Real-time Facial Emotion Classification Using Deep Learning

Emre Dandıl[1,*], Rıdvan Özdemir[2]
[1]*Department of Computer Engineering,Faculty of Enginnering, Bilecik Seyh Edebali University, Gulumbe Campus, 11210, Bilecik, Turkey*
[2]*Electronics and Computer Engineering Department, Institute of Science, Bilecik Seyh Edebali University, Gulumbe Campus, 11210, Bilecik, Turkey*

***Abstract*—Facial emotion recognition has an important position in the computer vision and artificial intelligence field. In addition, real-time face recognition applications have to be able to be performed at high speed and accuracy rate in order to make human-computer interaction successful in increasing artificial intelligence and humanoid robot applications. In this study, we detected the faces on real-time video data to recognize the anger, fear, happy, surprise, sad and neutral emotions upon these detected faces using deep learning methods. We created our own dataset to use in this study for six different facial emotions. At first stage, we created a convolutional neural network and trained it over our dataset by scratching method and we achieved 50% accuracy rate. Then, we increased the number of images in our database by 3 times, and get better accuracy which is 62%. Thanks to transfer training method and AlexNet's pre-trained networks, we reached 74% accuracy rate after increasing the number of images 80% in the dataset. In addition, we achieved 72% accuracy rate when we test our network which is trained with our own dataset with the Compound Emotion dataset. The basic reason of this decrease can be angry emotion because there are differences poses between our dataset and Compound Emotion dataset for angry emotion images. However, we obtained 100% accuracy rate for happy emotion and 89% for sad emotion. It has been seen that the work we are doing gives successful results when tested with different people in different ambient and light conditions.**

***Keywords*—classification, convolutional neural network, deep learning, emotion recognition, face recognition.**

I. INTRODUCTION

Facial expressions are an instant form of expression of people's feelings. Facial expression based emotion recognition practices have a significant role in computer vision and artificial intelligence studies[1]. Although emotion recognition can be carried out with wearable sensors, it is more important and more flexible to carry out emotion recognition with visual inputs without having a physical connection [2]. As a result of advances in computer technology, the application of artificial intelligence systems has increased. It is very important to realize emotion recognition based on facial expressions in real-time applications where human-computer interaction such as humanoid robot applications using artificial intelligence systems. Applications in many areas such as medicine, shopping and entertainment sector have gained great advances as a result of progress in facial emotion recognition technology [3]. Facial emotion recognition is a difficult application to perform because most of the datasets used in facial emotion recognition studies do not have enough number of images.

Most of the studies on real-time emotion recognition have been proposed on static images. Number of researches on facial emotion recognition with computer vision techniques has shown a significant rise in parallel with the development of computer technology, when the studies in the literature are reviewed over the last two decades. Although there are many applications relating face recognition and detection, convolutional neural networks (CNNs) are not generally preferred for the visualization of visual properties in most of them [4]. Breuer and Kimmel [5] used visual techniques to understand a CNN model which is trained with a variety of datasets for emotion recognition. They test the performance of CNN on facial emotion recognition datasets and also on some facial emotion recognition applications. Jung et al. [6] developed a technique using two different types of CNNs. While one of them is feature extraction from the visual datasets, the second one is extraction of the geometric properties of the facial landmarks. Kahou et al. [7] who trains deep convolutional neural networks using a visual dataset and then applies it to video proposed an example of rare studies on video.

In this study, we propose an emotion recognition system based on facial expressions from video frames in real time with deep learning. Unlike similar applications, system keeps high accuracy results under different environmental conditions and with different models. An original dataset was created for six basic emotions such as angry, fear, happy, neutral, sad and surprise. Then, the CNN model was trained with this dataset. Experimental studies also were conducted with this dataset on AlexNet, pre-trained network. Moreover, the obtained CNNs were compared with a common test data.

We explain detailed information about the created dataset and features, data processing, face detection, CNN architecture and transfer learning techniques in Section 2. The experimental studies such as facial emotion recognition

Manuscript received November 27, 2018; accepted July 08, 2019.
* Corresponding author: emre.dandil@bilecik.edu.tr

on test dataset and real-time facial emotion recognition are given in Section 3. Finally, we obtain the overall results and possible studies in the future in Section 4.

