Efficient CNN Approach for Facial Expression Recognition

Gheyath Mustafa Zebari¹, Dilovan Asaad Zebari², Diyar Qader Zeebaree³, Habibollah Haron⁴, Adnan Mohsin Abdulazeez⁵ and Kamil Yurtkan¹

¹Cyprus International University Nicosia, Northern Cyprus, Turkey.
²Center of Scientific Research and Development Nawroz University, Kurdistan Region, Iraq.
³Director of Research Center of Duhok Polytechnic University, Kurdistan Region, Iraq.
⁴School of Computing, Universiti Teknologi Malaysia, Johor Bahru 81310, Malaysia.
⁵Presidency of Duhok Polytechnic University, Kurdistan Region, Iraq.

Email: dilovan.majeed@nawroz.edu.krd

Abstract. In the last decade, the Facial Expression Recognition field has been studied widely and become the base for many researchers, and still challenging in computer vision. Machine learning technique used in facial expression recognition facing many problems, since human emotions expressed differently from one to another. Nevertheless, Deep learning that represents a novel area of research within machine learning technology has the ability for classifying people's faces into different emotion classes by using a Deep Neural Network (DNN). The Convolution Neural Network (CNN) method has been used widely and proved as very efficient in the facial expression recognition field. In this study, a CNN technique for facial expression recognition has been presented. The performance of this study has been evaluated using the fer2013 dataset, the total number of images has been used. The accuracy of each epoch has been tested which is trained on 29068 samples, validate on 3589 samples. The overall accuracy of 69.85% has been obtained for the proposed method.

1. Introduction

Emotion recognition through utilizing facial expression has a key role in the interaction of intelligent social and this technology has been utilized in several application. Most applications that use the faces images that have been captured in the real world as it is input. Nonetheless, conventional techniques of recognition of facial expression have concentrated on the frontal face’s expressions. In the recent age, because of different poses the recognition of facial expression is considered as a challenging topic. Some of those poses are subtle expressions, changes of illumination, occlusions, and more so. Facial expressions are the essential identifiers for human sentiments, since it is correlated with feelings [1-3].

The facial expression of humans can be arranged simply into 7 essential emotions: happy, anger, surprise, sad, fear, disgust and neutral [4]. To categorize any image according to its photography can be a complex function for machines. Since, it is very simple for humans to distinguish the smiling and happy individual’s face, for computers when look to any image it will see a matrix of pixel values.

Facial Expression Recognition (FER) is more complicated than other image classification tasks [5]. It has been studied for an extensive stretch of time and acquired the advancement late decades. In spite of the fact that much advancement has been made, FER with a high precision still has difficulties because of its complexity and types of facial expressions [6,7].

In the recent age, in the domain of machine learning, computer vision methods of convolutional neural networks have investigated a key success in recognition of facial expression. Different from conventional methods, by taking deep learning methods for example Convolutional Neural Network (CNN) has ability to implement different tasks in an end-to-end manner, by training it can combine both steps of extraction features and classification together [8-10]. In contrast, recognition of facial expression development based on deep learning methods still suffer from several problems. Deep CNN may not be able to extract significant features from images that suffer from low-resolution or...
that have affected by different types of noises. Furthermore, these techniques may suffer from another
problem with data as well, the deeper CNN is the more weights there required to be specified. CNN is
a type of Neural Networks has been proved as very efficient in the scope of image recognition,
processing, and classification. Conventional techniques on FER can be classified into three main steps
which are facial detection, feature extraction and classification steps. The face detection step using
facial expression has become a well-evolved technology as well as it has been implemented on
different application of the real-world. The two of main components of FER are extracting most
relevant features as well as designing an efficient classifier. For representation of images expression,
methods-based feature and hand-crafted features are mostly utilized [4]. For instance, for capturing
different edges of images through various scales and orientations Gabor wavelets [5] may gain good
robustness. Moreover, it has been investigated that the Local binary pattern (LBP) for texture feature
extraction is beneficial in FER. A technique for recognition of facial expression has been introduced
by [8] which depends on binary patterns mapping and cognitive. A fusion technique between texture
feature using LBP and geometric features with a deep learning network method for FER has been
proposed by [9]. Based on their results it has been investigated that the proposed method is efficient in
recognition of facial expression.

