

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/329718637>

# On the Performance Analysis of APIs Recognizing Emotions from Video Images of Facial Expressions

Conference Paper · December 2018

DOI: 10.1109/ICMLA.2018.00040

CITATIONS

5

READS

1,476

4 authors:



**Ananya Bhattacharjee**

University of Toronto

23 PUBLICATIONS 71 CITATIONS

[SEE PROFILE](#)



**Tanmoy Sarkar Pias**

The University of Asia Pacific

22 PUBLICATIONS 76 CITATIONS

[SEE PROFILE](#)



**Mahathir Ahmad**

Bangladesh University of Engineering and Technology

1 PUBLICATION 5 CITATIONS

[SEE PROFILE](#)



**Ashikur Rahman**

Bangladesh University of Engineering and Technology

9 PUBLICATIONS 49 CITATIONS

[SEE PROFILE](#)

Some of the authors of this publication are also working on these related projects:



ServiceHub [View project](#)



Gender Recognition from Sensor signals [View project](#)

# On the Performance Analysis of APIs Recognizing Emotions from Video Images of Facial Expressions

Ananya Bhattacharjee<sup>1</sup>, Tanmoy Pias<sup>2</sup>, Mahathir Ahmad<sup>3</sup>, and Ashikur Rahman<sup>4</sup>

Department of CSE, Bangladesh University of Engineering and Technology, Dhaka-1000, Bangladesh  
{ananyab153258<sup>1</sup>, tanmoy.sarkar.pias<sup>2</sup>, sazerterus25<sup>3</sup>, }@gmail.com, ashikur@cse.buet.ac.bd<sup>4</sup>

**Abstract**—This work is meant to provide insights towards state-of-the-art *Application Programming Interfaces* (APIs) commercially available today for recognizing emotions from video footprint of facial expressions. We analyze and compare performance of four such integrable commercial APIs using standard Extended Cohn-Kanade Dataset (CK+) containing over 10,000 images with facial emotions and some randomly collected real life images. We also discuss issues to understand their limitations and adaptation as well as enhancement strategies. For performance analysis, we introduce the performance metric *moving average* to find the primary expression in the displayed emotion of a video sequence. Finally, using the popular *valence-arousal* dimensions, we show how to (lexicographically) order six basic emotions anger, happiness, sadness, surprise, fear and disgust. By providing neutral performance comparison, we hope to fill the existing notable gap between researchers, solution providers and application developers working on emotion based user modeling.

**Keywords**—Facial Expression Recognition; Emotion; API; Performance;

## I. INTRODUCTION

Emotion detection and recognition is a promising research field known as *affective computing* that has attracted many researchers from both academia and industry over the past decade. It allows tracking user's emotion at any given time and enables a machine to "read" the emotions of a human such as anger, sadness, fear, joy, disgust, and surprise, to name a few. By nature, human being conveys the message to others about their mood or emotion in mind by using facial expression, gesture and posture, or changing voice or other physiological parameters such as blood pressure, palpitation, breathing pattern and so on. Therefore, scientifically emotion detection and recognition requires an interesting blend of psychology and technology. The technology part needs the use of Internet of Things (IoT), wearable technology, and more often intelligent use of smartphones to detect the physical or physiological parameter changes in human body in response to a certain emotion in mind.

According to Gupta and Garg [10], multi-factors are involved in communicating a message from human to human to express emotional states of mind. The verbal part of the message contributes only 7% of its meaning as a whole; the vocal part contributes 38% while *facial* movement and the expression contributes 55% of the effect of that message. Thus, one can conclude that the *facial part* contributes the most in human communication about their mood/emotion.

A number of applications and API-accessible software exists online that parallels the human ability to detect and recognize emotions from the *facial expression*. These algorithm driven APIs combine advanced image processing techniques with sophisticated machine-learning algorithms that use facial detection and emotive analysis to interpret mood from photos, and videos of human faces. Some of the key players from industry in this domain are Affectiva, Beyond Verbal, Noldus Information Technology, Sentiance, Sightcorp, Realeyes, CrowdEmotion, Kairos AR, Inc., nViso SA., and SkyBiometry. Needless to say that the visual emotion detection market is expanding tremendously. A recent forecast shows the emotion detection and recognition market size will grow from USD 6.72 Billion in 2016 to USD 36.07 Billion by 2021 [1].

Despite a number of facial emotion recognition systems available online, the problem that many application developers face is lack of rigorous performance analysis and comparison of these APIs made by the research community. This situation motivated us to fill this notable gap by writing this paper and providing a fair comparison of four facial emotion recognition systems namely Sightcorp, Kairos, SkyBiometry, and Face API by Microsoft Azure. In another work, Bernin et al. [6] provide performance comparison of four other facial recognition APIs namely Affectiva, InSight, CERT, and Emotient. Thus, our work can be seen as a complementary research to their work. The major contributions of the paper are:

- (a) We provide neutral performance comparison of four emotion recognition APIs namely Sightcorp, Kairos AR, Inc., SkyBiometry, and Face API by Microsoft Azure on CK+ database [16]. In particular, we demonstrate confusion matrix of each of the APIs on 583 sequences of video images taken from CK+ repository.
- (b) Using a number of random real-life images collected from Internet we show how the performance of the four APIs degrades with real-life images.
- (c) We introduce a new metric *moving average* to combine the emotion scores of frame sequences generated from video footprints to deduce the displayed emotion of the entire video sequence.
- (d) Finally, using the popular *valence-arousal* dimensions we show how to (lexicographically) order six basic emotions anger, happiness, sadness, surprise, fear and disgust.

