Facial Expression Recognition through Machine Learning

Nikhil Jawade
Queen Mary University of London,
School of Electronic Engineering and Computer Science
n.jawade@se21.qmul.ac.uk

1. Introduction

Of all creatures, the human face is the most intricate and varied. One of the most potent, inherent, and common ways humans express their emotions and intentions is through facial expressions [1] [16]. Automatic facial expression analysis has been the subject of several studies due to its practical significance in sociable robotics, medical treatment, driver tiredness detection, and many other human-computer interface systems [10]. Different facial expression recognition (FER) techniques have been investigated to encode expression information from facial representations in the fields of computer vision and machine learning [10]. This paper attempts to examine the domain of FER in practical settings by discussing and analyzing several research works, and presenting hypotheses and their restrictions. Further, as a part of the task, data preparation and pre-processing techniques are performed on a subset of the Aff-Wild2 in-the-wild dataset as well as the results of these techniques are summarized.

2. Related Work

For a significant time, the study conducted in the facial expression recognition (FER) domain was on the data in a laboratory-controlled environment. Over time, the transition in FER has taken place from laboratory-controlled to more challenging in-the-wild conditions. Not only the data condition, but the transition has happened over the tasks in FER from category classification of six basic human emotions (i.e., happiness, surprise, disgust, fear, sadness, anger [2]) to multi-modal affect recognition. Conducting a literature review over an ever evolving domain is a humongous task at hand and should be done scrupulously.

The work of Shan Li and Weihong Deng [10] is a great place to start reading about the early works in the study of FER. This survey gives an overview of the evolution of datasets and techniques used in the domain. A summary of frequently used datasets like FER-2013, RAF-DB, ExpW, EmotioNet, AffectNet, etc., is presented containing the number of samples each of these dataset contains, expression distribution, and whether images are posed or spontaneous. A general deep FER system pipeline is represented which gives a newcomer an idea of the systematic framework that is followed in deep FER implementation. It further discusses the deep learning techniques apropos to the FER networks for static images and dynamic image sequences. The take from this survey was where the FER study has headed over the years and how complex it is to deal with different stems of FER.

FER as a study has a broad perspective. Hence further literature review in this paper is narrowed down to single task categorization of basic facial expressions. Each piece of literature is discussed on its merits and evaluated based on factors such as selection of dataset and its description, selection of deep learning technique and evaluation criteria chosen.

The early work of Yichuan Tang [15] on deep learning using Support Vector Machines (DLSVM) showed promising results in the classification of basic expressions in a relatively bigger dataset of 28,709 images of size 48x48. Their implementation of Convolutional Neural Network (CNN) with linear one-vs-all SVM achieved an accuracy of 71.2%. As this work was more focused on the results provided by DLSVM than the problem statement of FER, it failed to mention the distribution report of data based on expressions in each train, test and validation. The evaluation was based solely on the accuracy, which doesn't give a thorough outlook on the results.

In the study by Yu and Zhang [17], the CNN proposed contains five convolutional layers, three stochastic pooling layers and three fully connected layers. An interesting method of generating randomized perturbation was performed on the data before fetching it to the model. The proposed CNN were pre-trained over the FER-2013 dataset and fine-tuned over the SFEW dataset and accuracy evolution was presented over the number of epochs for both datasets. They were successful in building a much robust model but returned a mediocre accuracy.

Further, in a substantial review by Pramerdorfer and Kampel of existing CNN-based FER methods by highlight-

ing the differences between their architecture and performance impact in their study [13]. The results on CNN architectures from two studies mentioned above and another four from studies [5] [11] [18] [4] are summarized and highlighted them as one of the bottlenecks (comparatively shallow and basic) that would lead to improved performances. Three modern deep CNNs: VGG, Inception and ResNet were put forth and their accuracies were presented. Results showed that the best modern deep CNN outperforms the best shallow model by almost 2% under identical conditions. Even though they represented a fairly good observation, they failed to mention how to overcome the bottleneck which majorly hinders the FER performance that is the lack of publicly available dataset.

