Facial Expression Recognition with Convolutional Neural Networks

Shekhar Singh
Computer Science
University of Nevada Las Vegas
Las Vegas, USA
shekhar.singh@unlv.edu

Fatma Nasoz
Computer Science
University of Nevada Las Vegas
Las Vegas, USA
fatma.nasoz@unlv.edu

Abstract— **Emotions are a powerful tool in communication and one way that humans show their emotions is through their facial expressions. One of the challenging and powerful tasks in social communications is facial expression recognition, as in non-verbal communication, facial expressions are key. In the field of Artificial Intelligence, Facial Expression Recognition (FER) is an active research area, with several recent studies using Convolutional Neural Networks (CNNs). In this paper, we demonstrate the classification of FER based on static images, using CNNs, without requiring any pre-processing or feature extraction tasks. The paper also illustrates techniques to improve future accuracy in this area by using pre-processing, which includes face detection and illumination correction. Feature extraction is used to extract the most prominent parts of the face, including the jaw, mouth, eyes, nose, and eyebrows. Furthermore, we also discuss the literature review and present our CNN architecture, and the challenges of using max-pooling and dropout, which eventually aided in better performance. We obtained a test accuracy of 61.7% on FER2013 in a seven-classes classification task compared to 75.2% in state-of-the-art classification.**

Keywords—Facial Expression Recognition (FER), Convolutional Neural Networks (CNNs), Artificial Intelligence (AI), Facial Action Coding System (FACS), Pre-processing, Feature Extraction.

I. INTRODUCTION

Facial expressions are essential to human social communication, as this communication is both verbal and non-verbal. Facial expressions are one aspect of non-verbal communication, as the face expresses prominent signals of communication, which includes eye contact. Other aspects of non-verbal communication are gestures and body language. It is easy for humans to notice and understand faces and facial expressions. However, it still proves difficult to develop an automated system that accomplishes the same understanding. There are several problems related to this issue, such as the detection of an image segment as an actual face, due to occlusions or illumination, as well as variations in head poses, extraction of facial expression information, facial landmark detection, or classification of expression [1].

Facial Expression Recognition (FER) is an active research area in the field of Artificial Intelligence and applied in vast domains, such as security, monitoring and law enforcement, marketing and entertainment, e-learning and medicine, emotionally intelligent robotic interfaces, or social humanoid robots. Various fields, like data analytics, psychological research, social gaming, and others that include human-computer interactions, can benefit from the ability to recognize facial expressions automatically.

It has been determined that certain facial expressions have universal meaning. By 1978, Ekman and Friesen had finalized and developed the Facial Action Coding System (FACS), and the authors found that there are six facial expressions, including happiness, sadness, surprise, fear, anger, and disgust, that appear to be universal across all cultures [2]. Research challenges such as Kaggle's Facial Expression Recognition challenge present these same emotions for classification, along with the addition of a seventh emotion, which is the neutral emotion. However, it is challenging for computers to recognize these common human expressions in natural conditions because of differences in lighting various head poses.

A recent survey paper [3] shows the success of CNNs using the FER2013 dataset. We will discuss more the background and literature in the related work section. The objective of this paper is to develop a novel architecture, from scratch, to classify images of human faces into discrete emotion categories using CNNs, also illustrate the pre-processing and feature extraction techniques to further improve the accuracy on the FER2013 dataset.

II. RELATED WORK

Guillaume-Benjamin-Amand Duchenne de Boulogne was a French neurologist in the 19th century, who was interested in Physiognomy and wanted to understand how human face muscles work to produce facial expressions, as he believed that these were directly linked to a human's soul. To do this, he used electric probes to trigger muscle contractions, and then took pictures, using newly developed camera technology, of his subjects' faces showing the distorted expressions he was able to create. In 1862, he published his research and the photographs of the triggered facial expressions in the book "The Mechanism of Human Physiognomy " [4]. An example from his publication can be seen in Fig. 1, showing photographs of his subjects displaying a different expression on each side of their faces.