The rest of the paper is organized as follows. Section II

describes application scenarios of emotion recognition. Section III describes the challenges faced by emotion recognition systems. In Section IV we present an overview of the general methodology of most of the FER systems. Section V provides a generic description of the selected APIs which are used for analysis and comparison. In Section VI we present the findings and insights to the performance of all four APIs. Finally Section VII concludes the paper with pointers for future work.

## II. APPLICATION OF FACIAL EXPRESSION RECOGNITION

Facial expression recognition (FER) system has many applications in advertisements, health-care, education, wearable devices, and more. In this section we summarize some of its possible application domains many of which other researchers might already have pointed out.

**Education sector.** The automatic emotion detection can help to conduct better learning in education [22]. For example, while providing lectures in a class room full of students, if the teacher can see the emotional state of the students in real time, s/he can modify the lecture instantly as needed. When the students are getting bored, s/he can tell an interesting story to make them attentive. Again when the students are interested or attentive, the teacher can increase the content flow.

**Understanding kids' behavior.** Kids sometimes can not express their feelings and parents might have difficulties in taking appropriate measures. If the parents possess a system which can recognize the emotional state of the child, he can take proper actions to make his child happy again.

**Lie detection.** Emotion recognition can be used for lie detection [20] at the time of investigation or in other similar situations. British airport authorities are testing one such system based on Facial Action Coding System (FACS) [7].

**Better human-machine interaction.** FER can help in building better and richer interactions between computing applications and the users by including application context and emotional response of the user. For instance, emotion recognition from a user's face could be used for web page usability testing. When a potential customer is visiting a e-commerce site, customers' first five to ten seconds' expression could be captured to understand his/her impression on the website.

**Capturing viewers' response to new TV program.** Many media companies use emotion recognition software to test audience's reaction whenever they want to launch a new program. For example, CBS, an American English language commercial broadcast television, uses emotion analysis software at its Las Vegas laboratory for such purpose.

**Testing for video games.** Kolakowska et al. [15] has described some scenarios where emotion recognition can be applied in testing video games. An interesting application could be to observe how video game players react to external signals on different levels of immersion and understand which emotions are experienced at what points in the game. Often the video game players become so much immersed in virtual reality that they tend to ignore the real world. So it is important to notice when they stop responding to external stimuli. This observation will help to stop the addiction on the video games.

**Music and emotion.** Mikuckas et al. [19] discuss the impact of music on emotional state and vice-versa. A system can be developed to suggest songs based on a person's emotion.

**Banking.** Emotion recognition APIs can be used to gather emotions to make better financial decisions [2]. Financial advisors can use these APIs to gather insight into the minds of their clients, whom they may have never met.

**Healthcare.** Facial expressions can indicate mental health disorders including depression, anxiety and trauma.

## III. CHALLENGES IN FACIAL EXPRESSION RECOGNITION

Although human beings are empowered to detect and interpret facial expressions in the blink of an eye, recognizing facial expressions by a machine is a challenging task. According to [18], obtaining task-representative data, issues around obtaining ground truth, dealing with occlusions, and modeling dynamics are among the major challenges of recognizing emotions. Below we summarize some of those challenges.

**Use of controlled laboratory environment.** Most of the research works providing Facial Emotion Recognition (FER), use controlled environment inside laboratory while collecting data for training purposes. These lab conditions include controlled illumination conditions. The subjects direct their face towards the camera and show their obvious expressions. The real world condition may not have these controlled facilities. There may be occlusions to hide parts of a face and the illuminations may wildly vary. It is often difficult to create real-world like scenarios in a laboratory. Naturally research works conducted in idealistic environment often exclude the problems faced in real-world situations.

**Dependency on neutral expression.** Many works on FER assume that a frame displaying the subject's *neutral face* is available in the sequence. For example, [5] used an averaged neutral face to recognize expressions. These methods would fail if user's neutral face is unavailable or unknown.

**Facial expressions may not reflect actual emotions.** In the literature the term *facial expression* and *emotion* are used interchangeably. However, on many occasions facial expressions may possess different interpretation based upon the context behind the scene. For example, usually a person may cry in sadness. There are instances when a person cries in happiness or in fear. Thus, inferring emotion directly from unusual facial expressions could be erroneous as facial landmarks may convey wrong message.

**No commonly agreed set of features.** There is no universally accepted set of features that works best on FER. Researchers often base their work on arbitrarily set features.

**Dependency on ethnic group.** Some studies [9], [11], [14] show that basic emotions are not Universal. Western Caucasians use distinct sets of facial muscles to express six basic emotions, but for the East Asians the distinctions are less [11]. Consequently, the facial features need to be considered differently for judging emotions across various cultures.

