall. It is also possible to adjust it in a way that gives higher significance to either precision or recall.

$$F_1 = \frac{2}{\frac{1}{R} + \frac{1}{P}} = 2 \cdot \frac{P \times R}{P + R}. \quad (4)$$

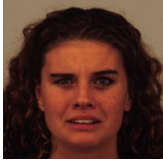

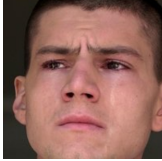
Alternatively, a weighted  $F_1$  score can be calculated by introducing a parameter representing the relative importance of recall over precision  $\beta$  (recall is  $\beta$  times as important as precision). The  $F_\beta$  score is then calculated as:

$$F_\beta = (1 + \beta^2) \cdot \frac{P \times R}{(\beta^2 \cdot P) + R}. \quad (5)$$

## 6 RESULTS AND DISCUSSION

We compared the performance of six facial emotion recognition services selected in Section 4 on images described in Section 3. This comparison was made first considering the overall accuracies of the considered services, and then we carried out a more detailed evaluation for each emotion. In this section we share the results of this analysis, point out the most important findings, and provide a guideline for how developers should interpret the results of such an analysis when selecting an emotion recognition service for their application. Sample misclassifications with confidence values provided by each service are shown in Table 3.

Table 4 shows the overall accuracies of each emotion recognition service over images selected from the various datasets. The accuracies were calculated with all emotions present in the selected images considered, irrespective of whether or not the given service was able to recognize them. For the accuracies calculated for only those emotions that the given service can recognize, please refer to the numbers in parantheses. If the two numbers are equal in a given cell, it means that the

KDEF (fear)	RaFD (anger)	AffectNet (sadness)
		

Service	Emotion (Confidence)		
Amazon Rekognition	calm/neutral (28.09)	confused (49.00)	disgusted (51.56)
Face++	sadness (96.90)	<b>anger (71.53)</b>	neutral (95.74)
Google Vision AI	joy/happiness and sorrow/sadness (LIKELY)	sorrow/sadness (UNLIKELY)	<b>sorrow/sadness (LIKELY)</b>
Microsoft Face API	sadness (0.95)	neutral (0.73)	neutral (0.89)
Sightcorp F.A.C.E. API	<b>fear (60)</b>	<b>anger (9)</b>	<b>sadness (84)</b>
Sighthound Cloud API	neutral (0.9999)	neutral (0.9640)	neutral (0.9866)

Table 3. Sample misclassifications from each dataset

service is able to recognize all emotions present in the images selected from the dataset.

	Dataset		
	KDEF	RaFD	AffectNet
Amazon Rekognition	63.618 (74.221)	65.473 (76.700)	35.189 ( <b>46.918</b> )
Face++	<b>76.156 (76.156)</b>	81.023 (81.023)	37.129 (42.433)
Google Vision AI	43.371 (75.899)	41.862 (73.259)	22.289 (44.578)
Microsoft Face API	<b>76.156 (76.156)</b>	<b>84.435 (84.435)</b>	<b>38.248 (38.248)</b>
Sightcorp F.A.C.E. API	62.282 (72.662)	69.723 (81.343)	23.916 (31.888)
Sighthound Cloud API	<b>76.156 (76.156)</b>	81.023 (81.023)	37.052 (42.346)

Table 4. Accuracy (in %) of emotion recognition services by dataset. Numbers in parentheses represent the accuracy calculated considering only images showing emotions the given service can recognize.

As expected, all services show higher accuracy over images from KDEF and RaFD that were taken under ideal conditions, where the faces are well-lit and the expressions are clear, while images from AffectNet come from various sources and show faces from different angles under different conditions. Google Vision AI had the

lowest accuracy across all sets of images, which is due to its being able to recognize four emotions compared to other services. If we consider accuracies normalized for the services' capabilities, we see that all services reached an accuracy around 75 % for images from KDEF, with Face++, Microsoft Face API, and Sighthound Cloud API having the highest accuracy of 76.156 %. The lowest normalized accuracy was achieved by Sightcorp F.A.C.E. API at 72.662 %.

Images selected from the RaFD dataset depict the same emotions as those from KDEF, however, they show people of different ethnicities and age groups. Despite this, all services apart from Google Vision AI had a higher accuracy (both for emotions in the dataset, and normalized accuracies). Microsoft Face API had the highest accuracy with 84.435 %, but Face++ and Sighthound Cloud API both reached an accuracy higher than 80 % at 81.023 %. The biggest difference can be observed for Sightcorp F.A.C.E. API with 72.662 % normalized accuracy for KDEF images, and 81.343 % for RaFD images, which was the second best accuracy from all evaluated services.

For images from AffectNet, all services showed an inferior performance. The highest accuracy when considering all emotions was reached by Microsoft Face API at 38.248 %, which is about half of the accuracy observed over KDEF images. Once again, Face++ and Sighthound Cloud API were the services that got close to Microsoft Face API's performance at accuracies somewhat bigger than 37 %. The lowest accuracy was reached by Google Vision AI (22.289 %), which is smaller than Sightcorp F.A.C.E. API's accuracy (23.916 %) only by little, even though the latter is able to recognize disgust and fear in addition to the emotions supported by Google Vision AI.

As for the normalized emotions, Amazon Rekognition was the best at 46.918 %. From the emotions present in the image set, it was unable to detect contempt and fear. Google Vision AI had a normalized accuracy of 44.578 %. The big difference between overall and normalized accuracy is due to the fact that Google supports only four emotions from the eight present in the dataset. We believe that the accuracy would have been higher still if we had accepted classification results where the emotion present in the image was recognized with a high probability along with another emotion as correct. Face++ and Sighthound Cloud API had a somewhat lower normalized accuracy at around 42 % (both unable to recognize contempt). From the six evaluated services only Microsoft Face API was able to recognize all emotions present, albeit at a lower normalized accuracy. Of all the tested services, Sightcorp F.A.C.E. API performed the worst on AffectNet when considering its normalized accuracy.

