

## SMILE-O-METER: A PILOT PROJECT TO MONITOR A PATIENT'S EMOTIONAL CHANGES THROUGH AN ON-LINE THERAPY SESSION

Marius-Eduard COJOCEA <sup>1\*</sup>

Robert-Costin BERCARU <sup>2</sup>

Costin-Anton BOIANGIU <sup>3</sup>

Mihai BRAN <sup>4</sup>

Ciprian Andrei APRODU <sup>5</sup>

Florin Cristian HUREZEANU <sup>6</sup>

Iuliana Andreea SICARU <sup>7</sup>

### ABSTRACT

*Smile-O-Meter is a pilot project that aims to track and analyze how a patient's emotional state changes throughout an online therapy session using a webcam. The main goal of the project is to see if a patient's mood has improved over time and this is performed by analyzing the frames of the webcam's video stream, with the help of facial recognition and emotion detection algorithms, coupled with a deep learning technique. The current approach aims to detect five different emotions. Results have shown that some emotions have a high detection rate (happiness, surprise), while others tend to be wrongly interpreted (sadness, disgust).*

**KEYWORDS:** *emotion recognition, facial recognition, online therapy, convolutional network*

### 1. INTRODUCTION

Psychologists and counselors are using the technology to make their daily tasks easier, mostly outside the therapy sessions. But because humans use emotions to convey messages [1], psychologists are starting to use the technology during the therapy sessions, the main reason being that patients are not always sincere. Also, psychologists might miss some features and expressions that may appear during the session. They may be recording the sessions, even go

---

<sup>1\*</sup> corresponding author, Engineer, PhD stud., "Politehnica" University of Bucharest, 060042 Bucharest, Romania, marius.cojoccea@cti.pub.ro

<sup>2</sup> Engineer, "Politehnica" University of Bucharest, 060042 Bucharest, Romania, robert.bercaru@cs.pub.ro

<sup>3</sup> Professor PhD Eng. "Politehnica" University of Bucharest, 060042 Bucharest, Romania, costin.boiangiu@cs.pub.ro,

<sup>4</sup> Psychiatrist, PhD Stud., Colțea Hospital, Bucharest, Romania, mihaibrans@gmail.com

<sup>5</sup> Engineer, "Politehnica" University of Bucharest, 060042 Bucharest, Romania, ciprian.aprodu@cs.pub.ro

<sup>6</sup> Engineer, "Politehnica" University of Bucharest, 060042 Bucharest, Romania, florin.hurezeanu@cs.pub.ro

<sup>7</sup> Engineer, "Politehnica" University of Bucharest, 060042 Bucharest, Romania, iuliana.sicaru@cs.pub.ro

through them after to see if they notice something new in a patient’s mood evolution or they can even provide online counseling or telepsychology, “the provision of psychological services using telecommunication technologies” [2]. Telepsychology has been practiced for a while now and comes in different forms: by phone, text messages, emails or webcams [3]. Monitoring a patient through these types of online sessions can be an automated process, by detecting, tracking and monitoring her/his emotions. The presented project is not an attempt at replacing the psychologist but rather is intended as an indication tool that can be used in order to improve the sessions. It is capable of giving punctual information, but it is more useful when providing statistics regarding the patient’s mood and tracking the evolution in time.

The first part of this paper presents some of the previous studies carried in this field. In the next part, the paper covers information about the technologies that were used during the implementation, after which information about the implementation itself is presented. The last part of the paper presents the results that were obtained using the presented application.