II. MATERIAL AND METHOD

The following steps are performed for this study. a dataset was created for six basic emotions from images which are available on public searches in the internet. Face detection process is applied to these images by Viola-Jones algorithm to find face positions on the images and crop them for creating new images files. The dataset is divided into two parts as a training dataset and test dataset. The CNN is trained with the training dataset and the performance of the network was tested with the test dataset. Finally, the resulting network is used to classify the facial emotions on the video frames. The flowchart of this study is denoted in Figure 1.

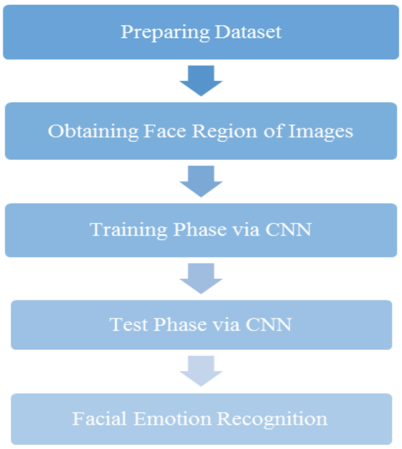


Fig. 1. Flowchart of proposed emotion recognition system

2.1. Dataset and Features

The dataset which is used in this study was created from public images [8] distributed for free. These images are labelled according to basic facial expressions like anger, fear, happy, neutral, sad and surprise. 600 images were obtained with 240 pixels smallest margin for each class after some of mislabelled images were eliminated. A dataset is created from these images which contain 3600 images in RGB format. After that, this dataset is divided into training and test datasets. 3360 images for the training dataset and 240 images for the test dataset were chosen. Some sample images of basic facial emotions in the training dataset are shown in Figure 2. Some of the images from the test dataset are given in Figure 3.

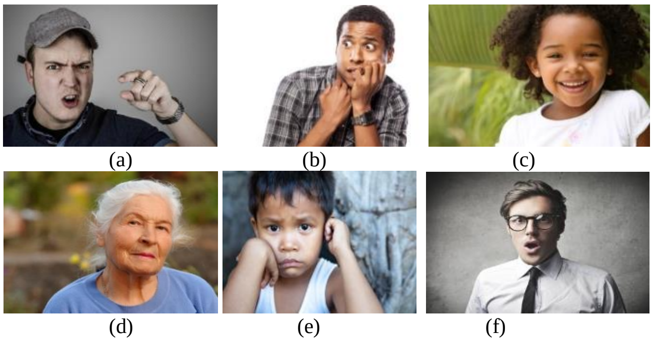


Fig. 2. Examples of basic facial emotions from the dataset, (a) angry, (b) fear, (c) happy, (d) neutral, (e) sad and (f) surprise

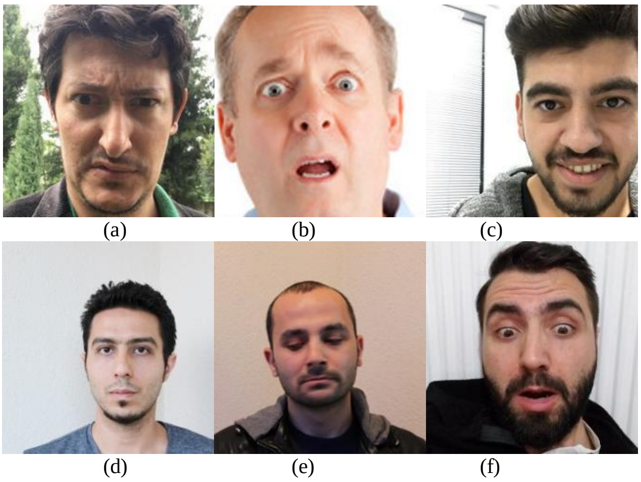


Fig. 3. Examples of basic facial emotions from the test dataset, (a) angry, (b) fear, (c) happy, (d) neutral, (e) sad and (f) surprise

2.2. Data Processing and Face Detection

Viola-Jones algorithm [9] was used for detection of a face in this study. The Viola-Jones algorithm uses rectangular features to determine the face in the image. Thus, it provides a very high speed operation as well as highly accurate results [9]. An example of face detection on the image using Viola-Jones algorithm is shown in Figure 4. Data processing was applied only the region of image where contains the faces instead of entire image thanks to Viola-Jones face detection algorithm. The images in the dataset were augmented in lower and higher contrast and brightness values after data processing in order to achieve better results in the training phase of the CNN. Thus, a dataset was obtained which has 1800 images for each basic facial emotion and 10800 images in total. After that, each image in the dataset is resized to 227 X 227 pixels. Data augmentation to enlarging training dataset using image processing techniques is presented in Figure 5.