Later, in 1872, Charles Darwin used this work as an important resource [5] for his book called "The Expression of Emotion in Man and Animals" in which he focused on the genetics of behavior. However, in recent years, Duchenne de Boulogne's book has been rediscovered by photographers as

978-1-7281-3783-4/20/$31.00 ©2020 IEEE

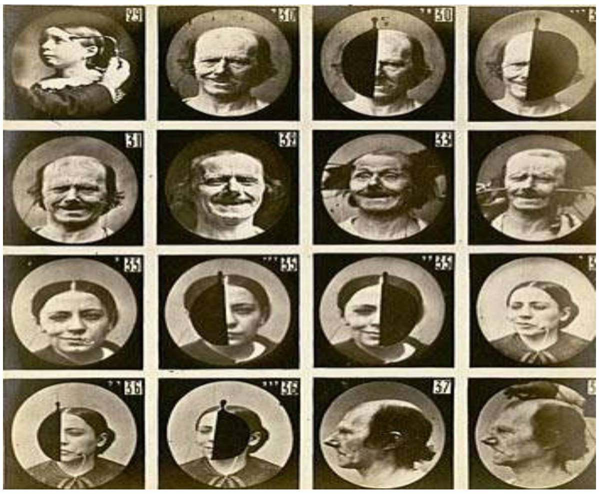


Fig. 1. The Mechanism of Human Physiognomy.

an important work of photographic art history. However, unquestionably, Ekman is the most influential researchers in the field of emotional expression of this century, as discussed in the introduction.

In 2016, Pramerdorfer and Kampel obtained state-of-the-art, which is 75.2% accuracy on the FER2013, using Convolutional Neural Networks (CNNs) [6]. The authors used an ensemble of CNNs using VGG, Inception, and ResNet with depths of 10, 16, and 33, with parameters of 1.8m, 1.6m, and 5.3m, respectively. The authors used the face images as given in the dataset, and for illumination correction, they used histogram equalization. They performed horizontal mirroring for training data augmentation and randomly cropped images to the size of 48 x 48 pixels. They also trained the architecture for up to 300 epochs and used stochastic gradient descent to optimize the cross-entropy loss, with a momentum value 0.9. The other parameters were fixed, like learning rate with 0.1, batch size with 128, and weight decay with 0.0001.

Zhang et al. [7] used a Siamese Network to introduce a method for understanding social relation behaviors from images and achieved a test accuracy of 75.1% on the challenging Kaggle facial expression dataset. The authors used multiple datasets, with various labels, to increase the training data; they also introduced a feature extraction method and patch-based registration, as well as working on feature integration via early fusion.

Kim et al. [8] proposed an ensemble of CNNs and demonstrated that during training and testing it is advantageous to use both registered and unregistered forms of given face images. The authors achieved a test accuracy of 73.73% on the FER2013 dataset. They also conducted Intraface for a conventional 2-D alignment, which is publicly available for landmark detector, and performed illumination normalization. To avoid the registration error, they performed registration selectively, based on the results of facial landmark detection.

Raghuvanshi and Choksi [9] experimented with different architectures using Deep CNNs and methods such as fractional max-pooling and fine-tuning, ultimately achieving an accuracy of 48% on Kaggle's Facial Expression Recognition Challenge dataset.

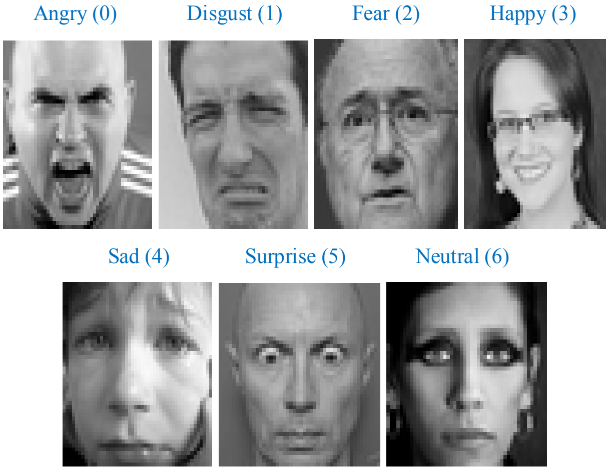


Fig. 2. Example images from the FER2013 dataset with labels.