Besides accuracy, we further looked at the class-specific metrics precision, recall, and  $F_1$  score, a detailed discussion of which now follows. We believe that it is exactly these metrics that can best help developers when selecting the service most appropriate for their application. *Precision* should be used when evaluating the service's predictive strength, i.e., how reliable the service's prediction is. This measure is especially important when the cost of misclassifications is high, and developers prefer having fewer, but correct classifications. *Recall*, on

the other hand, should be considered primarily if the expectation of the recognition system is that it will detect a specific emotion with large certainty, with a small cost of misclassifications for other emotions. For example, in an intervention system, it might be crucial that the service be able to recognize anger on all occasions. For a balanced view of both precision and recall, the  $F_1$  score should be considered, which is the harmonic average of precision and recall. Alternatively, a weighted measure can be calculated from precision and recall, as described in Section 5.

In the following, we present overview tables showing results for each emotion per dataset. The confusion matrices for the tested services along with all of the results for each emotion recognition system are available online<sup>1</sup>. For a quick overview of classification accuracy per emotion please refer to Table 5, which shows the calculated  $F_1$  score for each emotion and service over all tested images.

		Service					
		Amazon Rekognition	Face++	Google Vision AI	Microsoft Face API	Sightcorp F.A.C.E.	Sighthound Cloud API
Emotion	A	<b>0.360</b>	0.351	0.167	0.324	0.301	0.350
	C	NaN	NaN	NaN	<b>0.090</b>	NaN	NaN
	D	0.367	<b>0.412</b>	NaN	0.351	0.277	<b>0.412</b>
	F	NaN	0.308	NaN	<b>0.346</b>	0.237	0.308
	H	<b>0.671</b>	0.601	0.574	0.660	0.546	0.599
	N	0.368	<b>0.384</b>	NaN	<b>0.384</b>	NaN	<b>0.384</b>
	Sa	0.438	0.395	0.431	<b>0.483</b>	0.285	0.393
	Su	0.434	0.451	<b>0.488</b>	0.425	0.286	0.450
Overall		0.330	0.363	0.208	<b>0.383</b>	0.241	0.362

Table 5. Overall  $F_1$  score of individual services per emotion over images from the three datasets (A – anger, D – disgust, F – fear, H – happiness, N – neutral, Sa – sadness, Su – surprise). NaN signifies that the service was unable to predict the given emotion.

## 6.1 KDEF Results

Tables 6, 7 and 8 show the calculated metrics for images selected from the KDEF dataset. Services performed best for images depicting happiness; they reached the highest average precision (87.054%), recall (98.442%), and  $F_1$  score (0.923). From the six considered services, Microsoft Face API could recognize happiness the most, with a recall of 100% – all images depicting happiness were classified correctly. Precision was also the highest for the emotion with Microsoft Face API at 95.205% (only 7 images from other categories were classified as happiness). The resulting  $F_1$  score was 0.975. Of the four emotions recognized by all services, surprise was the one recognized the second most with the average recall of 92.806%. Here, Google Vision

<sup>1</sup> <https://github.com/ianmagyar/cloud-based-fer-analysis>

		Service					
		Amazon Rekognition	Face++	Google Vision AI	Microsoft Face API	Sightcorp F.A.C.E.	Sighthound Cloud API
Emotion	A	56.800	90.698	28.916	<b>92.593</b>	60.131	90.698
	C	–	–	–	–	–	–
	D	77.907	67.978	NaN	<b>92.727</b>	88.889	67.978
	F	NaN	88.235	NaN	<b>96.667</b>	58.407	88.235
	H	88.889	88.462	79.429	<b>95.205</b>	81.875	88.462
	N	57.604	<b>74.566</b>	NaN	62.332	NaN	<b>74.566</b>
	Sa	<b>71.533</b>	68.519	38.318	64.103	43.629	68.519
	Su	69.318	72.826	<b>78.161</b>	70.968	65.169	72.826

Table 6. Overview of precision (in %) per emotion over images from the KDEF dataset (A – anger, D – disgust, F – fear, H – happiness, N – neutral, Sa – sadness, Su – surprise). NaN signifies that the service was unable to predict the given emotion.

		Service					
		Amazon Rekognition	Face++	Google Vision AI	Microsoft Face API	Sightcorp F.A.C.E.	Sighthound Cloud API
Emotion	A	51.079	65.116	17.266	53.957	<b>66.187</b>	56.115
	C	–	–	–	–	–	–
	D	48.201	<b>87.050</b>	0.000	73.381	63.309	<b>87.050</b>
	F	0.000	21.583	0.000	20.863	<b>47.482</b>	21.583
	H	97.842	99.281	<b>100.000</b>	<b>100.000</b>	94.245	99.281
	N	89.928	92.806	0.000	<b>100.000</b>	0.000	92.806
	Sa	70.504	79.856	88.489	<b>89.928</b>	81.295	79.856
	Su	87.770	96.403	<b>97.842</b>	94.964	83.453	96.403

Table 7. Overview of recall (in %) per emotion over images from the KDEF dataset (A – anger, D – disgust, F – fear, H – happiness, N – neutral, Sa – sadness, Su – surprise). 0.000 signifies that the service was unable to recognize the given emotion.