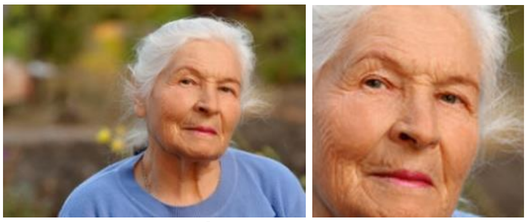


Fig. 4. Face detection in training images

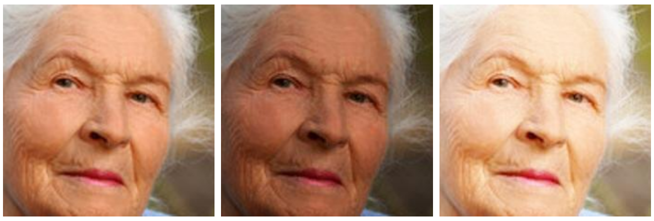


Fig. 5. Data augmentation using image processing techniques

2.3. *Convolutional Neural Network (CNN)*

CNN is a special model of multilayer artificial neural networks developed for computer vision applications. Architecture of CNN includes separate layers with distinctive tasks, such as convolution, pooling and fully connected. These structures are sequentially aligned to create CNN. In the first parts of this structure, feature extraction operations are performed and the classification process takes place in the final layers [10]. In this study, 5-layer CNN architecture was created, including three convolution layers, three pooling operations and two fully connected layers. The 5-layer CNN model trained with created training dataset. The architecture of the 5-layer CNN is shown in Figure 6.

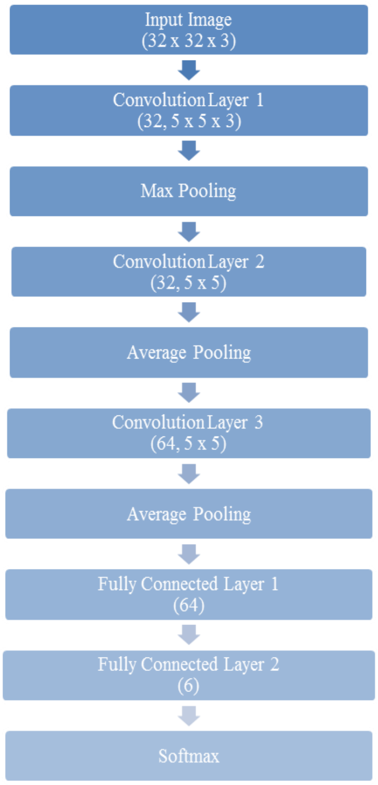


Fig. 6. The architecture of the 5-layer CNN.

2.4. Transfer Learning

Transfer learning is a widely used deep learning technique. It takes weights from a pre-trained network which is learned on a big dataset. Transfer those weights to various layers and retrain the last layers of network [11]. A dataset is required for training after creating a CNN architecture. If the dataset is not big enough the created CNN's test results will not be good as desired. In this case, the training dataset has to be enlarged. However, this is not possible and it may be necessary to work with a limited number of labeled training data in some cases. In these conditions transfer learning technique could be used to get best accuracy rate. In other words, transfer learning can provide high accuracy results with small datasets.

Maximum accuracy rate was 62% after experimental studies are carried out with 5-layer CNN. So, using transfer learning to get better performance results is considered. For this purpose, AlexNet has been selected because its input image size fits to the image sizes of the dataset's image size. AlexNet won the ImageNet competition in 2012, which reduced the lowest error rate of 26.2% to 15.4%. It contains 60 million parameters and 650 thousand neurons. The network structure consists of five convolution layers with max-pooling at the end of some and three fully connected layers and a 1000-path softmax feed at the end of them. The number of neurons in the last fully connected layer has been reduced from 1000 to 6 since there are six different class in this study.

III. EXPERIMENTAL RESULTS

In this study, experimental works are performed for facial emotion recognition on both the test dataset and real-time video by the proposed method.

3.1. Facial Emotion Recognition on Test Dataset

A dataset consisting of 3600 labeled images was created for six different facial emotion class. 3360 of the images were used in training dataset and 240 of the images were chosen for the test dataset. The 5-layer CNN model was trained with the dataset. The value of parameters that are used for training presented in Table I.

TABLE I. TRAINING PARAMETERS.