III. DATASET

The FER2013 [10] dataset was presented at the International Conference on Machine Learning (ICML) 2013 Workshop on Challenges in Representation Learning. FER2013 is a large dataset, which is publicly accessible on Kaggle's FER Challenge. The FER2013 dataset contains 35,887 face crops, including training, validation and testing images, with 28,709, 3,589 and 3,589, respectively. Fig. 2 illustrates the images of seven different emotions with their corresponding labels (0= anger, 1= disgust, 2= fear, 3= happiness, 4= sadness, 5= surprise and 6= neutral). All images are of 48x48 pixel resolutions and on grayscale. Ian Goodfellow determined the human accuracy of this dataset to be around 65.5% [10].

IV. TECHNICAL WORK

In this section, we will discuss our CNNs architecture and techniques to further improve the accuracy on the FER2013. The work is split into three different components, as follows:

A. Pre-processing

Pre-processing can be used to enhance FER system performance and can be done previous to the feature extraction process. Image pre-processing includes various processes, such as the detection and alignment of faces, correction of illumination, pose, occlusion, and data augmentation.

In the FER2013 dataset, the faces are registered automatically, so that they have similar space requirements and are more or less centered in the images. Fig. 3 is an illustration of face detection using the Haar Cascade classifier [11].

This study experiments with face detection to capture most of the face and implement various techniques on the face

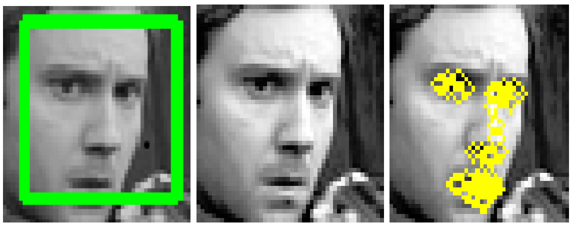


Fig. 3. Illustration of face detection (green square), the middle image shows illumination correction and the last image shows features extraction of the right eye, left eye, nose and mouth (yellow color).

image in order to improve accuracy in the future. We are further enhancing the face detection method to make a more robust algorithm that can address the occlusion, illumination, and head pose issues.

When images are captured in various types of light, expression features are sometimes inaccurately detected, and therefore, the expression recognition rate can be low and make feature extraction more difficult. In Fig. 3, the middle image is an illustration of illumination correction done by histogram equalization.

B. Feature Extraction

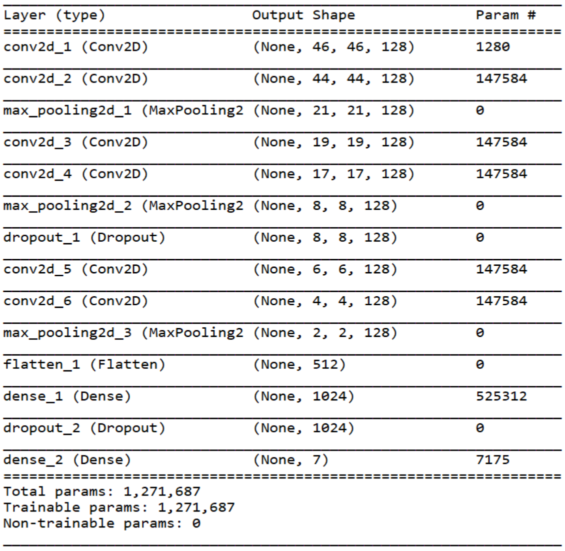
The extraction of facial features requires translating the input data into a set of features. By using feature extraction, researchers can reduce an immense amount of data down to a relatively small set, which allows for faster computation. We applied dlib facial landmark detector pre-trained on iBUG 300-W dataset [12], [13], [14] for feature extraction, and extracted the eight most prominent parts of a face, including both eyebrows, both eyes, the nose, the inner and outer outlines of the mouth, and the jaw. In Fig. 3, the last image is an illustration of feature extraction, in which we extracted the right and left eyes, nose, and inner and outer outline of the mouth, which are marked with yellow color.