		Service					
		Amazon Rekognition	Face++	Google Vision AI	Microsoft Face API	Sightcorp F.A.C.E.	Sighthound Cloud API
Emotion	A	0.538	<b>0.693</b>	0.216	0.682	0.630	<b>0.693</b>
	C	–	–	–	–	–	–
	D	0.596	0.763	NaN	<b>0.819</b>	0.739	0.763
	F	NaN	0.347	NaN	0.343	<b>0.524</b>	0.347
	H	0.932	0.936	0.885	<b>0.975</b>	0.876	0.936
	N	0.702	<b>0.827</b>	NaN	0.768	NaN	<b>0.827</b>
	Sa	0.710	0.738	0.535	<b>0.749</b>	0.568	0.738
	Su	0.775	0.830	<b>0.869</b>	0.812	0.732	0.830

Table 8. Overview of  $F_1$  score per emotion over images from the KDEF dataset (A – anger, D – disgust, F – fear, H – happiness, N – neutral, Sa – sadness, Su – surprise). NaN signifies that the service was unable to predict the given emotion.

AI performed the best at 97.842% recall. Average precision for images with surprise stood at 71.545% (precision was also highest for Google Vision AI at 78.161%), and the average  $F_1$  score was 0.808 (Google Vision AI's was 0.869). For images with sadness, the services had a high average recall (81.655%), but average precision was 59.104%, suggesting an overprediction of sadness among the tested images. Of individual services, Amazon Rekognition had 71.533% precision, while Microsoft Face API had 89.928% recall and 0.749  $F_1$  score. Anger was the last emotion supported by all six evaluated emotions, but they recognized this emotion considerably worse with an average  $F_1$  score of 0.583. Recall was at 51.62%, with a slightly higher precision at 69.973%. Sightcorp F.A.C.E. had the highest recall (66.187%), Microsoft Face API the highest precision (92.593%), and Face++ the highest  $F_1$  score (0.693).

Of the three further emotions present in the dataset, disgust is supported by five of the selected services. Both average precision and recall were above 70%, 79.096% and 71.798%, respectively. Microsoft Face API had the highest precision for the emotion at 92.727%, and Face++ the highest recall with 87.05%. The average  $F_1$  score was 0.736 with Microsoft Face API's being the highest at 0.819. Neutral emotion was recognized by four services (as calmness by Amazon Rekognition). Average precision was 67.267% and average recall was 93.885%. Once again, this suggests that services can recognize if the person on the image is calm, but somewhat overpredict the emotion. This fact is reflected by the  $F_1$  score of 0.781. With regard to precision, Face++ and Sighthound Cloud API both reached 74.566%, which also had the highest  $F_1$  score (0.827), while only Microsoft Face API reached a recall of 100% among all services. Detecting fear was supported by four services, which on average predicted it on a level comparable to happiness and disgust at 82.886%. However, the average recall was only 27.878%, which means that even though services could predict fear with great certainty, they failed to correctly recognize it in most images with this emotion, resulting in an  $F_1$  score of 0.39. The best example of such behavior is Microsoft Face API, which had the highest precision for the emotion at 96.667%, but the recall was only 20.863%, meaning that while the service misclassified an image as one depicting fear only once, it could correctly recognize the emotion fifth of the time. Recall for the emotion was highest for Sightcorp F.A.C.E., but even that failed to recognize more than half of the images (47.482%). Its  $F_1$  score thus remained at 0.524.

Based on the balanced  $F_1$  score, we can say that happiness was the emotion recognized the best, followed by surprise, both having clear facial expressions. The emotion recognized the least was fear. Microsoft Face API had the highest precision for four emotions (anger, disgust, fear, happiness), although the recall was low for anger, disgust and fear, resulting in overpredicting, and low precision for neutral, sad, and surprised expressions. For the remaining emotions, Face++ and Sighthound Cloud API had the highest precision for neutral, Amazon Rekognition for sadness, and Google Vision AI for surprise. Sightcorp F.A.C.E. had the highest recall for anger and fear, Sighthound Cloud API and Face++ had the highest recall



for disgust, Google Vision AI had the highest recall for surprise. Both Google Vision AI and Microsoft Face API correctly recognized all images depicting happiness. In addition to this, Microsoft Face API had a recall of 100% for neutral images, and also the highest recall for sadness.

### 6.2 RaFD Results

For images selected from the RaFD dataset (see Tables 9, 10 and 11), happiness was also the emotion recognized with the highest accuracy. The average precision was 97.612% (with Microsoft Face API having the best precision at 100%, and Sightcorp F.A.C.E. the worst at 93.897%), and the average recall reached 99.585%, resulting in an  $F_1$  score of 0.986. Three services – Amazon Rekognition, Google Vision AI, and Microsoft Face API – were able to correctly classify all images depicting happiness, while the other services all had a recall higher than 99%. Microsoft Face API turned out to be a perfect classifier for happiness over the selected RaFD images.

Of the four emotions commonly recognized by all tested services, surprise was once again the one classified the second best, at a level comparable to that over images from KDEF: 72.557% average precision, 99.005% average recall, and 0.833 average  $F_1$  score. Microsoft Face API had both the highest precision and recall, at 82.041% and 100%, respectively (the  $F_1$  score was 0.901). Sadness was recognized with higher accuracy over RaFD images compared to KDEF images: average precision was 67.579% (59.104% for KDEF), average recall was 84.743% (81.665% for KDEF), and average  $F_1$  score was 0.738 (0.673 for KDEF). For this emotion, Microsoft Face API performed once again the best at 89.216% precision, 90.547% recall and 0.899  $F_1$  score (the second-highest  $F_1$  score for this emotion was 0.778 reached both by Face++ and Sighthound Cloud API).