## 2. RELATED WORK









AU	Description	Facial muscle	Example image
<b><u>1</u></b>	Inner Brow Raiser	<i>Frontalis, pars medialis</i>	
<b><u>2</u></b>	Outer Brow Raiser	<i>Frontalis, pars lateralis</i>	
<b><u>4</u></b>	Brow Lowerer	<i>Corrugator supercilii, Depressor supercilii</i>	
<b><u>5</u></b>	Upper Lid Raiser	<i>Levator palpebrae superioris</i>	
<b><u>6</u></b>	Cheek Raiser	<i>Orbicularis oculi, pars orbitalis</i>	
<b><u>7</u></b>	Lid Tightener	<i>Orbicularis oculi, pars palpebralis</i>	
<b><u>9</u></b>	Nose Wrinkler	<i>Levator labii superioris alaquae nasi</i>	
<b><u>10</u></b>	Upper Lip Raiser	<i>Levator labii superioris</i>	

Figure 1. Example of Action units from the FACS system (Source [15])

*Affectiva* is a company that has made improvements to the field of emotion detection, their main focus being using this technology for market research. *Affectiva's* emotion detection solution measure 7 emotions: anger, contempt, disgust, fear, joy, sadness and surprise. The way this software works is presented in their official paper [13] which puts emphasis on the methodology that the team used to select the data training sets and perhaps, more importantly, on the “Efficient Non-Linear Kernel Approximation” [14] and “Active Learning Algorithm” [16] that seem to be the key to the software’s accuracy. The specific actions and signals that *Affectiva* searches for in a picture are AU02 (outer eyebrow raise), AU04 (eyebrow lower) and smiles. *Affectiva's* plans for the future include extending the search for more face markers and further improving the already accurate algorithm.

*Imotions* and *Emotient* are two other companies that have united in the purpose of creating a fully-developed solution software for recognizing emotions based on face movements and postures. They thrive to detect the same 7 emotions as the people at *Affectiva*, but also adding another two “advanced” emotions: confusion and frustration. Besides this, the program also extracts an overall feeling of a subject with positive, negative and neutral streams. This result was possible due to using *FACS* (Facial Action Coding System) [14], [15] which is a system for defining various human facial movements, based on facial regions of interest. There are 64 “Action Units”, each with a name, the muscle groups involved in obtaining that facial expression starting from a neutral, relaxed face and a picture showing that facial expression, as it can be seen in Figure 1.

Using this system to detect emotions is not something entirely new. The breakthrough comes from the fact that it is now used to categorize emotions automatically, whereas in the past actual doctors and scientists that were familiar with *FACS* studied images by hand to retrieve emotions.

In the not-so-distant future, *Imotions* and *Emotient* plan to combine the software that they created with Stimuli, Facial Expressions, EEG (electroencephalography), GSR (galvanic skin response) and more in order to achieve the best accuracy possible in.

### **3. USED TECHNOLOGIES**

Earlier methods for detecting facial features were based on extracting manually the features of interest from the image. These features were usually a collection of points mapped on a face and corresponded to facial areas that determine expressions. In general, these areas are mouth, eyes, nose, eyebrows and face contour. Although some of these methods had some success, they were based on heuristics and new mappings and rules had to be implemented for each new expression. Taking this into consideration, deep learning is a better alternative for approaching this task. Firstly, by using deep convolutional networks, the relevant features that have to be extracted are learned and not manually extracted. This means, that the model can extract the features that matter, even though some of these may seem esoteric. Secondly, the same model (maybe with a few changes in hyperparameters) can be used to detect new emotions, however, needing a dataset with labeled pictures of people experiencing the wanted facial emotions.

Convolutional networks are a better choice than fully connected networks for a couple of reasons. Since in images the localization and the neighborhood of every pixel are very relevant, two pixels that are outside each other's vicinity do not contain relevant information regarding each other, so instead of connecting each pixel in the input to each neuron in the first layer, it is better to use convolutional connections. A second reason is that convolutional networks are much faster, thanks to fewer connections and shared weights.

Amongst the most popular open-source deep learning frameworks is *Tensorflow* [4], which will be used for the purpose of this project as well. *Tensorflow* has been developed by *Google Brain Team*, the first version being released in February 2017. The API is available in multiple programming languages, such as Python, C++, Java and Go.

