Real-Time Deep Facial Expression Recognition using Feature Extraction and Facial Landmarks

Chenhao Wang

CMPT 757 Final Project Report

Abstract

Real-time facial expression recognition has been an active field of research over the past several decades. This work aims to classify physically disables people's emotion expressions based on feature extraction and facial landmarks using a convolutional neural network (CNN) that works effectively in uneven lightning and subject head rotation, different backgrounds, and various skin tones. At the same time, I also examined the benefits of different FER model designs, which include a combination of different pre-processing methods, network architectures, network design strategies, and learning strategies to improve the existing techniques. In this Project, I built a model that uses Convolution Neural Network to successfully classify faces as one of the core seven emotions: anger, contempt, disgust, fear, sadness, happiness, surprise and neutral.

Keywords: *Facial expression recognition, emotion classification, facial feature analysis, computer vision, image processing*

1 Introduction

One of the important ways humans display emotions is through facial expressions. Facial expression recognition is one of the most powerful, natural and immediate means for human beings to communicate their emotions and intentions. Therefore, I concentrated on building a successful emotion recognition model that can work in real-time. Due to the fast advancement of artificial intelligence and machine learning, its application is actively being used in many domains. Moreover, machine learning algorithms have played a significant role in pattern recognition and pattern classification problems, especially in facial expression recognition.

The implementation can be broadly categorized into four stages: dataset development, data pre-processing, convolutional neural network (CNN) construction and real-time interface development. An appropriate facial database was to be obtained which serves as our training and our testing dataset, essentially consisting of humans displaying labelled emotions with the images [1]. The first stage deals with data collection and cleaning that modify the dataset to fit my model. The second stage works on face area extraction and normalization, in some experiments it also extracts facial landmarks [2] as identified critical features for emotion detection. In Third stage, the VGG-like network is built. The input feature vectors are given as input to the VGG-like network that is trained to then classify what emotion is being shown by the human. The final stage is developing a real-time interface to detect a person's emotion either in a live stream camera or a video.

2 Related Work

Numerous studies have been conducted on automatic facial expression analysis because of its practical importance in sociable robotics, medical treatment, driver fatigue surveillance, and many other human-computer interaction systems. In the field of computer vision and machine learning, various Facial Expression Recognition (FER) systems have been explored to encode expression information from facial representations. As early as the twentieth century, Ekman and Friesen [3] defined six basic emotions based on cross-culture study [4], which indicated that humans perceive certain basic emotions in the same way regardless of culture. These prototypical facial expressions are *anger*, *disgust*, *fear*, *happiness*, *sadness*, *surprise*.

FER systems have been implemented in a multitude of ways and approaches. The majority of these approaches have been based on facial features analysis. Yu and Zhang [5] proposed a CNN architecture specialized on emotion recognition performance. They proposed two novel constrained optimization frameworks to automatically learn the network ensemble weights by minimizing the loss of ensemble network output responses. Ghimire et al. [6] used the concept of position-based geometric features and angle of 52 facial landmark points. For hybrid methods, some approaches [7] have combined geometric and appearance features to complement the positive outcomes of each other and in fact, achieve better results in certain cases. In video sequence, many systems [6, 8, 9] are used to measure the geometrical displacement of facial landmarks between the current frame and previous frame as temporal features.

Those studies use component-based facial features to combine geometric and appearance information of face. Conventional FER systems in general use considerably lower processing power and memory when balanced with deep learning based approaches and are thus, still being researched for use in real time mobile systems because of their low computational power and high degree of reliability and precision.

3 Approach

In order to accomplish my task of developing a real-time emotion detection system, I have to get several components working independently. Below summarizes my general approach:

3.1 Dataset Development

The dataset comes from Expression in-the-Wild (ExpW) data collection [1], which contains 106,962 RGB images downloaded using Google image search with 7 basic expressions, which are *angry*, *disgust*, *fear*, *happy*, *sad*, *surprise*, and *natural*. Each of the face images was manually annotated as one of the seven basic expression categories. The labels saved in one text file with information about image name, face id in image, face bounding box coordinate, and expression label.

To build the dataset for my own network, I removed images with no face and got 68,096 images after cleaning. then I divided labels into each single text file with the same name as its corresponding image. In the new label file, each row represents one detected face with expression label and bounding box coordinate. In the experiments, I used 58,096 images for training and 10,000 for testing results.