		Service					
		Amazon Rekognition	Face++	Google Vision AI	Microsoft Face API	Sightcorp F.A.C.E.	Sighthound Cloud API
Emotion	A	59.355	<b>88.235</b>	5.195	87.671	72.558	<b>88.235</b>
	C	–	–	–	–	–	–
	D	80.833	85.903	NaN	<b>99.476</b>	92.899	85.903
	F	NaN	88.679	NaN	<b>100.000</b>	64.331	88.679
	H	98.049	99.005	95.714	<b>100.000</b>	93.897	99.005
	N	<b>72.266</b>	70.803	NaN	59.118	NaN	70.803
	Sa	78.947	72.727	46.982	<b>89.216</b>	44.872	72.727
	Su	58.651	75.188	64.423	<b>82.041</b>	78.049	75.188

Table 9. Overview of precision (in %) per emotion over images from the RaFD dataset (A – anger, D – disgust, F – fear, H – happiness, N – neutral, Sa – sadness, Su – surprise). NaN signifies that the service was unable to predict the given emotion.

Anger was the emotion recognized the least from the four common emotions, with an average  $F_1$  score of 0.495, which was lowest across all emotions. Average

		Service					
		Amazon Rekognition	Face++	Google Vision AI	Microsoft Face API	Sightcorp F.A.C.E.	Sighthound Cloud API
Emotion	C	–	–	–	–	–	–
	D	48.259	<b>97.015</b>	0.000	94.527	78.109	<b>97.015</b>
	F	0.000	46.766	0.000	<b>74.129</b>	50.249	46.766
	H	<b>100.000</b>	99.005	<b>100.000</b>	<b>100.000</b>	99.502	99.005
	N	92.040	96.517	0.000	<b>100.000</b>	0.000	96.517
	Sa	74.627	83.582	89.055	<b>90.547</b>	87.065	83.582
	Su	99.502	99.502	<b>100.000</b>	<b>100.000</b>	95.522	99.502

Table 10. Overview of recall (in %) per emotion over images from the RaFD dataset (A – anger, D – disgust, F – fear, H – happiness, N – neutral, Sa – sadness, Su – surprise). 0.000 signifies that the service was unable to recognize the given emotion.

		Service					
		Amazon Rekognition	Face++	Google Vision AI	Microsoft Face API	Sightcorp F.A.C.E.	Sighthound Cloud API
Emotion	A	0.517	0.594	0.045	0.467	<b>0.750</b>	0.594
	C	–	–	–	–	–	–
	D	0.604	0.911	NaN	<b>0.969</b>	0.849	0.911
	F	NaN	0.612	NaN	<b>0.851</b>	0.564	0.612
	H	0.990	0.990	0.978	<b>1.000</b>	0.966	0.990
	N	0.810	<b>0.817</b>	NaN	0.743	NaN	<b>0.817</b>
	Sa	0.767	0.778	0.615	<b>0.899</b>	0.592	0.778
	Su	0.738	0.857	0.784	<b>0.901</b>	0.859	0.857

Table 11. Overview of  $F_1$  score per emotion over images from the RaFD dataset (A – anger, D – disgust, F – fear, H – happiness, N – neutral, Sa – sadness, Su – surprise). NaN signifies that the service was unable to predict the given emotion.

precision was 66.875% and average recall was 41.459%. Face++ and Sighthound Cloud API had the highest precision at 88.235%, although recall was about half of that at 44.776%, suggesting overprediction. Sightcorp F.A.C.E. was the service best equipped to recognize anger, correctly recognizing 77.612% of the images showing this emotion, and having an  $F_1$  score of 0.75. Google Vision AI, although supporting the recognition of anger, had a 5.195% precision and 3.98% recall. It was most often misclassified as sadness (95 images), or no one emotion was identified as the most likely (98 images).

In comparison with the KDEF dataset, disgust was recognized with higher precision (89.003% average) and recall (82.985% average) on images from RaFD with an  $F_1$  score of 0.849, making it the second-best recognized emotion for this dataset. Of the five services supporting the recognition of this emotion, Face++ had the highest recall at 97.015%, and Microsoft Face API the highest precision at 99.476%, the latter also having the highest  $F_1$  score of 0.969. Amazon Rekognition had both the lowest precision and recall, 80.833% and 48.259%, respectively. The neutral expres-

sion's recognition rate was comparable with results over KDEF, with 68.248% average precision (67.267% for KDEF), 96.269% average recall (93.885% for KDEF), and 0.797  $F_1$  score (0.781 for KDEF). Fear was recognized correctly more often than on images from KDEF; while the average precision did not change much (85.422% for RaFD, 82.886% for KDEF), there was a twofold increase in recall (54.478% from 27.878%), resulting in 0.66  $F_1$  score. Still, only anger was recognized less frequently on RaFD images than fear. For recognizing fear, Microsoft Face API was the best with 100% precision and 74.129% recall (the  $F_1$  score was 0.851).

Similarly to KDEF images, happiness was the emotion recognized with the highest accuracy from RaFD images, and fear and anger were once again the expressions recognized the least often. Microsoft Face API outperformed the other services more clearly and for more emotions: it had the highest precision for disgust, fear, happiness, sadness and surprise; and the highest recall for fear, happiness, neutral, sadness and surprise. It had a perfect  $F_1$  score of 1 for happiness, making it a perfect classifier over the dataset for this emotion, but the metric was high for disgust (0.969), surprise (0.901) and sadness (0.899) as well. Other services had higher precision only for anger and neutral. For anger, Face++ and Sighthound Cloud API had a precision of 88.235%, only slightly bigger than Microsoft Face API's 87.671%. When recognizing neutral expressions, Amazon Rekognition was more reliable in its predictions, however, Microsoft Face API classified all examples from the class correctly. It was only for the expression of anger, in fact, where Microsoft Face API did not have the highest precision nor recall.