C. CNN Architecture

CNNs have been widely used in a variety of computer vision applications, including FER. Early in the 21st century, several studies of FER literature [15], [16] determined that CNNs work well on changes of face location, as well as variations in scale. They were also found to work better than multilayer perceptron (MLP) when looking at face pose variations not seen previously. Researchers used CNN to help solve various facial expression recognition problems, such as translation, rotation, subject independence, and scale invariance [17].

Fig. 4 shows that finding an optimized architecture or network structure is a very challenging task. Our model was trained using the following characteristics:

- six convolutional layers using "RELU" as an activation function;
- three max-pooling: out of which the first two using pool size (3,3) and stride (2,2), and the third using pool size (2,2) and stride (2,2); every max pooling is followed by every two convolutional layers;



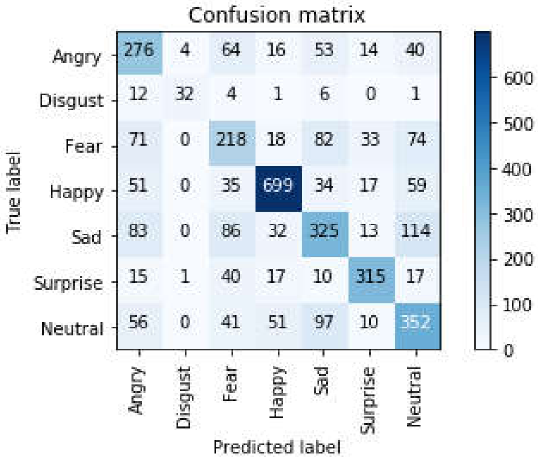
```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d_1 (Conv2D)            (None, 46, 46, 128)       1280
_________________________________________________________________
conv2d_2 (Conv2D)            (None, 44, 44, 128)       147584
_________________________________________________________________
max_pooling2d_1 (MaxPooling2 (None, 21, 21, 128)       0
_________________________________________________________________
conv2d_3 (Conv2D)            (None, 19, 19, 128)       147584
_________________________________________________________________
conv2d_4 (Conv2D)            (None, 17, 17, 128)       147584
_________________________________________________________________
max_pooling2d_2 (MaxPooling2 (None, 8, 8, 128)         0
_________________________________________________________________
dropout_1 (Dropout)          (None, 8, 8, 128)         0
_________________________________________________________________
conv2d_5 (Conv2D)            (None, 6, 6, 128)         147584
_________________________________________________________________
conv2d_6 (Conv2D)            (None, 4, 4, 128)         147584
_________________________________________________________________
max_pooling2d_3 (MaxPooling2 (None, 2, 2, 128)         0
_________________________________________________________________
flatten_1 (Flatten)          (None, 512)               0
_________________________________________________________________
dense_1 (Dense)              (None, 1024)              525312
_________________________________________________________________
dropout_2 (Dropout)          (None, 1024)              0
_________________________________________________________________
dense_2 (Dense)              (None, 7)                 7175
=================================================================
Total params: 1,271,687
Trainable params: 1,271,687
Non-trainable params: 0
_________________________________________________________________
```

Fig. 4. Illustration of CNN Architecture.

- two drop out with value 0.2;
- one flattened layer and two dense layers: one dense layer with "RELU," and the other with "Softmax" as an activation function;
- total parameters and trainable parameters are 1.2 million, respectively.

V. CONCLUSION

This study's six CNNs layer architecture performed competitively and achieved a FER2013 test accuracy of 61.7%, without involving any pre-processing or feature extraction techniques. The state-of-the-art test accuracy for the seven emotion categories using the ensemble of CNNs was 75.2%. We performed several experiments using different batch sizes and epochs but obtained the best test accuracy using a batch size value of 512 and 10 epochs.



	Angry	Disgust	Fear	Happy	Sad	Surprise	Neutral
Angry	276	4	64	16	53	14	40
Disgust	12	32	4	1	6	0	1
Fear	71	0	218	18	82	33	74
Happy	51	0	35	699	34	17	59
Sad	83	0	86	32	325	13	114
Surprise	15	1	40	17	10	315	17
Neutral	56	0	41	51	97	10	352

Fig. 5. Confusion matrix on the FER2013 test dataset.