**Label subjectivity.** While annotator labels an image with an emotion, consistency becomes a concern. Manual annotation requires very high inter-rater reliability. As often there is

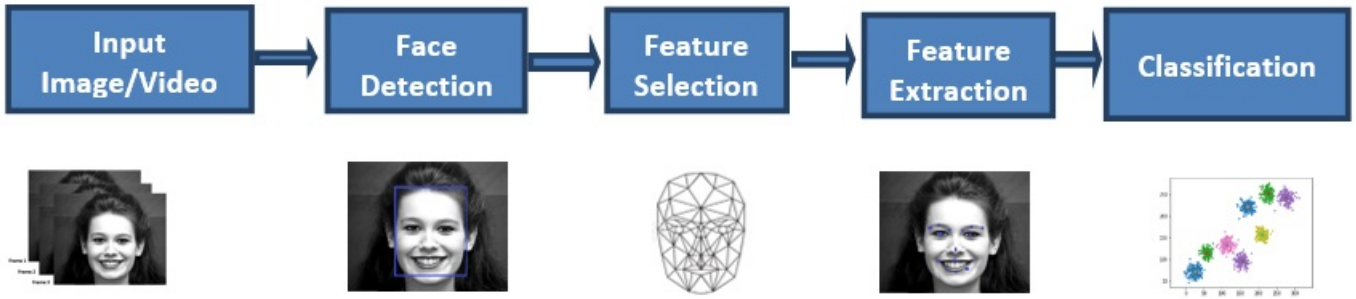


Figure 1: Work-flow of generic facial expression recognition algorithms

choice among the universal six emotions, annotator does not have the choice of free labeling. This is quite a laborious task and also error-prone, if the annotators are not well-trained.

**Need to consider multi-factors in emotion recognition.** Currently different perspectives are separately analyzed to detect an emotion, although in real life these perspectives are analyzed together. Only facial expression does not carry enough information to recognize a particular emotion, it needs to be integrated with head pose variations along with audio and video data.

#### IV. FACIAL EXPRESSION RECOGNITION ALGORITHMS

A number of research works has been conducted for emotion recognition from facial expression. In this section we provide a review of some related works. The generic work-flow of facial expression recognition is shown in Fig. 1. The entire process begins by feeding a video sequence or still image in the system. Then some pre-processing like extracting frames from the video or down-scaling the images etc. are needed. The processed image is then used for face detection. There is a bunch of algorithms for face detection in a picture, among them one or some combination of the algorithms are used. Once the face is detected features are extracted. A face contains different sets of features but most of the time only one type of feature set is needed. For example facial landmarks or action units (AU) could be the feature set. After selecting or defining a feature set, features are extracted from the face and feature vectors are constructed. Then a classification algorithm is trained with those feature vectors. Once trained, the same model is used for classification purposes of the new video image. Sometimes, before invoking the classification algorithm an optimization algorithm is run to reduce the size of the feature set. This is the most basic work flow which can be found in many research works ([4], [12], [17], [20]).

#### V. SELECTED APIS FOR ANALYSIS

We have chosen four state-of-the-art FER algorithms based on their general availability and practicality for performance analysis. All four are capable of both static and realtime processing and available in commercially accessible systems. Below we describe selected FER algorithms in detail.

##### A. Sightcorp

We have used the paper [4] as a reference for unearthing the working principle of Sightcorp API. Similar to other emotion

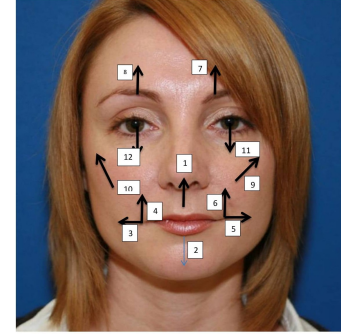


Figure 2: Facial motion units used in Sightcorp

recognition software, the API provided by sightcorp takes an image as input, detects human faces on the image and extracts the features referred to as motion units (MU), which are very similar to action units. According to [4], there are 12 motion units as shown in Fig. 2. They are:

- 1) vertical movement of the center of upper lip
- 2) vertical movement of the center of lower lip
- 3) horizontal movement of left mouth corner
- 4) vertical movement of left mouth corner
- 5) horizontal movement of right mouth corner
- 6) vertical movement of right mouth corner
- 7) vertical movement of right brow
- 8) vertical movement of left brow
- 9) lifting of right cheek
- 10) lifting of left cheek
- 11) blinking of right eye
- 12) blinking of left eye

Once the face is detected, a wireframe model is constructed and fitted in the face. Head motions and local deformations of facial features such as eyes, eyebrows, nose, lips, etc. can then be easily tracked. These local deformations are then expressed as magnitudes of the motion units. We can think of the recovered deformations as a 12-dimensional vector, where the unit vectors are the motion units. This is the feature vector which is then fed into a classifier. The classifier then outputs the confidence value of each emotion.

##### B. Face API By Microsoft Azure

The Face API of Microsoft cognitive services detects the emotions within faces in an image. Its main algorithm is based on a method which was submitted for the Emotion Recognition

Table I: Technical details about APIs

API	Platforms	Labeled Expressions	Output
Sightcorp	Windows, OS X, Linux, iOS and Android	Anger, Disgust, Fear, Happiness, Sadness, Surprise	Probability(0 - 100)
Face	Windows, Linux, Android and iOS	Anger, Disgust, Fear, Happiness, Sadness, Surprise, Contempt, Neutral	Probability(0.0 - 1.0)
SkyBiometry	Windows, Linux, Mac OS X, iOS and Android	Anger, Disgust, Fear, Happiness, Sadness, Surprise, Neutral	Probability(0 - 100)
Kairos	Windows, Linux, Android, iOS, OSX, and Red Hat	Anger, Disgust, Fear, Happiness, Sadness, Surprise	Probability(0 - 100)

in the Wild Challenge (EmotiW) 2015. The method classifies a set of static images into seven basic emotions. This method contains a face detection module based on the ensemble of three face detectors, followed by a classification module based on multiple deep convolutional neural networks (CNN). Below is the description of each module.

