Facial Expression Analysis: Pre-processing Methods on modified VGG16 Network*

Ali Azmoudeh
dept. Computer and Informatic Engineering
Istanbul Technical University
Istanbul, Turkiye
azmoudeh22@itu.edu.tr

***Abstract*—Over the last multiple decades, Facial Expression Analysis(FEA) has been one of the challenges in the Computer Vision Field. By growing the CNN models, there are many Achievements in providing more accurate models. Although many works are available in this field, FEA is still challenging. According to the image processing methods, which impact the accuracy of the models in computer vision tasks. Cropping the face region, rescaling all available data, and aligning them show three necessary steps for pre-processing the face figure for Convolutional neural network models to classify the expression accurately. Instead of these steps, local binary patterns were compared and evaluated on VGG16 Network for FEA in this work.**

***Index Terms*—Computer Vision, Convolutional Neural Networks, VGG16, Face Alignment, Image Processing**

I. Introduction

Facial expressions provide a great deal of information about human emotion and are an essential type of nonverbal communication. Automatic facial expression analysis(FEA) has been an exciting and challenging research problem for the last two decades. Facial emotions are produced by facial muscle contractions, resulting in temporally distorted facial features such as eyelids, brows, nose, lips, and skin texture, typically exposed by wrinkles and bulges. It has many vital applications, such as human behavior recognition, sign language recognition, and human-computer interaction. FEA is an effective way of communication between humans to humans and human computers. In the real world, by looking at a person's facial expression, you can observe their feelings about every phenomenon they have encountered. The machine's purpose is to monitor these human emotions.

Automatic facial expression recognition systems play a crucial role in the Human-Computer Interaction community; it is recognized as a non-verbal communication method where machines can help to convey the proper message after decoding the emotion from the face. [1] Another application of these model have been used extensively in the medical area, e.g., FEA used for Autism cases [2], and or it used for recognition of disgusted facial expression in depression [3], in some cases, by considering the facial expression, the machine can detect pain intensity from facial expression images. [4]

FEA can be used to evaluate the level of satisfaction between two humans and detect boredom between two individuals. It is related to the human to human communication. In addition, by using facial expression analysis, companies can detect the level of satisfaction when consumers use their products, and they can improve their services to enhance the quality of their services.

Theoretically, an infinite number of facial expressions are possible, which leads to an infeasible problem of recognizing all the expressions in a finite memory system. Recently, Many enhancements have been achieved in this field, and many methods have been produced. The latest works achieved significant results parallel to this enhancement by growing deep neural networks. The introduction of convolutional Neural Networks(CNN) was a revolution in the field of computer vision; thus, after introducing new methods by CNN, AlexNet was one of the influential first methods in the field of CNN. Moreover, by following the AlexNet, VGG16 [8] was introduced, and it can achieve the best training model on ImageNet in 2014; these achievements always continued researchers from seeking new approaches. This approaches help many researchers and save time because CNNs are computationally costly. It can take much time each time we want to train the model from the first layer in the large Networks. In this case, by considering transfer learning, we can save time and fine-tune our models on each dataset by using pre-trained models, which is achieved by training on a famous dataset such as ImageNet [12] with 1 Million data and with 1 thousand classes.

Before utilizing the training model, preparing the data for models such as VGG16 is a vital step. If we don't consider these steps. The CNN models will be failed to give an accurate result. Indeed, it can be disappointing during training and fine-tuning. Owing to the necessity of preprocessing and various impacts within them. It is necessary to know the efficiency of these models.

In this work, by considering the VGG16 as a model for training, we compared the pre-processing methods when they were used individually and together on two distinct datasets with significant differences on their attributes; after fine-tuning the model, the data, when cropped and implemented alignment on the face figure. Performs better than other results by achieving 66.2% accuracy compared to the other pre-processing methods, such as *Local Binary Pattern* on the

FER2013 dataset. However, the impact of the LBP seems very helpful by significant accuracy of 99.2% on the JAFFE dataset.