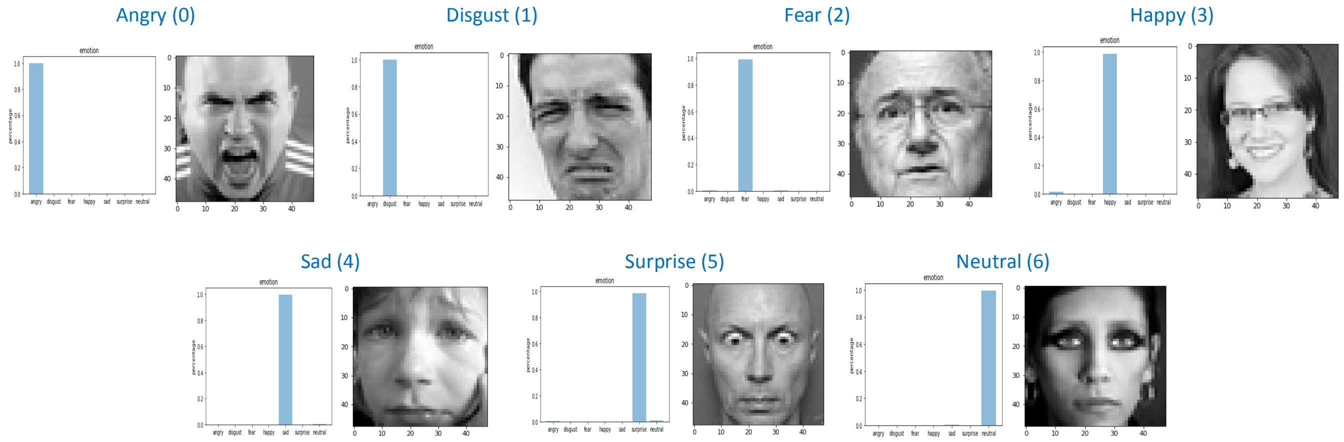


Fig. 6. Illustration of images with correctly predicted respective emotions.

All the classifiers were implemented in Keras, using Tensorflow as a backend, and trained on a personal laptop configuration of CPU I7 64 bit, with 32 GB RAM and dedicated 8 GB NVIDIA GPU GTX 1080, and best model and weights were saved in HDF5 format.

Fig. 5 illustrates the confusion matrix on the test dataset, where rows represent true labels and columns represent the predicted label. There are 467 “angry” instances in the test set, and we correctly predicted 276 “angry” instances. In another example, there are 895 ”happy” instances, and we correctly predicted 699 of those instances.

Fig. 6 shows correctly predicted emotions from the training dataset referred to in Fig. 2. On the emotion prediction bar chart, the y-axis represents percentage and the x-axis represents emotions. The prediction bar shows the percentage of respective emotions.

We analyzed the confusion matrix, from which the most misclassified images came from the emotions of fear and sad with 43.95% and 49.77% accuracy, respectively. Fig. 7 shows the misclassified images with their actual and predicted labels. As an example, one can consider the third image, whose actual label is ‘sad,’ but the model predicts fear, sad, and angry with 43%, 40%, and 15% of accuracy, respectively, as well as disgust, surprise, and neutral with 2% accuracy altogether.