1) *Face Detectors*: Three face detectors are used in the detection module. They are:

- (a) **Joint Cascade Detection and Alignment Detector (JDA)**: It is able to return detected faces with high alignment and accuracy and detection precision. But, for profile faces, its accuracy leaves much to be desired. So, this detector is used first in the detection module.
- (b) **Deep-CNN-Based Detector (DCNN)**: Unlike the JDA, it shows excellent performance for non-frontal and even profile faces. Both combined, returns the largest face in a frame where it detects multiple faces.
- (c) **Mixture Of Trees (MoT)**: It is used last in the detection module in case the first two fails to detect faces. Even though it gives accurate face alignment results under many different challenging conditions, the results still contain a lot of missing or false positive faces. That is why, it is used in combination with the previous two for better results.

2) *CNN Module*: The network in CNN module contains five convolutional layers, three stochastic pooling layers and three fully connected layers. Here, stochastic pooling layers are used instead of max pooling layers since stochastic pooling introduces randomness by randomly sampling a response. It reduces the risk of network overfitting. The CNN model is trained on the combined FER dataset formed by training, validation and test set. The network is then fine-tuned on SFEW training set using perturbation and voting strategies. Here, perturbation means that we randomly perturb the input faces with additional transforms.

### C. SkyBiometry

SkyBiometry introduces a free, cloud-based face detection and recognition API. The company was officially launched as a spin-off of Neurotechnology. Neurotechnology is contributing to object recognition and computer vision technologies along with high-precision biometric identification for more than 25 years. They provide cloud-based biometric software as a service (SaaS).

SkyBiometry uses VeriLook algorithm [3], which in turn uses robust digital image processing algorithms based on deep neural networks. However, they do not disclose the algorithm. Advanced face localization, enrollment and matching can be

implemented using this algorithm. SkyBiometry detects 68 points including mouth, nose, eyes and other facial features. It can also extract the points as a set of their coordinates during face template extraction. Each of these points is given a fixed sequence number. For example, number 31 will always correspond to a nose tip. Again, VeriLook can detect certain facial attributes including smile, open-mouth, closed-eyes, glasses, dark-glasses, beard and mustache. This algorithm is configured to detect emotion from a human face. Six basic emotions- anger, disgust, fear, happiness, sadness and surprise are analyzed. The algorithm returns a confidence value for each of the emotions. Emotion with the larger confidence value is the emotion displayed in the face.

### D. Kairos

Kairos, an artificial intelligence company, focuses mainly on face recognition. However, they also offer the feature of emotion and demographic analysis. According to initial research conducted by Kairos and IMRSV (acquired by Kairos in 2015), Ekman's universal emotions [8] do not consistently have distinct facial expressions. For example, it is hard to distinguish anger and disgust. Kairos found these traditionally accepted facial expression of universal emotions to be generic and exaggerated. They believe those analytics to be unreliable, as these expressions are rarely seen in real world environment.

Kairos's face detection algorithms are based on a learned face pattern. An arrangement of pixels that resembles the usual pattern of a human face is detected. Their Anonymous Video Analytics (AVA) Technology uses patterns such as pixel density around the eyes, nose, and mouth. After detection of faces, redundant objects are ignored.

Although Kairos wrote the ultimate face recognition white paper, they are yet to publish any white paper on emotion analysis. Kairos provides five features in addition to emotion analysis—Attention Measurement, Emotion Detection, Facial Expression Detection, Gender Detection, and Age Detection. For emotion detection, the API looks for faces in images and videos and analyzes the facial features and expressions using proprietary face analysis algorithms. Finally, it returns values for the six universal emotions of the faces found. It also returns values for age, gender, and other useful meta data.

## VI. EXPERIMENTAL ANALYSIS

In this section we provide performance of four APIs.

### A. Utilized database

For analyzing performance of APIs, we use 583 videos from the Extended Cohn-Kanade Dataset (CK+) [16]. This



Figure 3: An Example Frame Sequence of Happiness

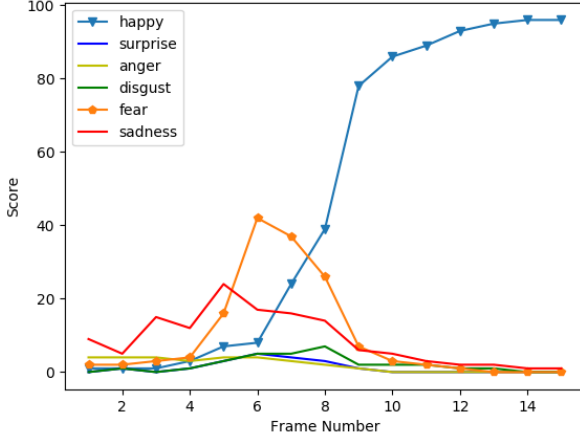


Figure 4: Emotion scores in video sequence of Fig. 3 using Sightcorp API

dataset provides sequences of video images, each of which starts with a neutral expression and ends in a peak expression. An example sequence of happiness is shown in Fig. 3.