3.2 Data Pre-processing

Before training the deep neural network to learn meaningful features, pre-processing is required to align and normalize the visual semantic information conveyed by the face. The network accepts face feature vectors as input and gets the expression label in result. There are different pre-processing steps in experiments. The experiments before milestone, the network accepted normalized 48 by 48 grayscale face images as input, so the pre-processing including: crop face areas based on bounding box, convert and resize to 48 by 48 grayscale images. In order to balance the data distribution in each emotion label, the class-based data augmentation is also applied, which contains horizontal flipping, small degree rotation, image shearing, and etc. The Figure 1 shows the data distribution in each emotion label before class-base augmentation, and the Figure 2 indicates the occurrence of each emotion label after class-based augmentation. It is obvious that the augmentation reduce the differences between labels to balance the distribution of data. But due to the lack of dataset in label *fear*, after augmenting the data, the number of data in *fear* is lower than others.

Moreover, I got inspiration from milestone discussion with classmates that add facial landmarks as extra features into the network. Facial landmarks are used to localize and represent salient regions of the face, such as eyes, eyebrows, nose, mouth and jawline. Facial landmarks have been successfully applied to face alignment, head pose estimation, face swapping, blink detection and much more [10]. In this project, I applied the FacemarkLBF model [2] to extract facial landmarks as an additional input layer. The Figure 3 shows the identified critical landmarks using the FacemarkLBF model. In testing and live-stream video detection processing, I applied face alignment using OpenCV Cascade classifier to detect faces then removed

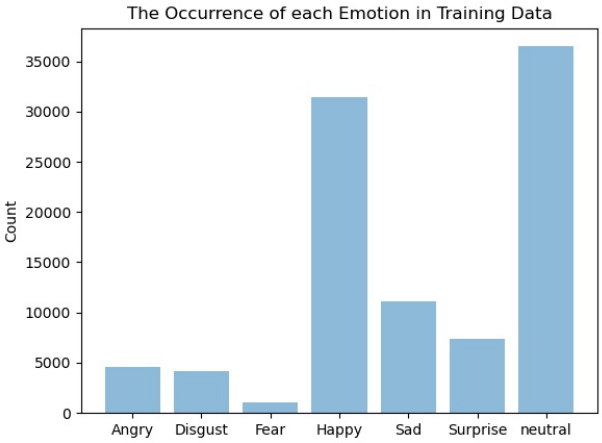

Figure 1. The Occurrence of each Emotion in Training Data before Augmentation

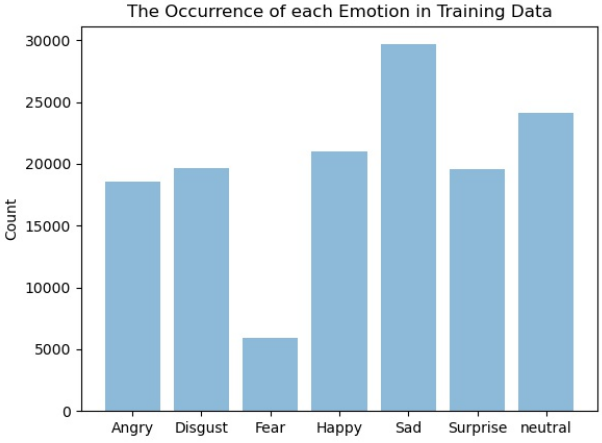

Figure 2. The Occurrence of each Emotion in Training Data after Augmentation

background and non-face areas instead of using bounding box. Face detection using Haar feature-based pre-trained cascade classifiers is an effective object detection method proposed by Paul and Michael [11]. Then the cropped face areas would be fed into my facial expression recognition network.

3.3 CNN Construction

CNN has been extensively used in diverse computer vision applications, including FER. The convolution operation is associated with three main benefits: local connectivity, which learns correlations among neighboring pixels; weight sharing in the same feature map, which greatly reduces the number of the parameters to be learned; and shift-invariance to the location of the object. The initial network architecture is inspired by the family of VGG networks [12]. All the convolutional layers in the network are using 3 by 3 kernels. The number of fil-

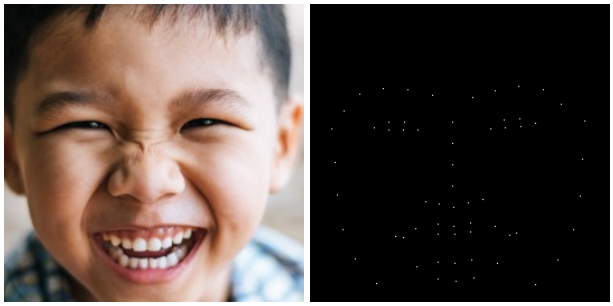

Figure 3. The Extracted Facial Landmarks using FacemarkLBF

ters learned by each convolutional layer will be doubled as the network becomes deeper. The reason I used VGGNet is that it significantly outperforms models in the classification task. I added another two groups of conv + activation + batchnorm layers with one max pooling layer and dropout on the network to make the network deeper, and I added L2 regularization in the network to avoid overfitting.