Parameter	**Value**
Optimizer	SGDM
Initial Learn Rate	0.001
Max Epochs	20
Mini Batch Size	100

As a result of the classification operations carried out on the test dataset with 5-layer CNN, the accuracy rate of 50% has been reached as seen in Table II. As it can be seen from this table, the most successful facial emotion was happy in which 30 of the 40 images in the test were correctly predicted, while the most inaccurate facial emotion was fear that only 13 images were correctly predicted. In the next step, the accuracy rate that can be seen from the confusion matrix in Table II increased to 62%, after the CNN model was trained with the dataset which was enlarged by data augmentation. The most successful facial emotion was happy with 82% accuracy and the most unsuccessful was fear with 49%. As it can be seen from the confusion matrix in Table II (b), the greatest error was between the class of fear and surprise cause of similar features. On the other hand, the highest success rate was achieved in the "happy" class.

TABLE II. 5-LAYER CNN: CONFUSION MATRIX OF NETWORK TRAINED WITH 6X560 DATASET (ACCURACY 50%)

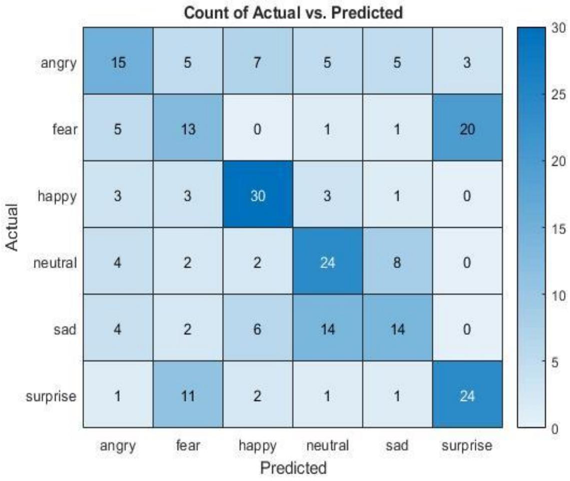
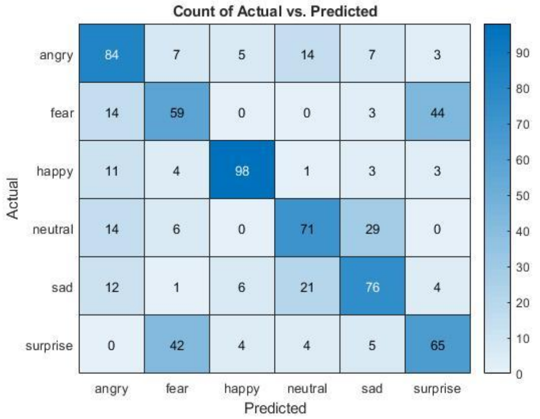


TABLE III. . 5-LAYER CNN: CONFUSION MATRIX OF NETWORK TRAINED WITH 6X1680 DATASET (ACCURACY 62%).



Because the result of the test (accuracy 62%) is below the desired level, transfer learning method was decided to use. Transfer learning is a method that allows obtaining high accuracy with small training dataset. AlexNet architecture is chosen for transfer learning. 1000-way last fully connected layer of AlexNet was changed to 6-way for 6 basic facial emotion class and fine tuning operation was performed with the training dataset. The value of parameters, which are used for training, presented in Table IV.

TABLE IV. TRAINING PARAMETERS.

Parameter	Value
Optimizer	SGDM
Initial Learn Rate	0.001
Max Epochs	10
Mini Batch Size	64

The network which was obtained by transfer learning reached 74% accuracy rate after it was performed on the test dataset. Confusion matrix of this test can be seen in Table V. As it can be seen from the table, the most successful class is neutral with 35 correct predictions in 40 images. The worst class is fear that only 19 images were correctly predicted. In the next step, fine tuning was performed with the enlarged training dataset to network and the same parameter values were used like the previous training. As it can be seen from the confusion matrix in Table VI, the accuracy rate of the network increased to 80%. The most successful facial emotion class is sad with 93% accuracy, while the worst fear with 60% accuracy.