We have conducted a black box testing [21], as these APIs do not publish any white paper on their algorithms. We use these APIs' developer libraries and apply those on the dataset images. These APIs provided emotion scores along with other facial features, gender, ethnicity etc. We have extracted the emotion scores from those results returned by the APIs.

### B. Classification and decision metric

In a video, the expression changes over time and consequently the emotion score of the frames also changes from one to another. As the task is to identify the primary expression over a period of time in a video footprint, a mechanism is needed to combine these emotion scores on individual frames. To further clarify, let's see the emotion scores in Fig. 4 produced by Sightcorp on the frame sequence of Fig. 3. Although the overall expression should be categorized as *happy*, the happy scores of initial frames are very low (less than 5%). Thus, a metric formulaion is needed to combine the emotion scores of each frame sequence. In this context, we introduce the following metric.

**Moving Average:** In order to find the score of an emotion, we calculate exponential moving average from emotion scores of the individual frame sequences of a video according to

Table II: Moving average of Emotion scores in video sequence of Fig. 3 using Sightcorp API

Anger	Disgust	Fear	Happiness	Sadness	Surprise
0.03	0.34	0.45	94.63	1.46	0.03

Equation 1, where  $n$  represents the frame number within the video sequence. Each emotion's moving average value is independent and calculated separately. While finding moving average of an emotion, we set  $\alpha = 0.5$  to equally emphasize the emotion score of the current frame and the average score of all previous frames.

$$\mathcal{M}(x, n) = \begin{cases} x_1 & \text{if } n = 1 \\ \alpha x_n + (1 - \alpha)\mathcal{M}(x, n - 1) & \text{else if } n > 1 \end{cases} \quad (1)$$

Sightcorp, SkyBiometry and Kairos treat emotions independently, and the scores of the emotions do not rely on each other. On the other hand, the Face API treats emotions to be dependent to each other, consequently the scores for all emotions must sum to 1.

Let us explain the scoring process using an example. When we use Sightcorp to detect emotions in the image sequences of Fig. 3, it generates the results as shown in Fig. 4. For simplicity, we number these images from 1 to 15. When we combine the emotion scores using moving average we get the results shown in Table II. Notably, moving average has the highest values for happiness. So, we conclude Sightcorp correctly classifies this sequence.

### C. Results

We present confusion matrices for all four APIs using moving average as a metric in Tables III, IV, V and VI. Sightcorp, SkyBiometry and Kairos can detect basic six emotions, but Face API has additional capability of detecting contempt. We included all seven emotions in our confusion matrices as we wanted to observe what emotions Sightcorp, SkyBiometry and Kairos label in case of actual contempts. In these matrices there is an extra label X, which indicates the API could not classify the images in a video sequence. It may happen when none of the emotions have a value greater than zero or the API could not even detect the face in the image.

### D. Discussion

Table VII summarizes the performance accuracy of detecting correct emotion labels of APIs. As we can see, although Sightcorp does not have the option to detect contempt, it has

Table III: Confusion matrix for video sequences using Sightcorp API

	Predicted Emotion (Threshold Metric: Moving Average)								
Actual Emotion		X	Anger	Contempt	Disgust	Fear	Happiness	Sadness	Surprise
	Anger	0	68 (97.14%)	0	0	1 (1.43%)	0	1 (1.43%)	0
	Contempt	0	7 (29.17%)	0	2 (8.33%)	1 (4.17%)	4 (16.67%)	6 (25%)	4 (16.67%)
	Disgust	0	2 (2.7%)	0	71 (95.95%)	0	1 (1.35%)	0	0
	Fear	0	0	0	0	63 (91.30%)	5 (7.25%)	1 (1.45%)	0
	Happiness	0	0	0	0	0	112 (99.12%)	0	1 (0.88%)
	Sadness	0	3 (2.73%)	0	1 (0.91%)	1 (0.91%)	1 (0.91%)	104 (94.55%)	0
	Surprise	0	0	0	0	1 (0.81%)	2 (1.63%)	1 (0.81%)	119 (96.75%)

Table IV: Confusion matrix for video sequences using Face API

	Predicted Emotion (Threshold Metric: Moving average)								
	X	Anger	Contempt	Disgust	Fear	Happiness	Sadness	Surprise	
Actual Emotion	Anger	1 (1.43%)	37 (52.86%)	11 (15.71%)	0	0	4 (5.71%)	14 (20%)	3 (4.29%)
	Contempt	0	1 (4.17%)	9 (37.5%)	0	0	10 (41.67%)	3	1 (4.17%)
	Disgust	1 (1.35%)	17 (22.97%)	2 (2.7%)	49 (66.22%)	0	2 (2.7%)	3 (4.05%)	0
	Fear	0	6 (8.7%)	0	5 (7.25%)	11 (15.94%)	21 (30.43%)	17 (24.64%)	9 (13.04%)
	Happiness	0	1 (0.88%)	0	1 (0.88%)	0	110 (97.35%)	1 (0.88%)	0
	Sadness	5 (4.55%)	4 (3.64%)	3 (2.73%)	1 (0.91%)	0	8 (7.27%)	78 (70.91%)	11 (10%)
	Surprise	1 (0.81%)	1 (0.81%)	2 (1.62%)	0	0	14 (11.38%)	4 (3.25%)	101 (82.11%)