To train the network, labeled images are needed. They will be used by the neural network to learn to differentiate between the desired classes of emotions. After the training is done and the model provides the desired performance, the model can be saved and loaded when the project is run. The training of the model is time-consuming, taking hours, maybe days, depending on the sizes of the dataset and the neural network. However, the trained model can classify images very fast. Depending on the size of the neural network and the available hardware, the response time may vary between a few milliseconds to a few seconds. This is enough for real-time detection of emotions during sessions with a psychologist.

#### 4. CONVOLUTIONAL NEURAL NETWORKS

The main layers in a convolutional neural network are convolutional layers, in which every neuron is connected to a window of neurons in the previous layer, as it can be seen in Figure 2. This allows the network to extract features automatically, the layers acting like a multitude of filters that are applied to the input. Initially, the filters are randomly initialized, but during training they should converge to the desired values.

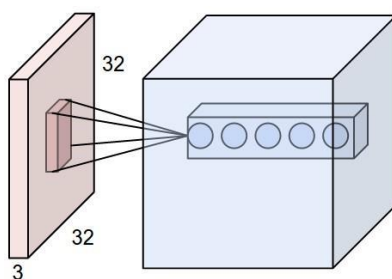


Figure 2. Example of how neurons are connected in a convolutional layer (Source [18])

Usually, convolutional layers induce a certain depth to the information. This means that the depth of the layer is equal with the number of filter types that will be learned, as it can be seen in Figure 3.

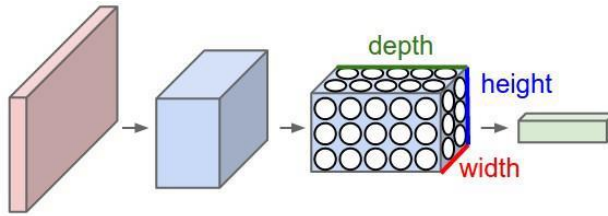


Figure 3. Example of how information changes shape through a convolutional layer (Source [18])

Pooling is usually applied after a convolutional layer in order to reduce information size and to offer more invariability to rotations, translations, and small variation in features. There are multiple ways to apply pooling, but the most common is max-pooling which chooses the biggest value in a 2x2 window of pixels and sends it forward, as it can be seen in Figure 4.

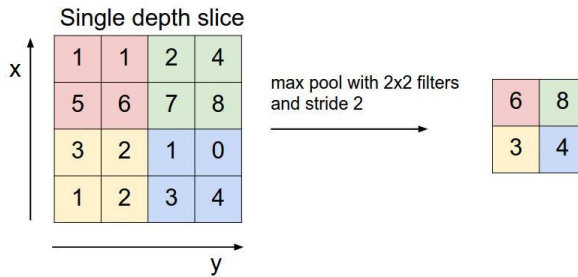


Figure 4. Example of max pooling (Source [18])

The first layers will learn basic features such as edges, regions, when the latter layers will learn complex features such as facial parts, faces, emotions etc. Figure 5 exemplifies this process.

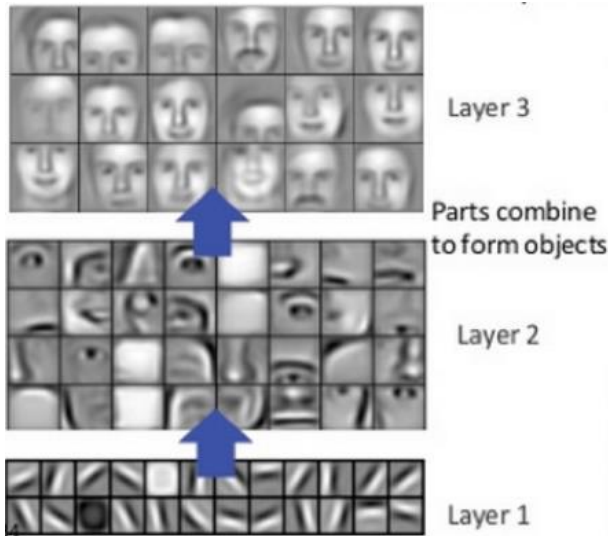


Figure 5. Examples of the features learned by different layers (Source [19])

## **5. IMPLEMENTATION DETAILS**

When the application is first run, the neural network will be initialized with a file which contains the configuration of the trained network. Afterward, the network will only be used for interrogations. The application allows input from a specific folder or from a video stream (webcam). The output can be saved on the hard disk (used for manual labeling) or the application can directly output an emotion classification. In both situations, the user has to choose a sampling period at which the images are extracted. In the situation of having an online stream, the sampling will be done based on time, whereas if the input is from a specific folder (offline), the sampling will be done by establishing a specific frame-rate.