TABLE V. CONFUSION MATRIX OF NETWORK TRAINED WITH 6X560 DATASET (ACCURACY 74%),

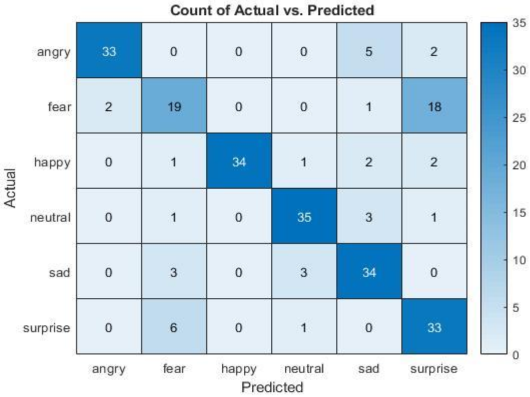


TABLE VI. CONFUSION MATRIX OF NETWORK TRAINED WITH 6X1680 DATASET (ACCURACY 80%).

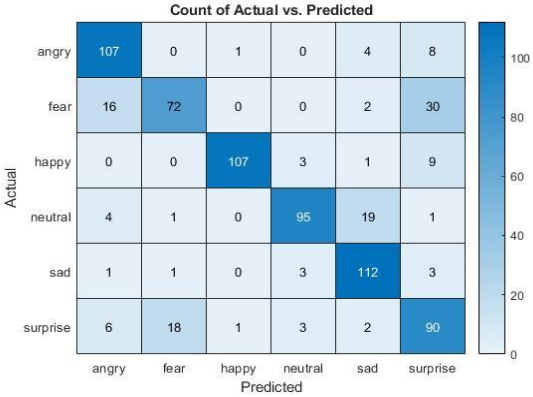


Figure 7 shows the facial emotion classification on a sample image. All the face is found in the test image via Viola-Jones face detection algorithm like in this example and emotion of these faces classified by CNN with the score. Finally the facial emotion class and it's score printed over the face which is detected by Viola-Jones algorithm.

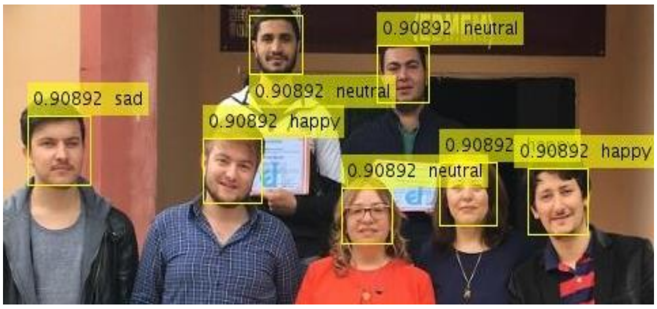


Fig. 7. Facial emotion recognition on an image.

3.2. *Real-time Facial Emotion Recognition*

The most successful classification was made as a result of using transfer learning method for real-time facial emotion classification, after the experimental work on the CNN with the test dataset. It was seen that the network, which is trained with enlarged training dataset, could give successful results also in the changing light and environmental conditions. The faces in the image were detected individually by applying Viola-Jones algorithm on each frame of a video for a real-time facial emotion classification. After face detection, facial emotion recognition was performed and the classification results were printed on the frames in real-time. This process was applied to each video frame individually and the facial emotion recognition was performed on the real-time video. Figure 8 shows examples of facial emotion classification for angry, fear, happy, neutral, sad and surprise expressions that obtained in real-time on video frames.