Table V: Confusion matrix for video sequences using SkyBiometry API

	Predicted Emotion (Threshold Metric: Moving average)								
		X	Anger	Contempt	Disgust	Fear	Happiness	Sadness	Surprise
Actual Emotion	Anger	0	52 (74.29%)	0	6 (8.57%)	0	0	6 (8.57%)	6 (8.57%)
	Contempt	0	4 (16.67%)	0	3 (12.5%)	0	6 (25%)	0	11 (45.83%)
	Disgust	0	9 (12.16%)	0	64 (86.49%)	0	0	0	1 (1.35%)
	Fear	0	4 (5.8%)	0	5 (7.25%)	34 (49.28%)	11 (15.94%)	6 (8.7%)	9 (13.04%)
	Happiness	0	2 (1.63%)	0	3 (2.44%)	1 (0.81%)	103 (83.74%)	0	4 (3.23%)
	Sadness	0	37 (33.64%)	0	1 (0.91%)	12 (10.91%)	2 (1.82%)	33 (30%)	25 (22.73%)
	Surprise	0	2 (1.62%)	0	4 (3.24%)	7 (29.17%)	1 (0.81%)	1 (0.81%)	108 (87.8%)

Table VI: Confusion matrix for video sequences using Kairos API

	Predicted Emotion (Threshold Metric: Moving average)								
Actual Emotion		X	Anger	Contempt	Disgust	Fear	Happiness	Sadness	Surprise
	Anger	32 (45.71%)	24 (34.29%)	0	0	2 (2.86%)	0	8 (11.43%)	4 (5.71%)
	Contempt	14 (58.33%)	1 (4.17%)	0	0	3 (12.5%)	2 (8.33%)	4 (16.67%)	0
	Disgust	9 (12.16%)	23 (31.08%)	0	41 (55.41%)	0	0	1	0
	Fear	23 (33.33%)	2 (2.9%)	0	0	29 (42.03%)	6 (8.7%)	0	9 (13.04%)
	Happiness	16 (14.16%)	1 (0.88%)	0	2 (1.77%)	3 (2.65%)	89 (78.76%)	0	2 (1.77%)
	Sadness	49 (44.55%)	2 (1.82%)	0	0	12 (10.91%)	1 (0.91%)	38 (34.55%)	8 (7.27%)
	Surprise	26 (21.14%)	0	0	1 (0.81%)	6 (4.88%)	0	0	90 (73.17%)

Table VII: Accuracy of APIs using moving average

Sightcorp	Face	SkyBiometry	Kairos
92.11%	67.75%	67.58%	53.34%

the highest accuracy. If we do not consider the sequences labeled contempt, the accuracy rises up to 96.06%.

Face API shows poor performance in detecting fear (15.94%). It has marked more than 50% sequences of fear to be either happy or sad.

SkyBiometry shows more than 70% accuracy rate in detecting anger, disgust, happiness and surprise. It shows better performance than Face API in detecting fear. However, it shows only 30% accuracy in case of sadness.

Kairos could not classify 169 sequences. It shows poor performance in detecting faces. But it could detect around 75% sequences correctly, whenever it could classify.

All four APIs have shown good performance in detecting happiness and surprise (73% accuracy at least). For Sightcorp,

the accuracy on happiness videos is more than 96%.

### E. Performance analysis in valence-arousal space

According to the dimensional approach [13], emotions are related to one another in a systematic and orderly manner. Using this approach an emotion can be modeled as a point in a two-dimensional space defined by:-(i) *valence*, and (ii) *arousal*. The valence dimension refers to how positive or negative the emotion is (ranging from unpleasant to pleasant feelings) and the arousal dimension refers to how excited or apathetic the emotion is (ranging from sleepiness to total excitement). Using valence and arousal, all emotions can be plotted at various positions on a two-dimensional plane as shown in Fig. 5. It is easy to see that the emotional space consists of four quadrants: low arousal positive, high arousal positive, low arousal negative, and high arousal negative. After plotting all emotions, we can order them based on their vector distance from each other. For example, if we consider two vectors, one for “surprise” and the other for “joy” as shown in



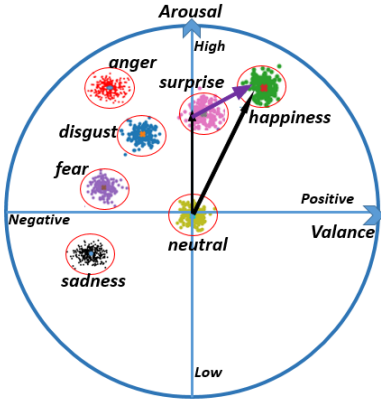


Figure 5: Emotions in valence-arousal space [13]

Table VIII: Emotions’ adjacency list in valence-arousal space

Emotion	Adjacent Emotion(s)
Anger	Disgust, Surprise
Disgust	Fear, Surprise, Anger
Fear	Sadness, Disgust
Happiness	Surprise
Sadness	Fear
Surprise	Happiness, Disgust

Fig. 5, then based on their vector distance we can conclude that “surprise” is adjacent emotion to “joy” in this valence-arousal space. On the other hand, the vector distance from “joy” to “sadness” is very large. Thus, if an API classifies an image under “joy” category to “sadness” is much worse than another API classifying “joy” to “surprise”. Using this approach we can easily find adjacent emotion(s) of all emotion. Table VIII shows the complete adjacency list based on this valence-arousal concept. Contempt is not included in this model.