In the current state of the project, the sampling of the video stream from a webcam is being done every  $t$  seconds, by setting up a timer that calls the method responsible with the interrogation of the video stream, thus obtaining each frame. This is an important step because by eliminating redundant data, the load of the network decreases as well. Each frame will be processed and sent over to the neural network to be analyzed for emotions.

As an alternative to sampling a video stream, a file system watchdog was implemented. This will monitor a specific directory for new files and changes to existing files as well. When a new image or video file is created, it waits for it to be fully written on the disk. This is done by inspecting the file size until 2 consecutive results are equal. If the file is an image, then the specific image is loaded and it is being preprocessed in order to be fed to the neural network. Otherwise, if the file is a video, the watchdog extracts frames at each second of the video and creates images with them in the same input folder so they can be processed as a new image would be.

### **5.1. Emotion detection**

The emotion recognition step is done using convolutional networks. These networks “can accurately interpret semantic information available in faces in an automated manner without hand-designing of features descriptors” [5] and they are capable of learning different characteristics from an image, by feeding them with matrices of pixel values for each image, with minimum preprocessing. They also considerably reduce the computational cost [6] due to the fact that they are not fully-connected networks and weight sharing. Usually, convolutional neural networks contain mainly convolutional layers, coupled with pooling. In a convolutional layer, each neuron is connected to a windows of neurons from the previous layer and all neurons share the same convolutional weights. This way, a lot fewer connections are needed compared to fully connected layers. Also, in images, pixels are relevant only to a limited vicinity, not to the whole image, which makes convolutional connections very useful. The pooling helps as well, reducing the size of data while keeping the essence of the information and giving the images invariance to rotation and distortions.

In the early stages, this algorithm was supposed to detect 7 emotions. Due to the lack of relevant dataset “fear” has been removed from the classes. Also, due to the high similarity with every other class, the “neutral” class was also removed. The currently used algorithm detects 5 emotions: disgust, sadness, anger, surprise, and happiness.

## 5.2. Preprocessing

Preprocessing each frame is important to increase the quality and the response speed of the neural network. The preprocessing step involves applying face detection and extraction algorithm on each frame (Figure 6), followed by a resizing of the image to a 128x128 dimension if a face has been detected, and, finally, converting the image into a grayscale one (Figure 7).



Figure 6. Example of an image before the preprocessing step. (Source: [7])

Also, the images are randomly flipped left-right, rotated and blurred. This is done as a way of giving images invariance to various transforms. This also helps with generalizing the information found in the dataset.

The face extraction step is done using trained Haar cascades classifiers [8]. The image is first converted to gray-scale, after which the algorithm marks the coordinates of a detected face (front or profile). The images containing faces will be scaled to a given dimension by using a Lanczos4 interpolation [9].



Figure 7. Example of a cropped, gray-scale photo obtained from the original photo.

The face detection and extraction steps are necessary because of several problems encountered when no face extraction algorithms were applied and the results were incorrect because of the background of the image being mostly constant in the training dataset, while in reality, the background can vary a lot.

Another preprocessing step involves reformatting the data to the correct input format for the neural network: converting integers' values [0, 255] to floats [0.0, 1.0] and modifying the matrix which contains the 2D data to a 3D one by wrapping each value in an array with only one element. This can be achieved easily using numpy, by creating a new axis to a numpy array. The network uses tensors in order to move data inside. Although an image maps visually to a 2D tensor, it is actually a 3D tensor, where the third dimension is the number of color channels. Only after this can the data be used to interrogate the network.