### 6.3 AffectNet Results

As Tables 12, 13 and 14 show, all metrics had a significantly lower value for images from the AffectNet dataset, which show faces from various angles and under different lighting conditions. Of the four commonly supported emotions, the order by accuracy did not change: happiness was recognized with the highest accuracy (average  $F_1$  score was 0.585, this being the only emotion with an  $F_1$  score above 0.5), followed by surprise (0.376), sadness (0.372) and anger (0.29). Happiness was recognized relatively successfully with a recall of 77.685% on average. However, precision was down to 47.355%, hinting at overprediction of the emotion. While Microsoft Face API recognized happiness with the highest recall with 86.087% (closely followed by Google Vision AI at 85.422%), Amazon Rekognition was more reliable in its predictions with 57.482% precision, also having the highest  $F_1$  score of all the tested services at 0.648.

For the emotion surprise, Google Vision AI performed the best with 48.077% precision (42.304% on average across all services), and 42.532% recall (34.463% average), resulting in an  $F_1$  score of 0.451. For sadness, Microsoft Face API reached the highest precision at 58.801% (41.998% average), while Sightcorp F.A.C.E. had the highest recall at 39.105% (35.205% average), although it also had the lowest precision with 19.47%. On a balanced evaluation based on the  $F_1$  score, Microsoft Face

API was the most accurate for images depicting sadness (0.442  $F_1$  score). Similar data was observed for images depicting anger, with Microsoft Face API having the highest precision at 55.374 % (41.839 % on average), and Sightcorp F.A.C.E. having the highest recall (33.018 %; 23.951 % average) and lowest precision (22.972 %) at the same time. For recognizing anger, Amazon Rekognition had the highest  $F_1$  score at 0.343. Just like on images from RaFD, Google Vision AI had both low precision (37.049 %) and recall (11.304 %); of the 3 910 images selected from ImageNet depicting anger, it failed to identify a single prominent emotion on 2 538 images, predicted sadness on 449, happiness on 340, and surprise on 141.

		Service					
		Amazon Rekognition	Face++	Google Vision AI	Microsoft Face API	Sightcorp F.A.C.E.	Sighthound Cloud API
Emotion	A	47.283	44.264	37.049	<b>55.374</b>	22.972	44.094
	C	NaN	NaN	NaN	<b>44.318</b>	NaN	NaN
	D	48.047	52.082	NaN	<b>90.870</b>	25.594	52.082
	F	NaN	47.338	NaN	<b>87.799</b>	21.717	47.225
	H	<b>57.482</b>	45.287	40.816	50.708	44.700	45.139
	N	22.319	<b>24.234</b>	NaN	22.864	NaN	24.216
	Sa	47.533	41.555	43.258	<b>58.801</b>	19.470	41.368
	Su	43.282	41.387	<b>48.077</b>	46.664	33.055	41.357

Table 12. Overview of precision (in %) per emotion over images from the AffectNet dataset (A – anger, D – disgust, F – fear, H – happiness, N – neutral, Sa – sadness, Su – surprise). NaN signifies that the service was unable to predict the given emotion.

		Service					
		Amazon Rekognition	Face++	Google Vision AI	Microsoft Face API	Sightcorp F.A.C.E.	Sighthound Cloud API
Emotion	A	26.931	25.857	11.304	20.818	<b>33.018</b>	25.780
	C	0.000	0.000	0.000	<b>4.987</b>	0.000	0.000
	D	27.059	<b>27.187</b>	0.000	16.036	21.228	<b>27.187</b>
	F	0.000	<b>20.921</b>	0.000	18.772	20.639	20.895
	H	74.271	79.744	85.422	<b>86.087</b>	61.151	79.437
	N	81.023	71.560	0.000	<b>92.788</b>	0.000	71.483
	Sa	35.729	31.023	39.054	35.371	<b>39.105</b>	30.946
	Su	36.496	40.742	<b>42.532</b>	31.125	16.189	40.691

Table 13. Overview of recall (in %) per emotion over images from the AffectNet dataset (A – anger, D – disgust, F – fear, H – happiness, N – neutral, Sa – sadness, Su – surprise). 0.000 signifies that the service was unable to recognize the given emotion.

Of the emotions supported only by some services, neutrality was recognized the best, with an average precision of 23.408 % (all services performed similarly, with Amazon Rekognition having the lowest precision of 22.319 % and Face++ the

		Service					
		Amazon Rekognition	Face++	Google Vision AI	Microsoft Face API	Sightcorp F.A.C.E.	Sighthound Cloud API
Emotion	A	<b>0.343</b>	0.326	0.173	0.303	0.271	0.325
	C	NaN	NaN	NaN	<b>0.090</b>	NaN	NaN
	D	0.346	<b>0.357</b>	NaN	0.273	0.232	<b>0.357</b>
	F	NaN	0.290	NaN	<b>0.309</b>	0.212	0.290
	H	<b>0.648</b>	0.578	0.552	0.638	0.516	0.576
	N	0.350	0.362	NaN	<b>0.367</b>	NaN	0.362
	Sa	0.408	0.355	0.410	<b>0.442</b>	0.260	0.354
	Su	0.396	0.411	<b>0.451</b>	0.373	0.217	0.410

Table 14. Overview of  $F_1$  score per emotion over images from the AffectNet dataset (A – anger, D – disgust, F – fear, H – happiness, N – neutral, Sa – sadness, Su – surprise). NaN signifies that the service was unable to predict the given emotion.

highest with 24.234%), while recall was 79.213% on average, this being higher even than recall for happiness. On recall, Microsoft Face API was clearly the best with 92.788% (also higher than happiness). This means that neutral expressions were recognized on a level comparable to sadness and surprise ( $F_1$  score was 0.36). For disgust, the average precision was 53.735%, although Microsoft Face API was an outlier with a precision of 90.87%, which was the highest precision observed across all emotions and all services. Recall was however low at 23.739% on average and 27.187% at best for Face++ and Sighthound Cloud API, which also shared the best  $F_1$  score of 0.357. Microsoft Face API's recall was the lowest at 16.036%, clearly showing underprediction of the emotion. Fear was recognized at a comparable level, with the average precision of 51.02% (Microsoft Face API once again being an outlier at 87.799%), and the average recall of 20.307% (Face++ having the highest at 20.921%). There was no big difference in recall across services, and although Microsoft Face API had the lowest recall with 18.772%, it still had the highest  $F_1$  score with 0.309. The expression of contempt is recognized only by Microsoft Face API, although the results show that this is done at low accuracy: precision was at 44.318% and recall at 4.987%. Of the 3910 images with this emotion, only 195 were classified correctly, while 2102 were classified as showing a neutral expression, and 1562 as ones showing happiness.