The Figure 4 indicates the general pipeline of my deep facial expression recognition system. The RGB images would be applies face alignment to remove non-face areas. Although face detection is the only indispensable procedure to enable feature learning, further face alignment using the coordinates of localized landmarks can substantially enhance the FER performance . This step is crucial because it can reduce the variation in face scale and in-plane rotation. After pre-processing stage, the normalized image and facial landmarks would be accepted by VGG-like network to predict the related emotion label with the highest confidence.

3.4 Real-Time Interface

OpenCV allowed us to get images from our laptop's webcam. There is a function that reads one frame from the video source and returns the actual video frame read. For each frame, I applied a Cascade classifier aligning the face area. The face images were pre-processed in the same manner as training and testing the dataset: convert it to grayscale, resize to 48 by 48, and extract the facial landmarks. A prediction is produced for each frame, and the face bounding box and emotion label would be shown on the image at the same time.

4 Experiments and Results

In this project, I implemented four experiments to compare the performance with different pre-processing methods, network architectures, network design strategies. Then the results are evaluated by both quantitative metrics and qualitative performance.

4.1 Experiments Implementation

The model was constructed based on Keras. The loss function is categorical cross-entropy which is good in multi-class classification tasks with imbalancing data distribution. I used

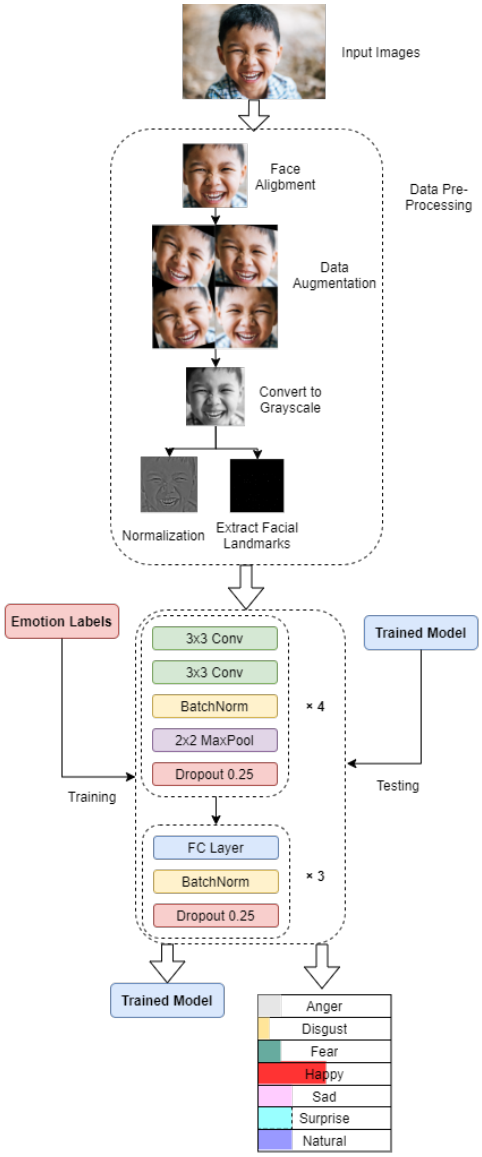

Figure 4. The General Pipeline of Deep Facial Expression Recognition System.

an SGD optimizer with a base learning rate of 1e-5 for all experiments. In the training process, the model split 20 percentages of the dataset for validation. The evaluation is based on precision, recall accuracy, and f1-score, then save the model with best validate f1-score.

Experiment 1: Baseline The first experiment trained the baseline model, the initial network takes 48 by 48 grayscale images as input with 128 batches. In the pre-processing, it only used horizontal flip to augment data. The model takes one channel input without facial landmarks.

Experiment 2: Baseline with Class-based Augmentation The second experiment carried out with same network architecture and hyper-parameters, but implemented class-based data augmentation before training. The class-based augmentation is used to balance the data in each emotion label. In my dataset,

the *natural* is the most common expression, so the augmentation would be skipped under this label. However, the label *fear* has few images, so I augmented data multiple times. The less amount of images in that label, the more times augmentation would be applied. In the next experiments, the class-based augmentation is also applied.