### 5.3. Network response

After the analysis of the input data, the network will provide a dictionary, which consists of each possible emotion with its associated probability. These probabilities will sum up to 1.0. The emotion with the highest probability will be considered as the detected emotion.

```
{
  "time stamp": "2018-01-29 21:08:18",
  "emotions": {
    "anger": "0.0000",
    "sadness": "0.0000",
    "happiness": "1.0000",
    "surprise": "0.0000",
    "disgust": "0.0000"
  },
  "id": "pixels-photo-713312"
}
```

Figure 8. Example of network response.

For the alternative implementation that is using a watchdog, the results are being written to disk as a JSON file, as shown in Figure 8. The file name will be the image id, which is the original image file name. The file will contain the id of an image, which may be needed for future functionalities, the timestamp at which the file was written and the probability for each emotion.

## 6. RESULTS

For testing purposes, we have used Cohn-Kanade dataset [10], [11]. The complete dataset consists of 593 sequences gathered from 123 subjects. All sequences begin from the neutral face and end on the peak expression. There are only 327 from the 593 sequences that display emotion sequences, with 8 different emotions available: neutral, anger, contempt, disgust, fear, happiness, sadness and surprise.



### General Detection Rate

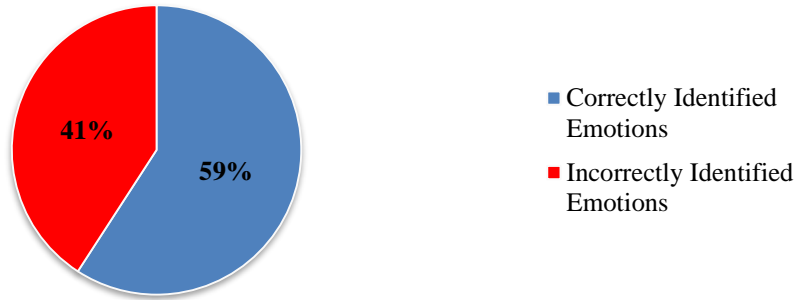


Figure 9. The general emotion detection rate of the algorithm.

As previously mentioned, the current implementation detects only 5 emotions out of the 8 available in the dataset: anger, disgust, sadness, surprise, and happiness. Each image from the set has been labeled with the appropriate emotion that needs to be detected. Out of the 327 images, only 284 contain an emotion that can be detected by the algorithm and out of the 284 images, 168 were correctly identified emotions, the algorithm having a ~59% success rate, as it can be seen in Figure 9.

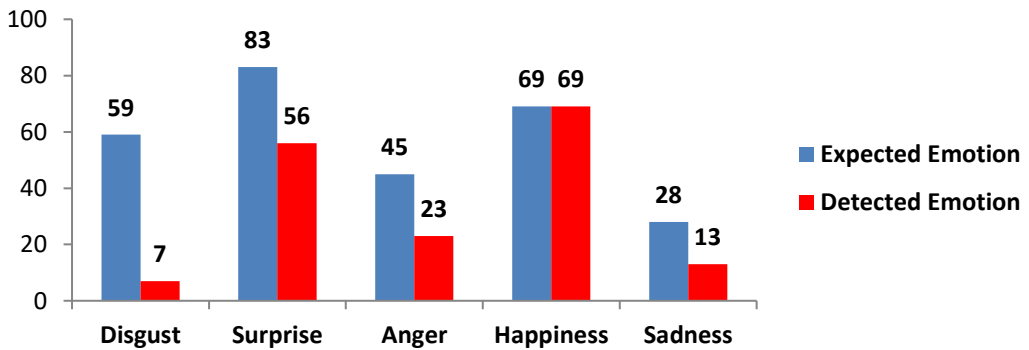


Figure 10. Number of correctly identified emotions for each image provided

Each emotion has a different detection rate, as it is shown in Figure 10. The highest detection rate was found for *happiness*, where 100% of the images were correctly identified and the lowest emotion detection rate was for *disgust*, where only 11.86% of the images were correctly identified. In between were *sadness*, *anger*, and *surprise* with a 46.43%, 51.11% and 67.47% detection rate, respectively.