After finding the adjacency lists, we can measure the accuracy of an API based on the following four metrics:

- Correctly classified:** Percentage of test videos that are classified to correct emotions.
- Adjacent classified:** Percentage of test videos that are classified to an emotion which is adjacent to correct emotions (adjacency is defined in Table VIII).
- Misclassified:** Percentage of test videos that are classified to an emotion which is neither correct nor adjacent to the correct emotion.
- Unclassified:** Percentage of test videos that could not be recognized or classified by an API.

With the above four performance metrics we show the accuracy of APIs based on the valence-arousal model in Fig 6. The same 583 videos from the Extended Cohn-Kanade Dataset (CK+) [16] have been used. The moving average is used to combine the scores of each video frame. The correct classification rates of Sightcorp is highest 96%. The FACE API and the SkyBiometry show more or less similar correct classification rate (69%, and 70% respectively). Kairos shows the worst performance with 50% correct classification rates. The adjacent classified rates of Sightcorp, FACE API, Sky-Biometry, and Kairos are 1%, 10%, 10% and 8% respectively. FACE API and SkyBiometry have the highest misclassification

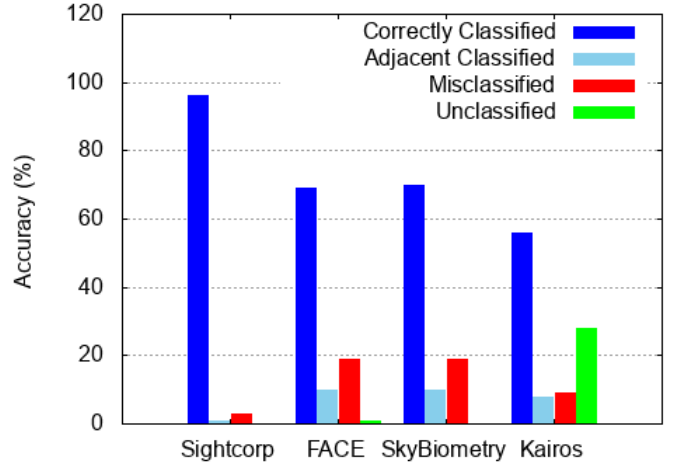


Figure 6: Accuracy of APIs using adjacent emotion model

rate which is about 10%. Misclassification rate of Sightcorp and Kairos is 3% and 8% respectively. Kairos shows a very high unclassification rates of about 28%. Unclassification rate of other three APIs is negligible.

#### F. Performance on Real Life Scenarios

As all APIs were trained on images taken under controlled lab environment, we investigate how they behave on real-life expressions. For this purpose, we collected 14 images of some of the real life expressions from the Internet and labeled them with our best judgment. We ran the API’s with these images and recorded their respective scores. We use maximum rule (alternatively known as “Winner takes it all”) to classify an image according to the score of a specific API. We can see the results in Table IX.

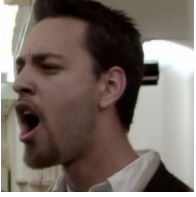
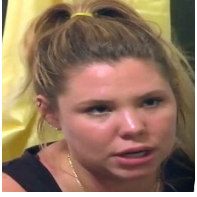

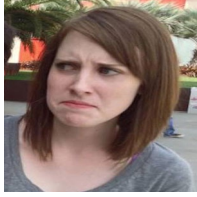


Surprisingly, none of the APIs could reach even 50% accuracy in detecting emotions. SkyBiometry performs best among four and could detect emotions in six images correctly. On the other hand Face API, Sightcorp and Kairos could correctly detect three, two and one image(s) respectively. None of the APIs could detect faces in Image 8 and Image 11 of the Table IX. In these images, the head orientations are different from those of controlled lab conditions. As mentioned in Section III, performances of these APIs degrade significantly in real life conditions which are different from lab environments.

## VII. CONCLUSION


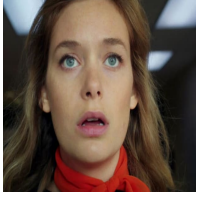
Facial expression recognition is essential for many applications ranging from the fields of entertainment to basic needs. Although many emotion recognition systems have been developed, they still have some rooms for improvement. We discuss the challenges faced by these systems and then analyze four state-of-the-art APIs’ performance. We propose a new metric moving average to boost the performance of the APIs. We also provide a (pseudo) ordering of emotions based on valence-arousal dimensions to better discuss their performance. Finally, we show APIs still suffer in recognizing emotions in real life environments. We hope our work will provide an insight into