Experiment 3: Additional input with Partial Facial Landmarks Third experiment extracts facial landmarks as additional input layer. It only keeps the partial landmarks as critical features, which include eyes, eyebrows, nose and mouth, and removes the landmarks of jawline. The reason I remove the jawline landmarks is that the jawline captures very little differences between emotions, it would not change too much in different facial expressions.

Experiment 4: Gaussian Smooth image with Partial Facial Landmarks In my last experiment, the inputs contain partial facial landmarks as previous experiment, and images with Gaussian filter on grayscale image to smooth-out the noise and zero center normalization. This pre-processing method was proposed by Fuzail [13] while using grayscale image and facial landmarks as inputs for facial expression recognition.

4.2 Results

For this project, I have two ways used to evaluate the performance of my model. The first is the obvious quantitative results, such as precision, recall, accuracy, and f1-score, in which I examine the statistical success of my model in a predicted labeled data set. The second success measure is how well it is able to classify test set and live-streamed images. Since there are no labels for live-streamed images, only us knowing which expression we are trying to portray, it is difficult to quantitatively examine these results without labels. In the below analysis, I describe the results with respect to both of these ways.

4.2.1 Quantitative Evaluation

After training the neural networks as explained above, the testing set of images was used to check the quantitative performance of each proposed FER system. The results of the test have been presented as quantitative evaluation tables as shown in Table 1 and the f1-score per emotion as shown in Table 2.

Exp No.	Accuracy	Precision	Recall	F1-score
1	0.69	0.64	0.66	0.64
2	0.66	0.69	0.72	0.70
3	**0.74**	**0.74**	**0.73**	**0.74**
4	0.68	0.66	0.68	0.66

Table 1. Quantitative Evaluations in Experiments.

From Table 1, it is observed that the experiment using critical facial landmarks as additional input gets the best results compared to other experiments. The accuracy is 0.74 which is higher than others, and f1-score is the highest as well. The highest f1-score indicates that the additional facial landmarks could improve the performance of convolutional FER systems. Besides that, from other results, compare to the results from experiment 1 and experiment 2 as example, the accuracy of experiment 1 is higher than experiment 2, however, the f1-score of experiment 1 is lower than experiment 2. It is indicated that for the classification tasks with imbalanced multiple classes, only the accuracy may not be enough to evaluate the model. Also, the class-based data augmentation improved the performance of the FER system. In addition, the experiment 4 using Gaussian smoothed image and zero-center normalization would not increase the score as I expected.

Emotions	Exp 1	Exp 2	Exp 3	Exp 4
Anger	0.60	0.48	**0.70**	0.50
Disgust	0.00	0.07	**0.34**	0.08
Fear	0.00	0.06	**0.41**	0.12
Happy	**0.84**	0.79	**0.84**	0.80
Sad	0.50	0.42	**0.60**	0.45
Surprise	0.61	0.52	**0.68**	0.55
Natural	**0.77**	0.72	**0.77**	0.73

Table 2. F1-Score of Each Emotion in Experiments.

The Table 2 shows the specific f1-score in each emotion label. It is clearly to see that the emotions of *happiness* and *natural* are being detected relatively well with 0.7+ f1-score, which indicates that the facial reactions to when a person is *happiness* or *natural* are more uniform than the other emotions leading to high success rates. The emotions of *fear* and *disgust* have their score below 0.4 indicating that people may have the same facial reaction for two different emotions as well as having different facial reactions for the same emotions.

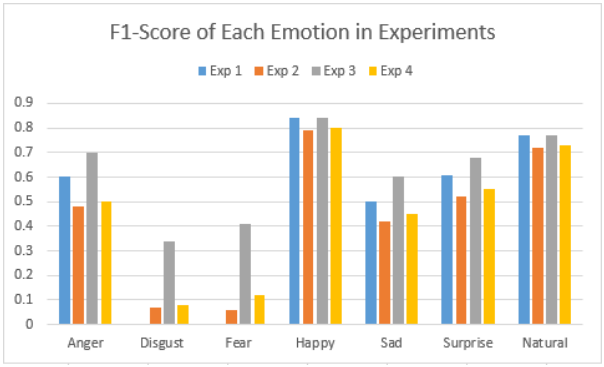

Figure 5. F1-Score of Each Emotion in Experiments.