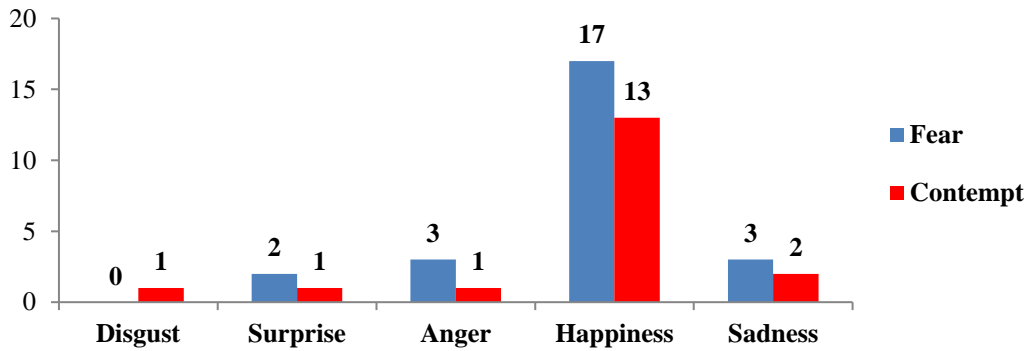


Figure 11. Incorrect detection of unavailable emotions

As previously stated, the dataset contains images with emotions that the current implementation cannot detect. Two of these emotions are *fear* and *contempt*. It has been noticed that both of them are generally interpreted as *happiness*, 68% of the time for *fear* and 72% of the time for *contempt* (Figure 11).

Aside from *happiness*, none of the other available emotions have a perfect detection rate. In Table 1 there are shown the complete results from the dataset. In general, when it is not correctly detected, each emotion is interpreted as *happiness*.

Table 1. Table of complete results for each image.

		Output				
		<i>Disgust</i>	<i>Surprise</i>	<i>Anger</i>	<i>Happiness</i>	<i>Sadness</i>
Input	<i>Disgust</i>	7	1	15	30	6
	<i>Surprise</i>	1	56	6	17	3
	<i>Anger</i>	3	0	23	14	5
	<i>Happiness</i>	0	0	0	69	0
	<i>Sadness</i>	2	1	2	10	13
	<i>Fear</i>	0	2	3	17	3
	<i>Contempt</i>	1	1	1	13	2

## 7. CONCLUSION

The approach presented in this pilot project will be integrated on an online platform and used by a team of doctors from a municipal hospital in Bucharest during the therapy sessions conducted there. Future work may include monitoring the results by a certified therapist, improving the neural network by continuously feeding it with data from the

therapy sessions and creating an UI to easily monitor the emotional changes throughout a session. The idea here is that, if willing, the therapist could “approve” or “disapprove” with the results of the software for a specific image or frames by entering his/her choice in the UI and so the neural network will learn from this good/bad review and evolve while being used. There will also be further experiments with different type of topologies for the neural network. Besides this, a bolder plan for the future is to add emotion detection based on the gestures of subjects. This could be achieved by adding another neural network that would be trained with different postures that signify specific emotions. Then the results would be combined with the emotions detected from the facial movement so that a more accurate final result could be obtained.