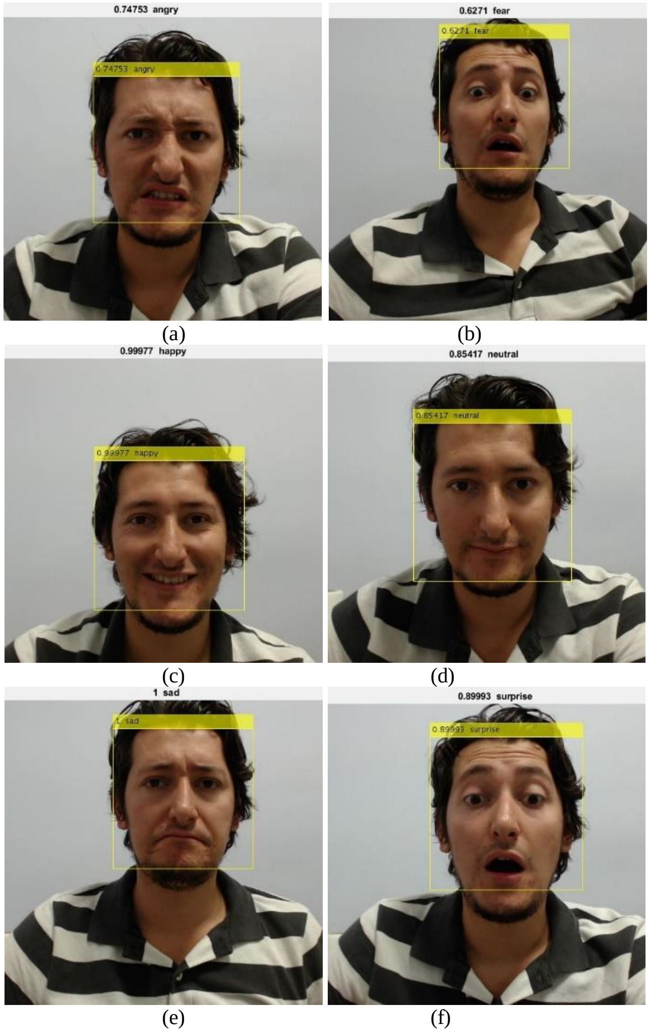


Fig. 8. Results of application on a real-time video (a) angry, (b) fear, (c) happy, (d) neutral, (e) sad and (f) surprise

IV. RESULTS AND DISCUSSION

In the experimental studies, 50% accuracy was achieved with 5-layer CNN. It was seen that the performance was increased to 62% after the test was repeated by the network, which was trained with enlarged training dataset. In order to increase this accuracy rate, the experimental studies were carried out using transfer learning method with AlexNet. It increased accuracy to 74%. After the new network was trained with the enlarged dataset, the performance is risen to 80%.

The faces on the image or video, were found using Viola-Jones face detection algorithm then emotion of these faces were recognized by created CNN. This application was tested in real-time on the video, and the successful results were obtained in changing light and environment conditions. The maximum number of error occurred between fear and surprise classes because of the similarities in the features of fear and surprise facial emotions. The number of errors can be reduce by reviewing the training dataset and removing misclassified images in it.

The number of facial emotions could be increased to seven by adding disgust to the states of emotion in the following studies. In addition, some characteristic information could be accessed through people by real-time facial emotion classification applications. Also an application might be used to estimate whether a positive or negative incident occurred from the total facial emotions of the persons in the group.

REFERENCES

[1] Shan C Shaogang G and Peter W. M 2009 *Facial expression recognition based on local binary patterns: A comprehensive study* (Image and vision Computing 27 6) p 803-816.

[2] Chul Ko B 2018 *A Brief Review of Facial Emotion Recognition Based on Visual Information* (Sensors 18 401) p 1-2

[3] Kołakowska A Landowska A Szwoch M Szwoch W and Wr´obel M R 2014 *Human-Computer Systems Interaction: Backgrounds and Applications* (Cham: Springer International Publishing) chapter 3 p 51–62

[4] Hoang Le T 2011 *Applying Artificial Neural Networks For Face Recognition* (Advances in Artificial Neural Systems vol 2011 Article ID 673016) p 15

[5] Breuer R Kimmel R 2017 *A Deep Learning Perspective on the Origin of Facial Expressions* (arXiv 1705.01842)

[6] Jung H Lee S Yim J Park S Kim J Joint 2015 *Fine-Tuning in Deep Neural Networks For Facial Expression Recognition* (In Proceedings of the IEEE International Conference on Computer Vision Santiago Chile) p 2983–2991

[7] Kahou S E Pal C Bouthillier X Froumenty P Gülçehre C Memisevic R Vincent P Courville A Bengio Y Ferrari R C et al 2013 *Combining modality specific deep neural networks for emotion recognition in video* (In Proceedings of the 15th ACM on International conference on multimodal interaction ACM) p 543–550

[8]https://stock.adobe.com/search/images?load_type=search&native_visual_search=&similar_content_id=&is_recent_search=&k=facial+emotions

[9] Viola P and Jones M J 2004 *Robust Real-Time Face Detection* (International Journal of Computer Vision vol 57 no 2) p 137-154

[10] Aydilek I B 2017 *Derin Öğrenme ile Tüketilen Besin İçeriklerinin Yaklaşık Kestirimi* (International Conference on Computer Science and Engineering) p 2

[11] Savoiu A Wong J 2017 *Recognizing Facial Expressions Using Deep Learning* (Stanford University) p 4

[12] Krizhevsky A Sutskever I and Hinton G E 2012 *ImageNet Classification with Deep Convolutional Neural Networks* (NIPS) p 5