The focus of previous research has mostly been on identifying the six universal expressions, which are intuitive and basic, and are unable to convey complicated affective states [8]. Such implementation will fail in real-life scenarios where emotions are continuous in a visual queue. This progression happened in the study [6], in which a large-scale in-the-wild data Aff-Wild was utilized for estimation of continuous emotion dimensions based on visual cues using deep neural architecture. Initially, the then-novel Aff-Wild database is discussed being the one of a kind to be annotated in terms of valence-arousal. The database of 1.18M frames is split into 900K for training and 300K for testing. The images are then resized to 224x224 for RGB channels. In order to get the 2-D prediction of valence and arousal, three deep neural network architectures were evaluated. Two sets of models are selected: for CNN architectures - ResNet50, VGG-Face and VGG-16, and for CNN-RNN architecture - changes in RNN either LSTM or GRU. The loss functions for evaluation and training included the Concordance Correlation Coefficient (CCC) and Mean Squared Error (MSE). More weightage was given to optimizing CCC which evaluates the agreement between two time series by scaling their correlation coefficient with their mean square difference [9]. The best results obtained out of three only-CNN architectures were using the VGG-Face net while the architecture containing the GRU neuron model was the best out of two CNN-RNN architectures. A comprehensive study with a novel dataset has shown promising results and laid a path for extending the analysis to simultaneously interpret the behaviour of multiple objects appearing in a video.

3. Hypotheses-Restrictions

Automatic facial expression recognition is a study in the ambiguous area where humans themselves have a different perspective for a single facial expression. Throughout the literature review conducted in this paper, it is followed that the FER study has evolved into three major segments: the task at hand, the database used and deep neural architectures. In this section, all these three segments are discussed, and then restrictions faced in each are discussed.

The task at hand began with category classification of six basic human emotions (i.e., happiness, surprise, disgust, fear, sadness, anger [2]) and later evolved to multi-modal affect recognition. While humans are capable of a significant amount of emotions, and their facial expressions vary for all of them, just classifying six basic expressions seemed quite a restricted task. Additionally, all the past work over this task was only on static frames (images) making it a failure to a continuous detection of expressions. Hence to overcome this, in recent works, the task at hand focused more on the affect recognition through valence-arousal values. Even this task has its limitations, mainly from the perspective of databases.

Very few databases existed when the FER category classification study using deep CNNs began. But with time more databases started to roll out but most of them were annotated by just the six basic expressions and neutral and action units. Majors issues with such databases were related to the images as most of them were captured in a more controlled environment. These images lacked variety of subjects of different age group and ethnicities, head poses, occlusions and illumination conditions. In attempt to resolve this, database like Aff-Wild2 was created and was annotated with valence-arousal values [8].

As the FER learning progressed towards affect recognition, the algorithms used in them got complex and evolved from shallower CNNs to more deeper ones. Deep CNNs used in recent works for affect recognition contains at least 9 blocks of combination of convolutional and fully-connected layers [6]. Deep neural network is ongoing study and with years passing better models trained over significantly larger dataset could help generalizing the problem task of FER.

4. Data Preparation

As a part of task 'Facial Expression Recognition through Machine Learning', a subset of Aff-Wild2 [8] in-the-wild dataset was provided. All the frames in Aff-Wild2 data are annotated by valence-arousal values, and partly annotated in terms of seven basic categorical expressions including Neutral, action units. The subset provided contains 43,435 images in total of size 128x128 of RGB channels belonging to six classes of basic expressions (i.e., happiness, surprise, disgust, fear, sadness, anger [2]). Sample images of each of the six expressions is shown in Figure 1.

The distribution of six expression is shown in the bar graph 2 with respective numbers. It could be seen from the graph that data is imbalanced. Three classes: anger, happiness and sadness being the dominant in data containing 70% of the total, while two classes: disgust and fear being the smallest in data containing only 13% of complete data.