Table IX: Performance on Real Life Scenarios

	1	2	3	4	5	6
						
Actual	Anger	Anger	Anger	Disgust	Disgust	Fear
Sightcorp	Face not detected	Surprise	Face not detected	Disgust	Face not detected	Disgust
Face API	Neutral	Neutral	Face not detected	Sadness	Sadness	Neutral
SkyBiometry	Anger	Anger	Face not detected	Anger	Face not detected	Anger
Kairos	Face not detected	0 for all emotions	0 for all emotions	Anger	Sadness	Disgust

	7	8	9	10	11	12
						
Actual	Happiness	Happiness	Happiness	Sadness	Sadness	Sadness
Sightcorp	Fear	Face not detected	Disgust	Sadness	Face not detected	Happiness
Face API	Neutral	Face not detected	Sadness	Sadness	Face not detected	Sadness
SkyBiometry	Surprise	Face not detected	Surprise	Sadness	Face not detected	Surprise
Kairos	Fear	Face not detected	Sadness	Face not detected	Face not detected	Face not detected

	13	14
		
Actual	Surprise	Surprise
Sightcorp	Anger	Sadness
Face API	Neutral	Surprise
SkyBiometry	Surprise	Surprise
Kairos	Surprise	Face not detected

the future works needed to be done in this arena to make these systems practical and perform well in real life scenarios.

## REFERENCES

- [1] <https://www.marketsandmarkets.com/ResearchInsight/emotion-detection-recognition-market.asp>. Accessed: 2018-06-20.
- [2] nviso helping financial advisors understand their clients' true financial needs with emotional intelligence. <https://www.ibm.com/case-studies/t338899k54153u05>. Accessed: 2018-06-20.
- [3] Verilook sdk: Face identification for stand-alone or web applications. <http://www.neurotechnology.com/verilook.html>. Accessed: 2018-06-20.
- [4] AZCARATE, A., HAGELOH, F., VAN DE SANDE, K., AND VALENTI, R. Automatic facial emotion recognition.
- [5] BAZZO, J., AND LAMAR, M. Recognizing facial actions using gabor wavelets with neutral face average difference., 01 2004.
- [6] BERNIN, A., MÜLLER, L., GHOSE, S., VON LUCK, K., GRECOS, C., WANG, Q., AND VOGT, F. Towards more robust automatic facial expression recognition in smart environments. *Proc. of the 10th Int. Conf. on Perv. Tech. Related to Assistive Environments* (2017), 37–44.
- [7] COUNCIL, N. R., ET AL. The polygraph and lie detection. committee to review the scientific evidence on the polygraph. division of behavioral and social sciences and education. Washington, DC: *The National Academic Press*. Retrieved 7, 7 (2003), 09.
- [8] EKMAN, P., FRIESEN, W. V., AND ELLSWORTH, P. *Emotion in the Human Face: Guide-lines for Research and an Integration of Findings*. Pergamon, 1972.
- [9] ELFFENBEIN, H., AND AMBADY, N. Universals and cultural differences in recognizing emotions. *Current Directions in Psychological Science* 12 (10 2003), 159–164.
- [10] GUPTA, A., AND GARG, M. A human emotion recognition system using supervised self-organising maps, 03 2014.
- [11] JACK, R. E., GARROD, O. G. B., YU, H., CALDARA, R., AND SCHYNS, P. G. Facial expressions of emotion are not culturally universal. *Proceedings of the National Academy of Sciences of the United States of America* 109, 19 (2012), 7241–7244.
- [12] JAIN, S., HU, C., AND AGGARWAL, J. K. Facial expression recognition with temporal modeling of shapes. *IEEE*, pp. 1642–1649.
- [13] JIN, X., AND WANG, Z. An emotion space model for recognition of emotions in spoken chinese. Springer, pp. 397–402.
- [14] KILBRIDE, J. E., AND YARCZOWER, M. Ethnic bias in the recognition of facial expressions. *Journal of Nonverbal Behavior* 8, 1 (1983), 27–41.
- [15] KOŁAKOWSKA, A., LANDOWSKA, A., SZWOCH, M., SZWOCH, W., AND WRÓBEL, M. Emotion recognition and its applications. *Advances in Intelligent Systems and Computing* 300 (07 2014), 51–62.
- [16] LUCEY, P., COHN, J. F., KANADE, T., SARAGIH, J., AMBADAR, Z., AND MATTHEWS, I. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression.
- [17] MAJUMDER, A., BEHERA, L., AND SUBRAMANIAN, V. K. Emotion recognition from geometric facial features using self-organizing map. *Pattern Recognition* 47, 3 (2014), 1282–1293.
- [18] MARTINEZ, B., AND VALSTAR, M. Advances, challenges, and opportunities in automatic facial expression recognition, 04 2016.
- [19] MIKUCKAS, A., MIKUCKIENE, I., VENČKAUSKAS, A., KAZANAVICIUS, E., LUKAS, R., AND PLAUSKA, I. Emotion recognition in human computer interaction systems. *Elektronika ir Elektrotechnika* 20 (12 2014), 51–56.
- [20] OWAYJAN, M., KASHOUR, A., N HADDAD, A., FADEL, M., AND G SOUKI, A. The design and development of a lie detection system using facial micro-expressions, 12 2012.
- [21] PATTON, R. *Software testing*. Pearson Education India, 2006.
- [22] PICARD, R. W. Affective computing: challenges. *International Journal of Human-Computer Studies* 59, 1-2 (2003), 55–64.