II. Datasets

A. The Jappanese Female Facial Expression(JAFFE) Dataset

This is one pioneer dataset for emotion detection from facial behavior. Moreover, This dataset consists of 10 female subjects from Japan, and each subject expresses each emotion multiple times in different ways. This dataset has seven basic facial expression classes. Overall, this dataset has 213 figures with a resolution of 256x256. Figure 1 shows some instances from this dataset.

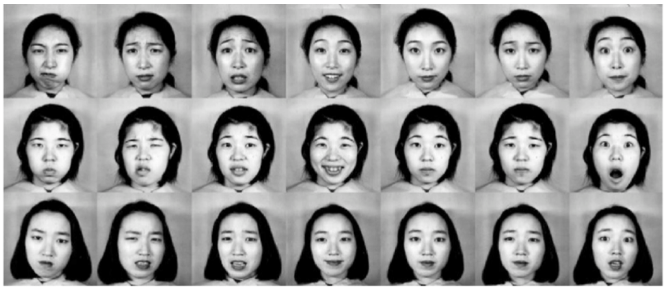


Fig. 1. Examples of JAFFE dataset

B. FER2013 Dataset

Various Datasets for FEA are available, all of them have distinct attributes, and they are different from each other. One of the widespread datasets for emotion recognition is FER2013. This dataset is publicly available on Kaggle, and the training set consists of 28,709 instances with 48*48 resolution. Moreover, FER2013 has seven facial expression labels: anger, disgust, fear, sadness, happiness, surprise, and neutral emotions. However, the number of figures in each class is not equal. For example, there are 436 data for the disgusting label, while the happy emotion label has 7,215 images. Figure 2 shows some examples of this database.

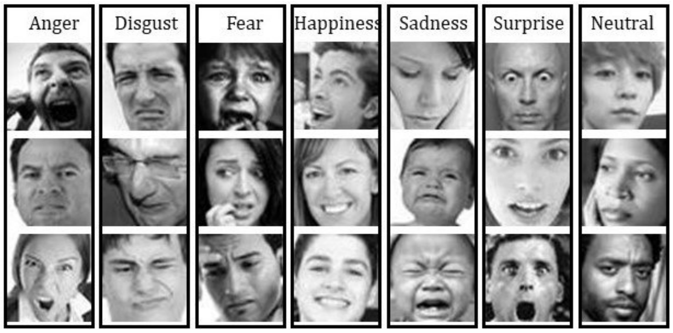


Fig. 2. Examples of FER2013 dataset [9]

III. Related Works

Many works overcome the FER2013 and JAFFE datasets and achieved varied accuracies for FEA in the In this field. In this part, we will introduce some of these works; we consider their methods in our work.

- John et al. [1] Proposed some pre-processing methods and evaluate them in. This work reveals the importance of the cropping face region. They achieved a significant enhancement in accuracy when they cropped the face region successfully. However, their approach to crop the face region differed from ours.
- Diah et al. [3] explain some preprocessing methods, such as cropping and rescaling the face figure analogous to previous work. In addition, this paper discussed some ways to enhance the CNN model, e.g., global contrast normalization, local normalization, histogram equalization, and Noise. However, their experiment was restricted to only six facial expressions and they didn't consider neutral class, and they didn't consider FER2013 in their work.
- Feng et al. [4] He proposed a way to implement local binary pattern images and use it as a feature extraction method and process the images for facial expression recognition. and they achieved 93.8% accuracy in their experiments.
- Gede et al. [5] proposed a method to fine-tune and upgrade the VGG16 network for use on FER2013 dataset. In their work, it is obvious that they manipulate the vgg16's classifier portion. In our work, we considered this model as our main model and train it on both datasets.

IV. Methodology

In this part, we will explore the method we use for implementing the preprocessing methods and VGG16 model. Before diving into the architecture of the model, we will discuss the preprocessing and augmentation methods step by step.