When comparing these calculated metrics for each service, we see that Microsoft Face API had the highest precision for five emotions (anger, contempt, disgust, fear, sadness – being the only one recognizing contempt) and the highest recall for three (contempt, happiness, neutral). Precision was highest for happiness with Amazon Rekognition, and with Face++ for neutral expressions. Recall was highest with Sightcorp F.A.C.E. for anger and sadness, and with Face++ for disgust (with Sighthound Cloud API) and fear. Google Vision AI was best at recognizing surprise both with regard to precision and recall.

## 7 CONCLUSIONS AND FUTURE WORK

In this paper we provided a comparative analysis of six selected cloud-based facial emotion recognition services: Amazon Rekognition, Face++, Google Vision AI, Microsoft Face API, Sightcorp F.A.C.E. API, and Sighthound Cloud API. This evaluation was done using images from three different datasets. Two of these datasets – Karolinska Directed Emotional Faces (KDEF) and Radboud Faces Database (RaFD) – had frontal images taken under conditions ideal for emotion recognition, while the third one – AffectNet – had images from different sources, showing faces from different angles and different lighting conditions. We considered eight emotions: anger, contempt, disgust, fear, happiness, neutral, sadness, and surprise. We compared the services using four main metrics: accuracy for general comparison, and precision, recall, and  $F_1$  score for emotion-specific comparison.

We showed that there is a large difference in the accuracy of these services over the selected images. While all services had an accuracy above 70 % for images from KDEF for supported emotions, and a slightly higher accuracy around 80 % for images from RaFD, accuracy was around 40 % for images from AffectNet. From the six tested services, Microsoft Face API had the highest accuracy when considering eight emotions over all datasets: 76.156 % for KDEF (the same accuracy was reached by Face++ and Sighthound Cloud API), 84.435 % for RaFD, and 38.248 % for AffectNet. However, when adjusting accuracy for only emotions that the service was able to recognize, Amazon Rekognition outperformed Microsoft Face API for AffectNet with 46.918 % – although it is unable to recognize contempt and fear.

Happiness was the emotion recognized with the greatest accuracy across all image sets with an  $F_1$  score of 0.923 for KDEF, 0.986 for RaFD and 0.585 for AffectNet. Surprise was also recognized with a relatively high accuracy, with 0.808, 0.833 and 0.376  $F_1$  scores, respectively. Of the four emotions supported by all tested services, sadness was recognized fairly well ( $F_1$  scores of 0.673, 0.738 and 0.372), while anger was often misclassified, especially as neutral or sadness, resulting in lower  $F_1$  scores of 0.583, 0.495, and 0.29. Of further emotions supported only by some services, neutrality and disgust were recognized at a comparable level (0.781 and 0.736  $F_1$  score for KDEF, 0.797 and 0.849 for RaFD, and 0.360 and 0.313 for AffectNet), while fear was classified poorly (0.39 for KDEF, 0.66 for RaFD, 0.275 for AffectNet). Recognizing contempt was supported only by Microsoft Face API; we tested this capability on images from AffectNet with an  $F_1$  score of 0.09.

Since cloud-based facial emotion recognition services are constantly under development and improved upon, these values are subject to change. However, the approach presented in this paper can be reproduced by developers and researchers to select the facial emotion recognition services most suitable for their applications. Precision should be used as a primary measure when the service's prediction reliability is crucial, while recall should be preferred if correctly identifying an emotion is essential to the functioning of the application.

Future evaluations of emotion recognition solutions should use further datasets and emotion recognition systems including those that are not set up as cloud-based

services to determine if there are differences between these types of systems with regard to performance. Additionally, it would be beneficial to investigate approaches such as combining the strengths of each system to enhance emotion assessment capabilities to achieve higher levels of accuracy. Alternatively, the possibility of adjusting a cloud-based service's predictions to a specific set of images tailored to a concrete application could be explored.

## **Acknowledgment**

This research work was supported by the Slovak Research and Development Agency under the contract No. APVV-015-0731, supported from 07-2016 to 06-2020, by the National Research and Development Project Grant 1/0773/16 2016–2019 “Cloud Based Artificial Intelligence for Intelligent Robotics”, by the AI4EU project from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 825619 2019–2021, and the Maria Curie RISE LIFEBOTS Exchange Grant Agreement No. 824047 2019–2021.