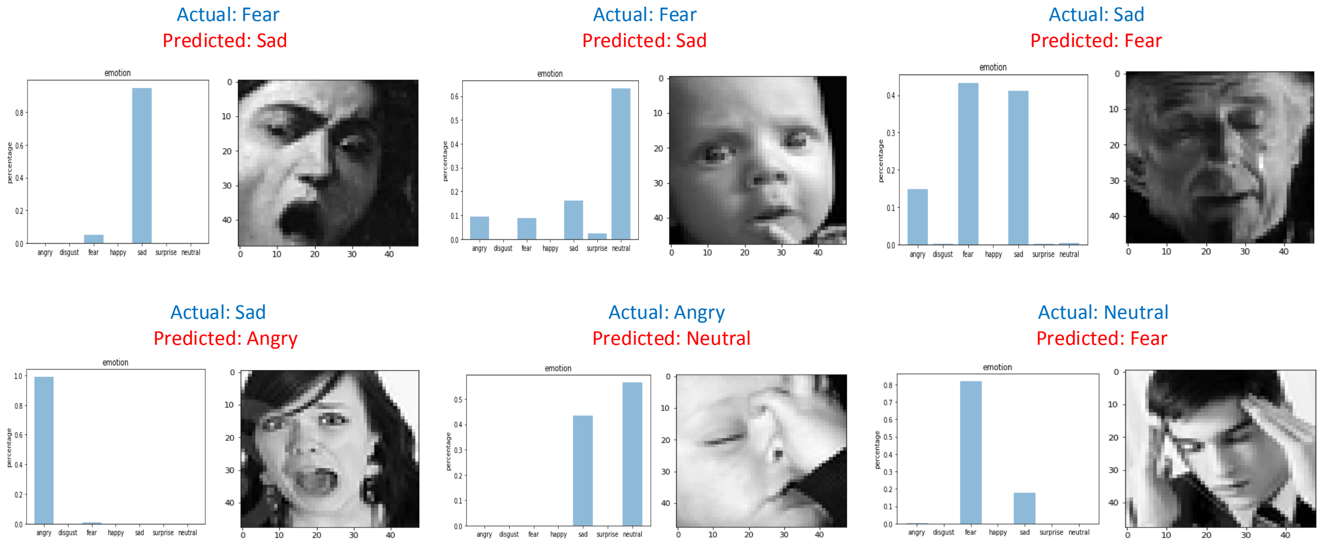


Fig. 7. Illustration of misclassified images on the FER2013 test dataset.

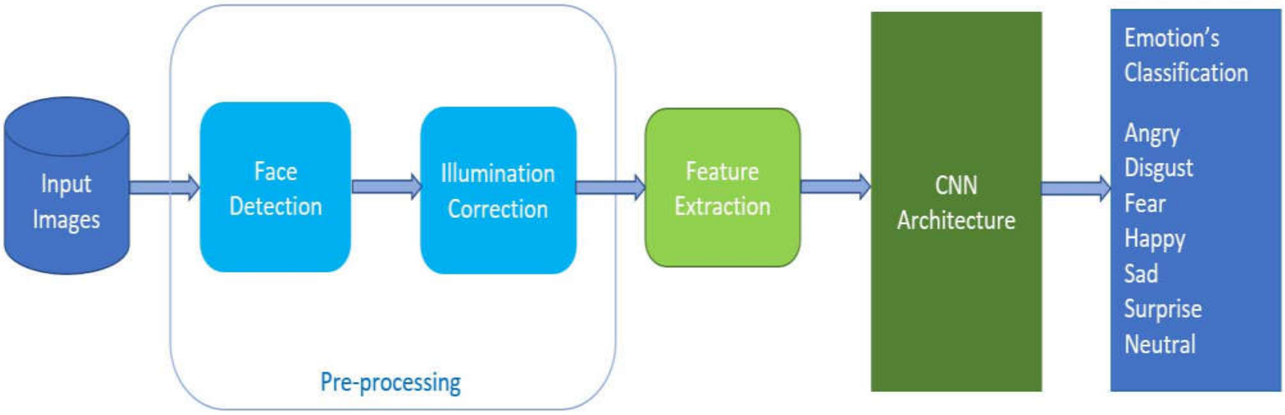


Fig. 8. Future work pipeline of FER.

VI. DISCUSSION AND FUTURE WORK

Our individual six CNNs layer architecture performed well and achieve a FER2013 test accuracy of 61.7% in a seven-classes classification task without requiring any pre-processing and feature extraction techniques.

In this section, we discuss the challenges and future work that could be done to further improve the test accuracy on the FER2013 dataset. Finding the best network architecture is a very challenging issue in deep learning. We used a heuristic approach to find CNNs and will work on finding a more robust network in the future, as noted in Fig. 8. We will also include pre-processing and feature extraction techniques, discussed in the technical work section, in order to achieve better accuracy. Additionally, we faced an overfitting issue due to 99.64% accuracy on the training dataset. Therefore, data augmentation is a vital step for deep FER. Usually, data augmentation is embedded in a deep learning toolkit to alleviate the overfitting issue. We believe that using the above-discussed techniques will allow for further improvement of test accuracy, with expectations of getting closer to the current state-of-the-art.