A. Cropping the face region

As a first preprocessing step, we crop the face region in images. Jhang et al. [11] give a solution for this approach. MTCNN is a tool that detects the face region automatically, and it performs better detection performance in comparison to the previous techniques. Furthermore, the advantage of using MTCNN on FER2013 is that this model is finetuned for this dataset and performs well. In addition, MTCNN provides efficient information about the human face; it can detect and localize the eyes, eyes, nose, and mouth, these are vital Information for further image processing and feature transformations. After cropping the face region and eliminating unnecessary information from the figure, we rescale the image to 224x224 pixels to be compatible to use in the VGG16 network. Figure 3 shows the MTCNN application cropping and rescaling on a figure from the JAFFE dataset.

B. Face Alignment

After cropping the face region, one of the problems is in each figure; the eyes are not in the exact location on the image, e.g., some figures are turned to the left or right minutely, or

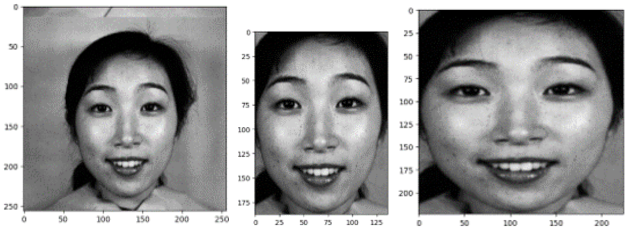


Fig. 3. (left) is the main figure in Jaffe with 256x256 resolution, (center) is cropped face region, (right) is rescaled cropped image with 224x224 resolution

the eyes are in the upper part of the images in comparison to the other figure. to solve this problem a vital approach is to put the eyes location and to scale same in all of the images. Consequently, all of the images' eyes will be in the same pixels, which can help the model to differentiate them easily. Figure 4 is an example of the alignment on the FER2013 dataset.

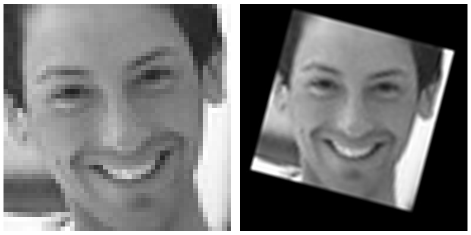
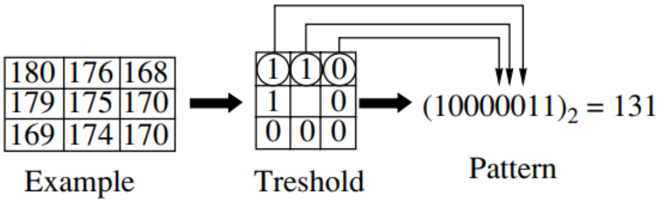


Fig. 4. (left) is the main figure in FER2013 with 48x48 resolution, (right) is aligned, rescaled image with 224x224 resolution

C. Local Binary Pattern(LBP)

Following the orders in previous works for local binary pattern implementation in previous works. The basic LBP transformation was performed the same as the figure 5. Moreover, LBP uses a 3x3 filter and transforms the pixels around the pixel in the center by adjusting a threshold, giving binary values to the values above and below the threshold, and then transforming them to decimals. In addition, this operator is used as a texture descriptor of the face region. In our work, because of the low dimensionality of the FER2013, it couldn't be implemented on this dataset. However, the implementation of the JAFFE dataset was successful. So we try this method on the JAFFE dataset.



180	176	168
179	175	170
169	174	170

1	1	0
1		0
0	0	0

$(10000011)_2 = 131$

Fig. 5. Local Binary Pattern operator [6]

D. Augmentation

The augmentation methods in this work consist of random horizontal flips and equalization for highlighting critical attributes in the image and making darker pixels evident on the face. Figure 3 clearly shows the augmentation process in this work. In addition, we normalize the data in each batch before training it in our models. In addition, we enlarge our datasets by implementing varied gaussian blur in each transformation.