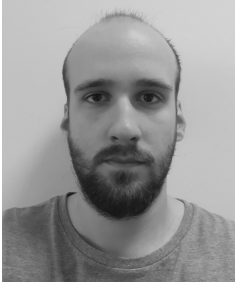
## **REFERENCES**

- [1] TOKUNO, S.—TSUMATORI, G.—SHONO, S.—TAKEI, E.—YAMAMOTO, T.—SUZUKI, G.—MITUYOSHI, S.—SHIMURA, M.: Usage of Emotion Recognition in Military Health Care. 2011 Defense Science Research Conference and Expo (DSR), IEEE, 2011, doi: 10.1109/DSR.2011.6026823.
- [2] ADAMS, E.: Models of Man, Social and Rational: Mathematical Essays on Rational Human Behavior in a Social Setting. *The Journal of Philosophy*, Vol. 59, 1962, No. 7, pp. 177–182, doi: 10.2307/2023734.
- [3] SIMCOCK, G.—MCLOUGHLIN, L. T.—DE REGT, T.—BROADHOUSE, K. M.—BEAUDEQUIN, D.—LAGOPOULOS, J.—HERMENS, D. F.: Associations Between Facial Emotion Recognition and Mental Health in Early Adolescence. *International Journal of Environmental Research and Public Health*, Vol. 17, 2020, No. 1, Art. No. 330, doi: 10.3390/ijerph17010330.
- [4] MANO, L. Y.—FAIÇAL, B. S.—NAKAMURA, L. H. V.—GOMES, P. H.—LIBRALON, G. L.—MENEQUETE, R. I.—FILHO, G. P. R.—GIANCRISTOFARO, G. T.—PESSIN, G.—KRISHNAMACHARI, B.—UEYAMA, J.: Exploiting IoT Technologies for Enhancing Health Smart Homes Through Patient Identification and Emotion Recognition. *Computer Communications*, Vol. 89-90, 2016, pp. 178–190, doi: 10.1016/j.comcom.2016.03.010.
- [5] CHEN, L.—ZHOU, M.—SU, W.—WU, M.—SHE, J.—HIROTA, K.: Softmax Regression Based Deep Sparse Autoencoder Network for Facial Emotion Recognition in Human-Robot Interaction. *Information Sciences*, Vol. 428, 2018, pp. 49–61, doi: 10.1016/j.ins.2017.10.044.
- [6] LIU, Z.—WU, M.—CAO, W.—CHEN, L.—XU, J.—ZHANG, R.—ZHOU, M.—MAO, J.: A Facial Expression Emotion Recognition Based Human-Robot Interaction

- System. *IEEE/CAA Journal of Automatica Sinica*, Vol. 4, 2017, No. 4, pp. 668–676, doi: 10.1109/JAS.2017.7510622.
- [7] PEREZ-GASPAR, L.-A.—CABALLERO-MORALES, S.-O.—TRUJILLO-ROMERO, F.: Multimodal Emotion Recognition with Evolutionary Computation for Human-Robot Interaction. *Expert Systems with Applications*, Vol. 66, 2016, pp. 42–61, doi: 10.1016/j.eswa.2016.08.047.
- [8] STAW, B. M.—SUTTON, R. I.—PELLED, L. H.: Employee Positive Emotion and Favorable Outcomes at the Workplace. *Organization Science*, Vol. 5, 1994, No. 1, pp. 51–71, doi: 10.1287/orsc.5.1.51.
- [9] ELFENBEIN, H. A.—FOO, M. D.—WHITE, J.—TAN, H. H.—AIK, V. C.: Reading Your Counterpart: The Benefit of Emotion Recognition Accuracy for Effectiveness in Negotiation. *Journal of Nonverbal Behavior*, Vol. 31, 2007, No. 4, pp. 205–223, doi: 10.1007/s10919-007-0033-7.
- [10] KRANEFELD, I.—BLICKLE, G.: Disentangling the Relation Between Psychopathy and Emotion Recognition Ability: A Key to Reduced Workplace Aggression? *Personality and Individual Differences*, Vol. 184, 2022, Art.No. 111232, doi: 10.1016/j.paid.2021.111232.
- [11] AL-OMAIR, O. M.—HUANG, S.: A Comparative Study on Detection Accuracy of Cloud-Based Emotion Recognition Services. *Proceedings of the 2018 International Conference on Signal Processing and Machine Learning (SPML '18)*, 2018, pp. 142–148, doi: 10.1145/3297067.3297079.
- [12] KHANAL, S. R.—BARROSO, J.—LOPES, N.—SAMPAIO, J.—FILIPE, V.: Performance Analysis of Microsoft's and Google's Emotion Recognition API Using Pose-Invariant Faces. *Proceedings of the 8<sup>th</sup> International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-Exclusion*, 2018, pp. 172–178, doi: 10.1145/3218585.3224223.
- [13] PICARD, R. W.: *Affective Computing*. Technical Report No. 321. M.I.T. Media Laboratory, Perceptual Computing Section, 1995. Available at: <http://www.media.mit.edu/~picard/>.
- [14] VIRČÍKOVÁ, M.: *Machine Empathy: Towards Artificial Emotional Intelligence with Active Personalization in Social Human-Robot Interaction*. Ph.D. Thesis, Technical University of Košice, Slovakia, 2014.
- [15] PORIA, S.—CAMBRIA, E.—BAJPAI, R.—HUSSAIN, A.: A Review of Affective Computing: From Unimodal Analysis to Multimodal Fusion. *Information Fusion*, Vol. 37, 2017, pp. 98–125, doi: 10.1016/j.inffus.2017.02.003.
- [16] GENDRON, M.—FELDMAN BARRETT, L.: Reconstructing the Past: A Century of Ideas about Emotion in Psychology. *Emotion Review*, Vol. 1, 2009, No. 4, pp. 316–339, doi: 10.1177/1754073909338877.
- [17] EKMAN, P.: An Argument for Basic Emotions. *Cognition and Emotion*, Vol. 6, 1992, No. 3–4, pp. 169–200, doi: 10.1080/02699939208411068.
- [18] EKMAN, P.: Basic Emotions. In: Dalglish, T., Power, M. (Eds): *Handbook of Cognition and Emotions*. Chapter 3. John Wiley and Sons, Hoboken, NJ, USA, 1999.
- [19] ROBERTS, K.—ROACH, M. A.—JOHNSON, J.—GUTHRIE, J.—HARABAGIU, S. M.: EmpaTweet: Annotating and Detecting Emotions on Twitter. *Proceedings of the*