REFERENCES

[1] Application: Facial Expression Recognition. In: Machine Learning in Computer Vision. Computational Imaging and Vision, vol 29. Springer, Dordrecht (2005).

[2] https://en.wikipedia.org/wiki/Paul_Ekman, Jan 2020.

[3] Deep FER: A Survey https://arxiv.org/pdf/1804.08348.pdf, Jan 2020.

[4] Duchenne, G.-B. (1862), Mécanisme de la physionomie humaine, ou analyse électro-physiologique de ses différents modes de l'expression. Paris: Archives générales de médecine, P. Asselin; vol. 1, p. 29-47, 152-174.

[5] Darwin, Charles Robert (1872), The expression of the emotions in man and animals. London: John Murray London.

[6] C. Pramerdorfer and M. Kampel, "Facial expression recognition using convolutional neural networks: State of the art," arXiv preprint arXiv:1612.02903, 2016.

[7] Z. Zhang, P. Luo, C.-C. Loy, and X. Tang, "Learning Social Relation Traits from Face Images," in Proc. IEEE Int. Conference on Computer Vision (ICCV), 2015, pp. 3631–3639.

[8] B.-K. Kim, S.-Y. Dong, J. Roh, G. Kim, and S.-Y. Lee, "Fusing aligned and non-aligned face information for automatic affect recognition in the wild: A deep learning approach," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2016, pp. 48–57.

[9] Raghuvanshi, A., & Choksi, V. (2016). Facial Expression Recognition with Convolutional Neural Networks.

[10] "Challenges in Representation Learning: A report on three machine learning contests." I Goodfellow, D Erhan, PL Carrier, A Courville, M Mirza, B Hamner, W Cukierski, Y Tang, DH Lee, Y Zhou, C Ramaiah, F Feng, R Li, X Wang, D Athanasakis, J Shawe-Taylor, M Milakov, J Park, R Ionescu, M Popescu, C Grozea, J Bergstra, J Xie, L Romaszko, B Xu, Z Chuang, and Y. Bengio. arXiv 2013.

[11] Viola, P. and Jones, M. Rapid object detection using boosted cascade of simple features. IEEE Conference on Computer Vision and Pattern Recognition, 2001.

[12] C. Sagonas, E. Antonakos, G, Tzimiropoulos, S. Zafeiriou, M. Pantic. 300 faces In-the-wild challenge: Database and results. Image and Vision Computing (IMAVIS), Special Issue on Facial Landmark Localisation "In-The-Wild". 2016.

[13] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, M. Pantic. A semi-automatic methodology for facial landmark annotation. Proceedings of IEEE Int'l Conf. Computer Vision and Pattern Recognition (CVPR-W), 5th Workshop on Analysis and Modeling of Faces and Gestures (AMFG 2013). Oregon, USA, June 2013.

[14] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, M. Pantic. 300 Faces in-the-Wild Challenge: The first facial landmark localization Challenge. Proceedings of IEEE Int'l Conf. on Computer Vision (ICCV-W), 300 Faces in-the-Wild Challenge (300-W). Sydney, Australia, December 2013.

[15] B. Fasel, "Robust face analysis using convolutional neural networks," in Pattern Recognition, 2002. Proceedings. 16th International Conference on, vol. 2. IEEE, 2002, pp. 40–43.

[16] F. Beat, Head-pose invariant facial expression recognition using convolutional neural networks, in: Fourth IEEE International Conference on Multimodal Interfaces, 2002. Proceedings, 2002, pp. 529–534.

[17] M. Matsugu, K. Mori, Y. Mitari, and Y. Kaneda, "Subject independent facial expression recognition with robust face detection using a convolutional neural network," Neural Networks, vol. 16, no. 5-6, pp. 555–559, 2003.