For the purpose of data preparation, data is divided into

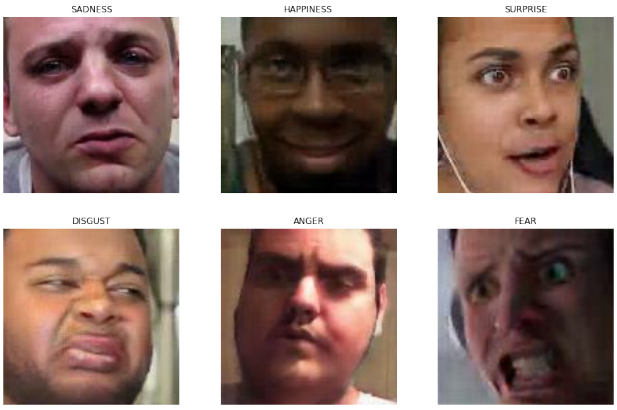


Figure 1. Images of six facial expressions

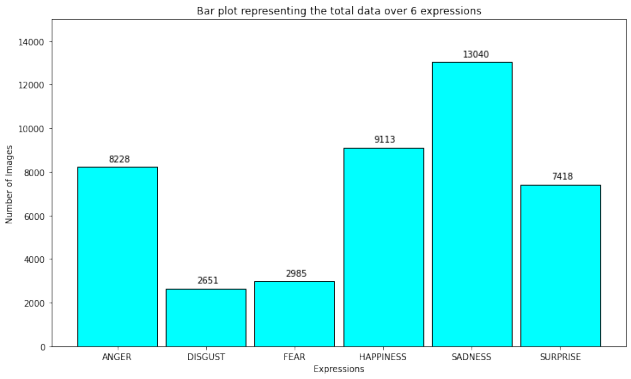


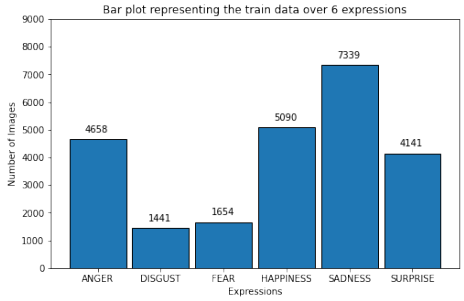
Figure 2. Test Data Distribution

three sets i.e. train, test and validation. The split is performed over the whole data of 43,435 into train and test at the percentage of 70-30 giving train data of 30,404 images and test data of 13,031. Validation data is necessary for fine-tuning of model. Hence, the train data is further split into train and validation sets at percentage of 80-20 giving train data of 24,323 images and validation data of 6,081. The distribution of expressions across all three datasets is shown in Figure 3.

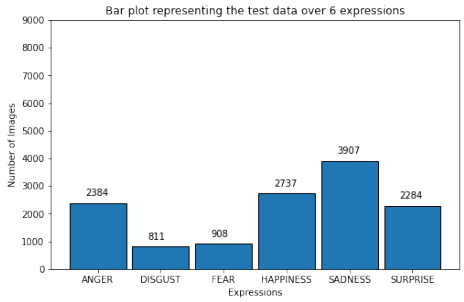
5. Data Pre-processing

Preparation of the Aff-Wild2 dataset is carried out by splitting all videos into images. The SSH detector [12] based on Resnet [3] was used to extract face bounding boxes from the images and later 5 facial landmarks (two eyes, nose and two mouth corners) were extracted to perform similarity transformation for purpose of face alignment [7]. Meaning that the subset of Aff-Wild2 dataset received for this study is well pre-processed data.

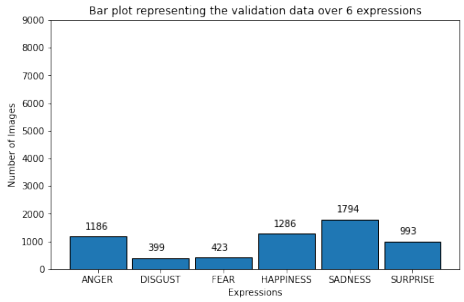
The significance of pre-processing after this is to train a CNN architecture making it more robust and general purpose in order to tackle real-world situations. This section mentions various image pre-processing techniques to achieve this.



(a) Train Data Distribution



(b) Test Data Distribution



(c) Validation Data Distribution

Figure 3. Distribution of dataset over Train, Test and Validation Sets

The first and one of the most important step to be followed in the pre-processing pipeline is to normalize the image pixel values sole purpose being scaling data in a range of [0, 1] for computational ease. Another step ensuing it is resizing. When fetching the image data to a CNN model, resizing images to a lower dimension (e.g. 32x32) helps keep up to the compute limitations. Hence it becomes very important to resize all the images before feeding it to a net. One of the samples is shown in the Figure 4 in which image of size 128x128 is resized to 32x32.

Data augmentation techniques is the further steps to be followed in the preparation pipeline. Based on previous studies of FER, augmentation techniques like rotation, flip, and addition of noise and blur turned to be more powerful towards the robust model achievement. Figures 5, 6, 7 and

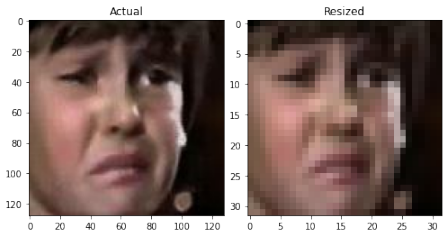


Figure 4. Resizing the image

8 shows rotation, flip and noise and blur addition to images respectively.