- Eighth International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey, 2012, pp. 3806–3813.
- [20] POSNER, J.—RUSSELL, J. A.—PETERSON, B. S.: The Circumplex Model of Affect: An Integrative Approach to Affective Neuroscience, Cognitive Development, and Psychopathology. *Development and Psychopathology*, Vol. 17, 2005, No. 3, pp. 715–734, doi: 10.1017/S0954579405050340.
- [21] DIENER, E.—SMITH, H.—FUJITA, F.: The Personality Structure of Affect. *Journal of Personality and Social Psychology*, Vol. 69, 1995, No. 1, pp. 130–141, doi: 10.1037/0022-3514.69.1.130.
- [22] DREWS, M.: Robert Plutchik's Psychoevolutionary Theory of Basic Emotions. 2007, available at: <http://adliterate.com/archives/Plutchik.emotion.theorie.POSTER.pdf>.
- [23] PLUTCHIK, R.: A General Psychoevolutionary Theory of Emotion. In: Plutchik, R., Kellerman, H. (Eds.): *Emotion: Theory, Research and Experience, Volume 1: Theories of Emotion*. Chapter 1. Academic Press, 1980, pp. 3–33, doi: 10.1016/B978-0-12-558701-3.50007-7.
- [24] CAMBRIA, E.—LIVINGSTONE, A.—HUSSAIN, A.: The Hourglass of Emotions. Technical Report. Available at: <http://sentic.net>.
- [25] LUNDQVIST, D.—FLYKT, A.—ÖHMAN, A.: Karolinska Directed Emotional Faces (KDEF). 1998, doi: 10.1037/t27732-000.
- [26] LANGNER, O.—DOTSCH, R.—BIJLSTRA, G.—WIGBOLDUS, D. H. J.—HAWK, S. T.—VAN KNIPPENBERG, A.: Presentation and Validation of the Radboud Faces Database. *Cognition and Emotion*, Vol. 24, 2010, No. 8, pp. 1377–1388, doi: 10.1080/02699930903485076.
- [27] MOLLAHOSSEINI, A.—HASANI, B.—MAHOOR, M. H.: AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild. *IEEE Transactions on Affective Computing*, Vol. 10, 2019, No. 1, pp. 18–31, doi: 10.1109/TAFFC.2017.2740923.
- [28] EKMAN, P.—ROSENBERG, E. L.: *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*. Oxford University Press, USA, 1997.
- [29] Amazon Rekognition. Available at: <https://aws.amazon.com/rekognition/>.
- [30] Face++ Cognitive Services – Emotion Recognition. Available at: <https://www.faceplusplus.com/emotion-recognition/>.
- [31] Google Vision AI. Available at: <https://cloud.google.com/vision/>.
- [32] Microsoft Face API. Available at: <https://azure.microsoft.com/en-us/services/cognitive-services/face/>.
- [33] F.A.C.E. API by Sightcorp. Available at: <https://face-api.sightcorp.com>.
- [34] Sighthound Cloud API. Available at: <https://www.sighthound.com/products/cloud>.



**Ján MAGYAR** received his Ph.D. in intelligent systems from the Technical University of Košice, Slovakia in 2021. He is Deputy Head of the Center for Intelligent Technologies at the Department of Cybernetics and Artificial Intelligence, Technical University of Košice. His research focuses on the use of artificial intelligence in adaptive and personalized systems in education.



**Peter SINČÁK** earned his Ph.D. in cybernetics in 1992 at the Czech Academy of Sciences, he is Professor of artificial intelligence since 2001. He is Head of the Department of Cybernetics and Artificial Intelligence, Technical University of Košice, and Chairman of the Scientific Council of the Slovak.AI platform. He has published more than 100 papers and has given invited lectures in Japan, China, and member countries of the EU.



**Ivan ČÍK** received his M.Sc. degree in intelligent systems from the Department of Cybernetics and Artificial Intelligence, Technical University of Košice, Slovakia, in 2019. His thesis focused on reinforcement learning and cloud computing. He is finishing up his doctoral studies at the same department. His research focuses on explainable artificial intelligence.



**Andrinandrasana David RASAMOELINA** received his M.Sc. degree in software engineering and database management at the Ecole Nationale d'Informatique de Madagascar. He is currently pursuing his doctoral degree at the Department of Cybernetics and Artificial Intelligence, Technical University of Košice, Slovakia. His research focuses on few-shot learning in computer vision.



**Cindy L. BETHEL** is Professor in the Computer Science and Engineering Department and holds the Billie J. Ball Endowed Professorship in Engineering at the Mississippi State University (MSU). Her research interests include human-robot interaction, human-computer interaction, affective computing, robotics, and artificial intelligence. Her research focuses on applications associated with robotic therapeutic support, information gathering from children, and the use of robots for law enforcement, search and rescue, and military.



**Filippo CAVALLO** is Assistant Professor of biomedical robotics at the Department of Industrial Engineering, University of Florence. The objectives of his research are to promote and evaluate novel service robotics for active and healthy aging, to identify and validate disruptive healthcare paradigms for neurodegenerative and chronic diseases, focusing on prevention and support for physical and cognitive declines, to optimize the management of working life for improving efficiency, security and QoL of workers in industrial settings.



**Ammar HAWBANI** is Associate Professor of networking and communication algorithms in the School of Computer Science and Technology at the University of Science and Technology of China, Hefei, China. He received the B.Sc., M.Sc. and Ph.D. degrees in computer software and theory from the University of Science and Technology of China (USTC), Hefei, China, in 2009, 2012 and 2016, respectively. From 2016 to 2019, he worked as Postdoctoral Researcher in the School of Computer Science and Technology at USTC. His research interests include IoT, WSNs, WBANs, WMNs, VANETs, and SDN.