2. Proposed method
This study presents an approach for facial expression classification. The main purpose of the study is
to improve the classification accuracy of the different types of facial expression. A deep learning
technique has been exploited to accomplish this work. Significant features have been extracted based
on CNN technique. Our proposed method is built based on several steps. In the first step, dataset has
been pre-processed by joining the initialization and conversions of raw data. In the next step, we have
used different layers of the CNN for different purposes. Finally, the classification step, this study has
used fully connected layer (FC) of CNN. Figure 1 shows the general architecture of the study.

2.1 Dataset
The total images of fer2013 dataset have been used in this study. It is an open-source dataset firstly,
created by Pierre-Luc Carrier and Aaron Courville for a continuing project, after that it shared as
public for a Kaggle competition. The dataset consists of 35,887 grayscale, 48x48 sized face images
with various emotions -7 emotions [10]: 8989 images- Happy, 4953 images- Angry, 5121 images-
Fear, 547 images- Disgust, 6077 images- Sad, 4002 images- Surprise, 6198 images- Neutral [13].

2.2 Pre-processing Image
Pre-processing dataset of in the area of recognition is a key stage for classification in image processing
field. CSV file (Comma-Separated Values) is plain-text file which use commas for separating values.
Lines of the file represents data records, and each record may consist of one or more fields. The
purpose of using this type of files it makes easier to deal with it. Also, importing this type of files into
any database storage or spreadsheets is simple, and massive amount of data could be organized better.
As the dataset is a CSV file which consist of 3 columns with almost 35887 rows of images, therefore
before feeding the dataset into CNN model, the initialization and conversions of raw data are required
in order to get a good classification performance and accuracy. There are many various pre-processing techniques each used in a specific case according to the problem. Such processing technique includes for instance: Image normalizations in which visual normalization is applied to solve the issue of very dark/light images, this done by computing the mean of each feature (image dimensions) in training samples then subtracting from each of images, in this case the organized data can be obtained by transforming the whole training set. Also, resizing images is a very important step to make the whole dataset images in the same size. In addition to that, dimension expanding helps to resize the dimension channel of each image height or width with preserving the aspect ratio of the original image.

2.3 Convolutional neural network (CNN) model
In this approach CNN is proposed as it is most common deep neural network learning structures. It simply consists of a list of layers which convert image volume into an output for example the seven score classes. Also, it is very effective model for image classification and recognition. There are four main units of CNN model: Convolution layer is a primary layer, which is used to extract features from the raw data (input image). By feeding 64 samples (images) each time to the network, which finally output the possibilities of seven different facial expression. The ReLU (Rectified Linear Unit) layer applies activation function per pixel, which replaces each negative pixel value in feature map into zero. Pooling Layer is also called subsampling layer, used to decrease the dimensionality of each feature and at the same time keeping the important information. Finally, fully connected layer (classification) computes the class scores using a traditional multilayer perceptron, in which each neuron is connected to all other neurons. It uses activation function SoftMax (multi class output), the main reason of this layer is taking the incoming input volume from convolution and pooling layer in to seven output emotion expression. In addition to that dropout of 50% is used within each convolution layer to prevent the network from overfitting during training data and batch normalization is added to standardize the inputs automatically which make the network faster and more stable to improve the accuracy of the model. Table 1 represents the configuration of neural network, here the batch size of input data is 64 x 64, which means feeding the network each time with 64 images. Kernel size or filter size is same for Convolutional and Pooling layers (3x3) and the stride is (2) that mean moving the filter matrix two pixels at a time.

| Layer       | Size   | Stride | Layer       | Size   | Stride | Layer         | Size   | Stride |
|-------------|--------|--------|-------------|--------|--------|---------------|--------|--------|
| Data        | 64 x 64|        | Conv_1      | 3 x 3  | 2      | Max_pooling_1 | 2 x 2  | 2      |
| Conv_2      | 3 x 3  | 2      | Max_pooling_2 | 2 x 2 | 2      | Conv_3       | 3 x 3  | 2      |
| Max_pooling_3 | 2 x 2 | 2      | Conv_4      | 3 x 3  | 2      | Max_pooling_4 | 2 x 2  | 2      |
| Fully connected layer | - |        |             |        |        |               |        |        |