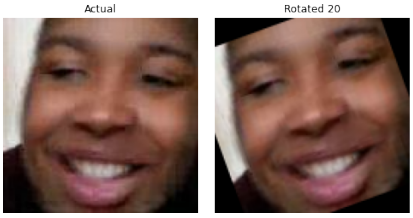


Figure 5. Rotating the image

Figure 6. Flipping the image Left-Right

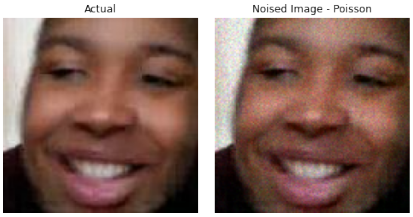


Figure 7. Adding the Poisson noise to the image

It is worth to mention data augmentation techniques used in study by Psaroudakis and Kollias [14]. An innovative approach of Mixup and MixAugment techniques for data pre-processing showed promising results on the data having significant variations in head poses, illumination condition, backgrounds and contexts. The mixup technique sample is shown in Figure 9.

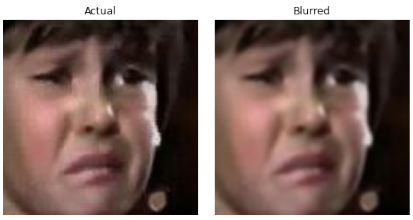


Figure 8. Blurring the image

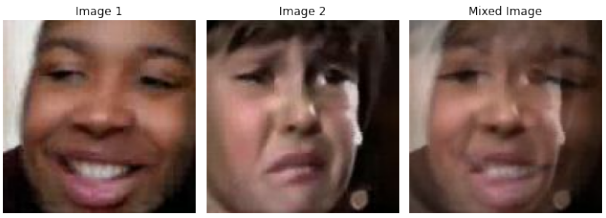


Figure 9. Mixing two image in the ratio of 50-50

References

[1] Charles Darwin and Phillip Prodger. *The expression of the emotions in man and animals*. Oxford University Press, USA, 1998. 1

[2] Paul Ekman. Basic emotions. *Handbook of cognition and emotion*, 98(45-60):16, 1999. 1, 2

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. 3

[4] Bo-Kyeong Kim, Suh-Yeon Dong, Jihyeon Roh, Geonmin Kim, and Soo-Young 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*, pages 48–57, 2016. 2

[5] Bo-Kyeong Kim, Jihyeon Roh, Suh-Yeon Dong, and Soo-Young Lee. Hierarchical committee of deep convolutional neural networks for robust facial expression recognition. *Journal on Multimodal User Interfaces*, 10:173–189, 2015. 2

[6] Dimitrios Kollias, Mihalis A Nicolaou, Irene Kotsia, Guoying Zhao, and Stefanos Zafeiriou. Recognition of affect in the wild using deep neural networks. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*, pages 1972–1979. IEEE, 2017. 2

[7] D Kollias, A Schulc, E Hajiyev, and S Zafeiriou. Analysing affective behavior in the first abaw 2020 competition. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)(FG)*, pages 794–800. 3

[8] Dimitrios Kollias and Stefanos Zafeiriou. Aff-wild2: Extending the aff-wild database for affect recognition. *CoRR*, abs/1811.07770, 2018. 2

[9] Dimitrios Kollias and Stefanos Zafeiriou. Affect analysis in-the-wild: Valence-arousal, expressions, action units and a

unified framework. *arXiv preprint arXiv:2103.15792*, 2021. 2

[10] Shan Li and Weihong Deng. Deep facial expression recognition: A survey. *IEEE transactions on affective computing*, 2020. 1

[11] Ali Mollahosseini, David Chan, and Mohammad H. Mahoor. Going deeper in facial expression recognition using deep neural networks. *CoRR*, abs/1511.04110, 2015. 2

[12] Mahyar Najibi, Pouya Samangouei, Rama Chellappa, and Larry Davis. SSH: Single stage headless face detector. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017. 3

[13] Christopher Pramerdorfer and Martin Kampel. Facial expression recognition using convolutional neural networks: State of the art. *CoRR*, abs/1612.02903, 2016. 2

[14] Andreas Psaroudakis and Dimitrios Kollias. Mixaugment & mixup: Augmentation methods for facial expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2367–2375, 2022. 4

[15] Yichuan Tang. Deep learning using linear support vector machines. *arXiv preprint arXiv:1306.0239*, 2013. 1

[16] Y-I Tian, Takeo Kanade, and Jeffrey F Cohn. Recognizing action units for facial expression analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 23(2):97–115, 2001. 1

[17] Zhiding Yu and Cha Zhang. Image based static facial expression recognition with multiple deep network learning. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*, pages 435–442, 2015. 1

[18] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Learning social relation traits from face images. *CoRR*, abs/1509.03936, 2015. 2