E. Models

Our Work considers the VGG16 model architecture, one of the well-known architectures in the computer vision field; this Cnn model gets the 224x224x3 input image and extracts the features with 3x3 convolutional layers. However, the model we used in this work differs from fully connected layers. Generally, the vgg16 model has three fully-connected layers with 4096 neurons; our model didn't consider any FC layers in the VGG16 model; we added an average pooling layer and a layer to classify seven facial expressions. Figure 6 is a representative of our CNN model.

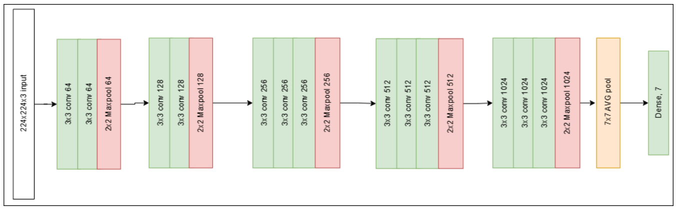


Fig. 6. The modified VGG16 Architecture has 14 layers, consisting of convolutional layers and max-pooling layers at the end of each block, after the last convolutional layer, we implement an average pooling layer with 7x7 filters and then classify the model directly after average pooling.

V. Experiments

A. Setup

Our training process in each try experiences almost the same attributes in the model and training criterion; we consider a batch size of 64 in both datasets for training and 32 for evaluation, and our loss function is cross entropy. In the first experiments of FER2013, we considered 50 epochs with a learning rate of 1e-3 with a weight decay of 1e-4 and a learning rate schedule with a gamma value of 0.33 for every 15 epochs. We considered constant 100 epochs for Jaffe, and the learning rate decay was the same as FER2013.

B. Result of JAFFE dataset

After fine-tuning the VGG16 model to train the dataset, we started our experiments. At first, non of the pre-processing methods were considered. In the later steps, we started to consider pre-processing methods step by step, e.g., on the second try, we considered cropping the face region from the whole image, and we experienced significant enhancement in training; on the 4th try, we started our augmentation to reduce the gap between training accuracy and validation accuracy, and it was a successful approach. When utilizing LBP, we first didn't consider any processing method except cropping the face region, and we achieved a disappointing accuracy of 39%. Still, after using heavy augmentation, finally, we achieved a satisfying accuracy of 98.4%. Table 1 shows our results for the JAFFE dataset.

TABLE I
RESULTS OF JAFFE DATASET

No	Norm	Aug	Crop	LBP	Align	Eql	Flip	Acc
1	✗	✗	✗	✗	✗	✗	✗	60.3%
2	✗	✗	✓	✗	✗	✗	✗	76.6%
3	✓	✗	✓	✗	✗	✗	✗	81.2%
4	✓	✓	✓	✗	✗	✗	✗	91.4%
5	✓	✓✓	✓	✗	✓	✓	✓	89.8%
6	✗	✗	✓	✓	✗	✗	✗	39%
7	✓	✓✓✓	✓	✓	✓	✓	✓	**98.4%**

the multiple ✓correspond to the heaviness of augmentation

C. Result of FER2013 dataset

For the FER2013 dataset, for the first four tries, we considered only 50 epochs for training to note the enhancement on each variation in steps, and for the remaining tries, we tried them on 100 epochs. Like the JAFFE, we didn't consider any pre-processing methods; after normalizing and augmenting the dataset weakly, we achieved 63.9% accuracy, which is high for a large dataset. At the final tries, we considered alignment on figures on FER2013; before the alignment, there was high disorderly in the dataset because in each image, faces were in a varied part of the figure and which did the model work classified the model as more inaccurate. Consequently, after implementing alignment, there is an enhancement in results, and considering more strong augmentation gives a satisfying result of 72% accuracy on the test set.