## **8. REFERENCES**

- [1] L. B. Krithika, K. Venkatesh, S. Rathore and M. Harish Kumar, "Facial recognition in education system," Conf. Series: Materials Science and Engineering, 3 December 2017.
- [2] "Guidelines for the Practice of Telepsychology," Association American Psychological, [Online]. Available: [http:// www.apa.org/ practice/ guidelines/ telepsychology.aspx](http://www.apa.org/practice/guidelines/telepsychology.aspx). [Accessed 18 January 2018].
- [3] Novotney, "A growing wave of online therapy," American Psychological Association, 2017. [Online]. Available: [http:// www.apa.org/ monitor/ 2017/ 02/ online-therapy.aspx](http://www.apa.org/monitor/2017/02/online-therapy.aspx). [Accessed 18 January 2018].
- [4] "TensorFlow," Google Brain Team, [Online]. Available: [https:// www.tensorflow.org/](https://www.tensorflow.org/). [Accessed 06 February 2018].
- [5] D. V. Sang, N. Van Dat and T. Do Phan, "Facial expression recognition using deep convolutional neural networks," in IEEE Xplore, Hue, Vietnam, 2017.
- [6] E. M. Cojocea, "Analiza comportamentului uman. Recunoasterea emotiilor faciale, Disseration Thesis," Unpublished Work, Bucuresti, 2017.
- [7] "Pexels," [Online]. Available: [https:// www.pexels.com/ photo/ close-up- photography-of-a-girl-smiling-713312/](https://www.pexels.com/photo/close-up-photography-of-a-girl-smiling-713312/). [Accessed 30 January 2018].
- [8] "Face Detection using Haar Cascades," OpenCV, [Online]. Available: [https:// docs.opencv.org/ 3.3.0/ d7/ d8b/ tutorial\\_py\\_face\\_detection.html](https://docs.opencv.org/3.3.0/d7/d8b/tutorial_py_face_detection.html). [Accessed 06 February 2018].
- [9] "Lanczos resampling," Wikipedia, [Online]. Available at: [https:// en.wikipedia.org/ wiki/ Lanczos\\_ resampling](https://en.wikipedia.org/wiki/Lanczos_resampling). [Accessed 06 February 2018].
- [10] T. Kanade, J. F. Cohn and Y. Tian, "Comprehensive database for facial expression analysis," in The Fourth IEEE International Conference on Automatic Face and Gesture Recognition, Grenoble, France, 2000.
- [11] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar and I. Matthews, "The Extended Cohn-Kanade Dataset (CK+): A complete expression dataset for action unit and emotion-specified expression," in The Third International Workshop on CVPR for Human Communicative Behavior Analysis, San Francisco, USA, 2010.

- [12] "About TensorFlow," [Online]. Available: [https:// www.tensorflow.org/](https://www.tensorflow.org/) . [Accessed 23 January 2018].
- [13] Thibaud Senechal, Daniel McDuff and Rana el Kaliouby, "Affectiva". Available at: [https:// www.affectiva.com/ wp-content/ uploads/ 2017/ 03/ Facial-Action-Unit-Detection-using-Active-Learning-and-an-Efficient-Non-Linear-Kernel-Approximation.pdf](https://www.affectiva.com/wp-content/uploads/2017/03/Facial-Action-Unit-Detection-using-Active-Learning-and-an-Efficient-Non-Linear-Kernel-Approximation.pdf) [Accessed 16 March 2018].
- [14] "Facial Action Coding System" by Imotions. Available at [https:// imotions.com/ blog/ facial-action-coding-system/](https://imotions.com/blog/facial-action-coding-system/). [Accessed 12 March 2018].
- [15] "Facial Action Coding System", [https:// www.cs.cmu.edu/ ~face/ facs.htm](https://www.cs.cmu.edu/~face/facs.htm) [Accessed 4 June 2018]
- [16] A. Rahimi and B. Recht, "Random features for large-scale kernel machines" found in "Advances in neural information processing systems", pages 1177-1184, 2007.
- [17] S. Tong and E. Chang, "Support vector machine active learning for image retrieval" found in Proceedings of the ninth ACM international conference on Multimedia, pages 107-118, 2001.
- [18] <http://cs231n.github.io/convolutional-networks/> [Accessed 4 June 2018]
- [19] [https:// ujjwalkarn.me/ 2016/ 08/ 11/ intuitive-explanation-convnets/](https://ujjwalkarn.me/2016/08/11/intuitive-explanation-convnets/) [Accessed 4 June 2018]