3. Experimental results
In experimental setup the FER2013 dataset has been used to evaluate the accuracy of the study. Accuracy is a common metric that utilized widely for monitoring the performance of any machine learning system. It can be defined the ratio of the cases that classified correctly to the whole number of cases. Fer2013 is an open-source dataset firstly, created by Pierre-Luc Carrier and Aaron Courville for a continuing project, after that it shared as public for a Kaggle competition. The dataset consists of 35,887 grayscale, 48x48 sized face images with 7 emotions. Beside the category of images, this study divided images into three different sets, training, validation, and testing. The total number of 28,709 samples have been used in the training whereas in validation and testing we have used 7178 samples: 8989 Happy, 547 Disgust, 4002 Surprise, 4953 Angry, 6077 Sad, 5121 Fear, and 6198 Neutral. Python programming language used to design and implement such algorithms in the study, as it contains a variety of libraries related to Neural Network Algorithms.
To verify the efficiency of the proposed study large set of images have been tested. The results based on CNN architecture with pre-processing techniques have been shown in Table 2. The achieved results tested on total images of fre2013 database using 70 epochs. Increasing the epochs needs long time for training (large number of iterations) because deep learning techniques need large number of datasets for training. The training and testing time of our proposed network for running on FER2013 dataset were taken about 48 hours. Table 2 presents the accuracy of each epoch of the study. Based on Table 2 obviously it has been shown that at the first epoch obtain a low accuracy about %24 while the epoch number 70 (last epoch which has been tested by our system) obtained higher accuracy %69. Thus, to obtain high accuracy the number of epochs should be increased. Figure 2 illustrates that testing our system obtained higher accuracy with a greater number of epochs.

![Figure 2: Graphical Analysis for Each Epoch](image)

![Table 2. Accuracy Results](image)

| Epoch | Accuracy | Epoch | Accuracy | Epoch | Accuracy | Epoch | Accuracy |
|-------|----------|-------|----------|-------|----------|-------|----------|
| 1     | 0.2452   | 10    | 0.3265   | 20    | 0.3995   | 30    | 0.4709   |
| 40    | 0.5688   | 50    | 0.6297   | 60    | 0.6697   | 70    | 0.6985   |

Furthermore, Figure 3 presents the results experiment for the proposed method using FER2013. As it is demonstrated in the Figure 3, the results are based on the confusion matrix. A confusion matrix is a matrix or table which holds information of actual and predicted classifications which is implemented by a system of classification. The input number of the predicted label is indicated by the main diagonal elements whereas the rest of the elements are considered mislabelled by the classifier. Maximization of the sum of values over the diagonal of the confusion matrix matches the system performance in classification. Based on the sum division over diagonal over the sum of all values in the confusion matrix, the classification accuracy can be achieved. The proposed method based on the CNN obtained high accuracy results for most of the labels of FER2013 dataset. The proposed method is performed on all categories of the facial expression dataset. As it is shown in the Figure 3, disgust and happy categories outperformed all other categories by achieving accuracy of 0.78% and 0.79%, respectively. This means that the features of both categories disgust and happy are more differentiable compared to other expressions. Moreover, the fear category obtained accuracy of 0.71% which considered as a high accuracy as well. However, the expression of both categories sad and surprise are easier to be known,
but lower accuracy results have been obtained 0.58% and 0.61% respectively. Finally, the overall accuracy of 69.85% for FER2013 dataset has obtained.

Table 3 depicts the performance evaluation of the proposed study with several studies in the literature. We have listed the accuracy of our proposed study compared with other state-of-the-art methods in Table 3 based on using FER2013 database. Several traditional descriptor features such as HOG, Gabor filters, and LBP were utilized widely in the literature for facial expression recognition. In contrast, it has been discovered that the traditional methods obtained lower accuracy results in recognition of facial expression compared to deep learning methods. The results of performance evaluation that obtained based on the computing methods using FER2013 dataset which is considered as a most challenging dataset in recognition of facial expression are presented in Table 3. Our proposed method obtained accuracy of 69.85% which is ranked higher that studies of [18, 19, 21] one the Table 3. However, the study of [13, 17, 20] obtained higher accuracy compared our proposed method slightly. Therefore, based on our evaluation results it has been investigated that the proposed method is robust and efficacy in recognition of facial expression for FER2013 dataset.