TABLE II
RESULTS OF THE FER2013 DATASET

No	Norm	Aug	Crop	LBP	Align	Eql	Flip	Acc
1	✗	✗	✗	✗	✗	✗	✗	53.2%
2	✓	✗	✗	✗	✗	✗	✗	56.6%
3	✓	✓✓	✗	✗	✗	✓	✓	63.9%
4	✓	✓✓	✓	✗	✗	✓	✓	63.4%
5	✓	✓	✓	✗	✓	✓	✓	66.1%
6	✓	✓✓✓	✓	✗	✓	✓	✓	**69.2%**

the multiple ✓correspond to the heaviness of augmentation

VI. DISCUSSIONS

In this work, varied pre-processing methods were compared. Some were effective on the FER2013 dataset, and some were effective on the JAFFE dataset because of the variation in the distribution and resolution in both datasets. e.g., higher resolution in JAFFE makes it perform better when utilizing LBP, and alignment makes significant differences on the FER2013 dataset. From the results of this work, the idea can be derived that there is less necessity to use cropping on FER2013.

There are some limitations in this work; other image processing methods could be considered compared to the implemented image processing techniques in this work. In addition, another dataset named AffectNet [13]. It is a massive dataset with almost 400 thousand instances and eleven classes, latest works were done on this dataset, and it could be one of the datasets in this work. Furthermore, other CNN Architectures could perform better than VGG16, and as a novel approach, Vision transformers could be considered, and we could evaluate the effectiveness of pre-processing on this dataset. As a feature work, we will consider all of the limitations and try to achieve the State-of-the-art(SOTA) for the FER2013 dataset and have a successful approach to the AffectNet dataset.

VII. CONCLUSION

In this work, pre-processing methods on a fine-tuned pre-trained VGG16 model were implemented, and the main Image processing methods, cropping, face alignment, and local binary patterns, were compared to each other by their performance, attributes, and scores on JAFFE and FER2013.

REFERENCES

[1] John, Ansamma, et al. "Real-time facial emotion recognition system with improved preprocessing and feature extraction." 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT). IEEE, 2020.

[2] Owada, Keiho, et al. "Quantitative facial expression analysis revealed the efficacy and time course of oxytocin in autism." Brain 142.7 (2019): 2127-2136.

[3] Douglas, Katie M., and Richard J. Porter. "Recognition of disgusted facial expressions in severe depression." The British Journal of Psychiatry 197.2 (2010): 156-157.

[4] Bargshady, Ghazal, et al. "Enhanced deep learning algorithm development to detect pain intensity from facial expression images." Expert Systems with Applications 149 (2020): 113305.

[5] Pitaloka, Diah Anggraeni, et al. "Enhancing CNN with preprocessing stage in automatic emotion recognition." Procedia computer science 116 (2017): 523-529.

[6] Feng, Xiaoyi, M. Pietikainen, and Abdenour Hadid. "Facial expression recognition with local binary patterns and linear programming." Pattern Recognition And Image Analysis C/C of Raspoznavaniye Obrazov I Analiz Izobrazhenii 15.2 (2005): 546.

[7] Kusuma, Gede Putra, Andreas Pangestu Lim Jonathan, and A. P. Lim. "Emotion recognition on fer-2013 face images using fine-tuned vgg-16." Advances in Science, Technology and Engineering Systems Journal 5.6 (2020): 315-322.

[8] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).

[9] Nunes, Ana Rita Viana. Deep Emotion Recognition through Upper Body Movements and Facial Expression. Diss. Master's Thesis, Aalborg University, 2019.

[10] Feifei Zhang, Tianzhu Zhang, Qirong Mao, and Changsheng Xu. A unified deep model for joint facial expression recognition, face synthesis, and face alignment. IEEE Transactions on Image Processing, 29:6574–6589, 2020. Project Proposal Page 9/9

[11] J. Xiang and G. Zhu, "Joint Face Detection and Facial Expression Recognition with MTCNN," 2017 4th International Conference on Information Science and Control Engineering (ICISCE), 2017, pp. 424-427, doi: 10.1109/ICISCE.2017.95.

[12] Deng, Jia, et al. "Imagenet: A large-scale hierarchical image database." 2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009.

[13] Mollahosseini, Ali, Behzad Hasani, and Mohammad H. Mahoor. "Affectnet: A database for facial expression, valence, and arousal computing in the wild." IEEE Transactions on Affective Computing 10.1 (2017): 18-31.