The Figure 5 plots the f1-score in different emotions. From the figure, we can easily know that the facial landmarks did great help in *fear* and *disgust* emotions, the scores from experiment 3 in those two emotions are obviously higher than scores from other experiments. feature extraction with facial landmarks shown better results in "more expressive" emotions. Since the dataset in those two emotion labels are smaller than others, even though the landmarks get great improvements, more data would lead to better results in the future.

4.2.2 Qualitative Evaluation

Based on the quantitative evaluations in different experiments, we could draw a conclusion that additional facial landmarks provide improvements of classification accuracy, however, the Gaussian smooth with zero center normalization is not the best choice in this project. Therefore, the quantitative evaluation would be based on the results from experiment 3. When observing the test set and live-stream of predictions being returned by the network, When portraying extreme expressions of *happy* or *sadness*, the model correctly classified them with near perfect accuracy. However, the model doesn't have good performance when predicting the expressions such as *anger*, *fear*, and *disgust*. The Figure 6 shows some samples in the test set, and the Figure 7 collects some screenshots about predictions in real-time.

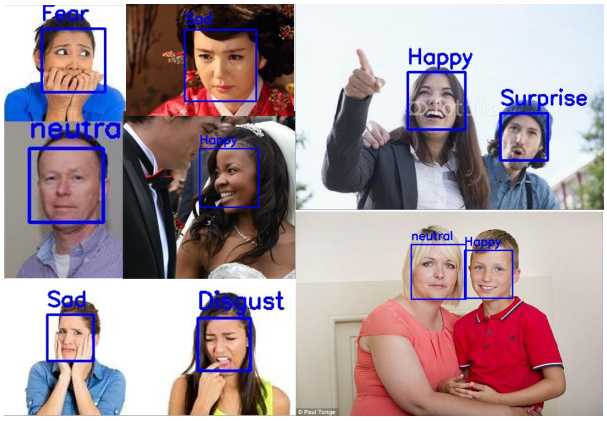

Figure 6. Sample Results from Test Set.

From the Figure 6 samples we could easily see that using the convolutional neural network with additional facial landmarks has good performance on facial expression recognition tasks. In addition, the proposed FER system works with multiple faces, different illumination environments, facial directions and different face poses.

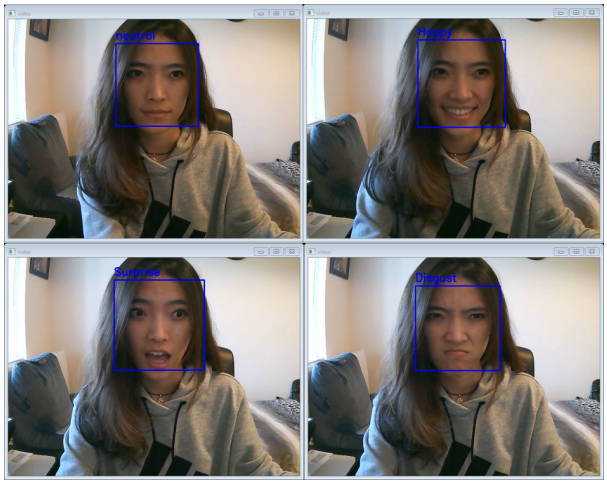

Figure 7. Sample Results from Live-stream Video.

The Figure 7 shows the random screenshots from real-time video. The network could easily capture the expressions such as *natural*, *happy* and *surprise*, and display the label on video immediately.

5 Discussion

In this project, a CNN-based emotion detection model is proposed that utilizes facial-detection software to accomplish its task. The final model resulted in f1-score reaching 0.74 on the testing dataset. In addition, the model also exhibits more balanced accuracy results across the emotion spectrum. According to the evaluation of experiments, the VGG-like network is working well on facial expression recognition problems, and class-based data augmentation decreased the influence of data imbalancing issues. Before I implemented facial landmarks extraction, I proposed using three channels images as input, the three channels inputs are converted to ycbcr color space, and applied illumination normalization which is proposed by [14]. However, the training f1-score was still lower than 0.5 even after 30 epochs. Moreover, three-channels inputs cost very expensive computation. Finally, I decided to abandon this method.

After trying three-channels inputs, I found a paper using CNN with facial landmarks [9] for expression recognition. So I tried this in my project, and got some improvements compared to previous experiments. In my last experiment, I examined a new pre-processing method which is proposed by [13], In this work. they applied a Gaussian filter to the images, and subtracted the mean-image of the training set from each image, then utilized pre-trained versions of AlexNet and LeNet in Caffe, where they re-trained the first and last layer. Unfortunately this method could not get better results in VGG network.