Table 3. Accuracy (%) Comparison between Our Method with Different Methods in the Literature

| Methods       | Accuracy (%) | Methods       | Accuracy (%) | Methods       | Accuracy (%) | Methods       | Accuracy (%) |
|---------------|--------------|---------------|--------------|---------------|--------------|---------------|--------------|
| [1]           | 65           | [11]          | 70.58        | [12]          | 70.47        | [13]          | 61.86        |
| [14]          | 57.10        | [15]          | 71.14        | [16]          | 62.99        | Proposed Method | 69.85        |

4. Conclusion

We proposed a facial expression recognition method based on the deep CNN technique and its performance has been analysed and evaluated in terms of accuracy using the FER2013 dataset. The proposed method consists of four layers within each layer max pooling and finally the fully connected layer. It takes the raw data (input image) and classifies it into one of the seven facial expressions: angry, fear, happy, neutral, disgust, sad, and surprise. In our performance evaluation, we have tested our proposed method using 70 epochs. The obtained results demonstrated that both learnings of the facial feature as well as classifying emotion could achieve better performance for a deeper evaluation model. However, the evaluation of experiments has implemented on 70 epochs, where the lower epoch obtained lower accuracy whereas the higher epoch obtained higher accuracy, which means that it could achieve a good score for more epochs in recognition of facial expression. Therefore, in future work, we could develop three different models for CNN by evaluation using more epochs.

References

[1] A. Agrawal and N. Mittal, The Visual Computer, 2020 36(2), 405-412
[2] D. Q. Zeebaree, H. Haron, A. M. Abdulazeez, and D. A. Zebari, 2019 International Conference on Advanced Science and Engineering (ICOASE), 2019: IEEE, pp. 106-111.
[3] D. A. Zebari, D. Q. Zeebaree, A. M. Abdulazeez, H. Haron, and H. N. A. Hamed, 2020 IEEE Access. 8 203097-203116
[4] G. Sikander, S. Anwar, and Y. Abd Djawad, 2017 5th International Symposium on Computational and Business Intelligence (ISCBI), 2017: IEEE, pp. 26-31.
[5] S. C. Bakchly, M. J. Ferdous, A. H. Sathi, K. C. Ray, F. Imran, and M. M. Ali, 2017 2nd International Conference on Electrical & Electronic Engineering (ICEE), 2017: 1-4.
[6] R. Zebari, A. Abdulazeez, D. Zeebaree, D. Zebari, and J. Saeed, Journal of Applied Science and Technology Trends, 2020 vol. 1, no. 2, pp. 56-70.
[7] D. A. Zebari, H. Haron, S. R. Zeebaree, and D. Q. Zeebaree, 2019 International Conference on Advanced Science and Engineering (ICOASE), 2019, 100-105.
[8] C. Qi et al., IEEE Access, 2018 vol. 6, pp. 18795-18803.
[9] D. Q. Zeebaree, A. M. Abdulazeez, D. A. Zebari, H. Haron, and N. A. Haza, Computers,Materials & Continua, 2021 Vol. 66, No.3, pp. 3363-3382.
[10] D. Q. Zeebaree, H. Haron, A. M. Abdulazeez, and D. A. Zebari, *2019 International Conference on Advanced Science and Engineering (ICOASE)*, 2019: IEEE, pp. 88-93.

[11] C. Pramer dorfer and M. Kampel, *Computer Science*, 2016 Corpus ID 16605821 arXiv preprint arXiv:1612.02903.

[12] J. Jeon et al., *Proceedings of the 10th International Conference on Ubiquitous Information Management and Communication*, 2016, pp. 1-4.

[13] G. Zeng, J. Zhou, X. Jia, W. Xie, and L. Shen, *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, 2018: IEEE, pp. 423-430.

[14] V. Tümen, O. F. Söylemez, and B. Ergen, *2017 International Artificial Intelligence and Data Processing Symposium (IDAP)*, 2017: IEEE, pp. 1-5.

[15] J. Shao and Y. Qian, *Neurocomputing*. 2019 vol. 355, pp. 82-92.

[16] S. Xie, H. Hu, and Y. Wu, *Pattern Recognition*, 2019 vol. 92, 177-191.