Limitations In this project, the proposed method recognizes the facial expressions accurately in most situations. But there are still some limitations. First, it needs a lot of pre-processing work, such as a pre-trained model for face alignment and facial landmarks extraction. A lot of pre-processing work would affect running speed in live-stream video recognition. Second, the performance in some similar and data-limited labels still needs to be improved. The Figure 8 shows some false labels in the test set. The system could not predict some labels perfectly, such as *disgust*, *sad*, or *anger*. The Figure 9 shows failure cases from live-stream video recognition. Sometimes the model predicts different emotion labels from the same expression.

Future Work One future area of work is to create a user interface where users can iteratively train the model through correcting false labels. This way the model can also learn more from real world users who express various emotions in different ways. In addition, including a layer in the network that accounts for class imbalance could provide additional improvements over the results.

Another area of interest to explore is predicting on a continuous scale the intensity of emotions being portrayed. Since the CNN already have a reliable recognition ability, adding some Recurrent Neural Network (RNN) layer such as Long Short-

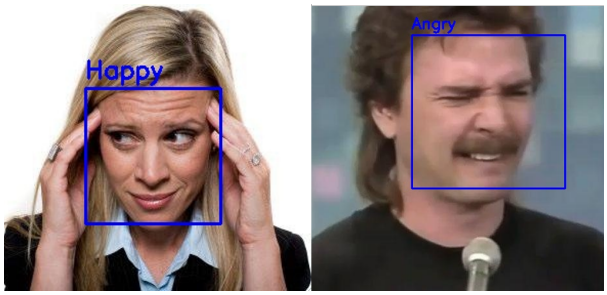

Figure 8. Failure Cases form Test Set.

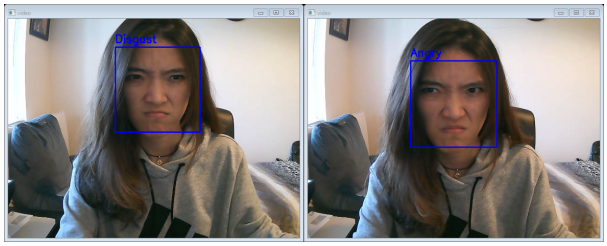

Figure 9. Failure Cases from Live-stream Video.

term Memory (LSTM) using the extracted facial features and facial landmarks between the current frame and previous frame as temporal features, a more powerful predictor for live-stream recognition tasks could be built.

References

[1] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "From facial expression recognition to interpersonal relation prediction," *CoRR*, vol. abs/1609.06426, 2016.

[2] S. Ren, X. Cao, Y. Wei, and J. Sun, "Face alignment at 3000 fps via regressing local binary features," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1685–1692, 2014.

[3] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion," *Journal of personality and social psychology*, vol. 17, 1971.

[4] P. Ekman, "Strong evidence for universals in facial expressions: A reply to russell's mistaken critique," *Psychological Bulletin*, vol. 115(2), 1994.

[5] Z. Yu and C. Zhang, "Image based static facial expression recognition with multiple deep network learning," IEEE - Institute of Electrical and Electronics Engineers, November 2015.

[6] D. Ghimire and J. Lee, "Geometric feature-based facial expression recognition in image sequences using multi-class adaboost and support vector machines," *Sensors (Basel, Switzerland)*, vol. 13, pp. 7714 – 7734, 2013.

[7] D. Ghimire, S. Jeong, J. Lee, and S. H. Park, "Facial expression recognition based on local region specific features and support vector machines," *CoRR*, vol. abs/1604.04337, 2016.

[8] M. Suk and B. Prabhakaran, "Real-time mobile facial expression recognition system – a case study," *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 132–137, 2014.

[9] F. Khan, "Facial expression recognition using facial landmark detection and feature extraction via neural networks," 2018.

[10] Y. Wu and Q. Ji, "Facial landmark detection: a literature survey," *CoRR*, vol. abs/1805.05563, 2018.

[11] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, vol. 1, pp. I–I, 2001.

[12] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[13] M. Özdemir, B. Elagöz, A. Alaybeyoglu, R. Sadighzadeh, and A. Akan, "Real time emotion recognition from facial expressions using cnn architecture," 10 2019.

[14] A. S. Sebyakin and A. V. Zolotaryuk, "Tracking emotional state of a person with artificial intelligence methods and its application to customer services," in *2019 Twelfth International Conference "Management of large-scale system development" (MLSD)*, pp. 1–5, 2019.