# Facial Emotion Recognition Using Shallow CNN

Sachin Saj T.K, Seshu Babu, Vamsi Kiran Reddy, Gopika.P, Sowmya.V, and Soman K.P

Center for Computational Engineering & Networking (CEN),Amrita School of Engineering, Coimbatore, Amrita Vishwa Vidyapeetham, India.
sachin96saj@gmail.com, v_sowmya@cb.amrita.edu

**Abstract.** Facial emotional recognition became an important task in the modern day scenario to understand the state of emotions of a human being by machines. With the development of computational power and deep learning techniques, facial emotion recognition (FER) became feasible, which contributed to a wide range of applications in modern day technology. In this paper, we propose a shallow convolutional neural network architecture with feature-based data, which can do this task more effectively and attained the state-of-the-art accuracy with less computational complexity (in terms of learnable parameters). The proposed architecture is shallow and gives comparable performance with all the existing approaches for FER in deep learning.

**Keywords:** Deep learning, CNN, Facial emotion recognition

## 1 Introduction

Facial expressions render more information, which can be used to understand the psychological state of a human being. The automatic facial emotion recognition (FER) is very important and it's a challenging problem in the community of computer vision[1]. This was successful in attracting attention in the modern world due to its wide potential applications in areas such as robotics, autonomous car, communication, health care etc [2]. Convolutional neural network (CNN) architectures in deep learning have achieved significant results in the field of computer vision [3].

There exists many FER methods from past many years. In [2], the authors have introduced two approaches for their FER method. One such method was the use of auto-encoders but, it was unable to generate proper results and the second method, CNN architecture with only 3 convolution layers, 3 max-pooling layers and 2 fully connected layers. Through this architecture, they were successful in generating state-of-the-art accuracy with JAFFE dataset. In [4], the authors proposed a method based on a weighted mixture deep neural network, which process both grey scale facial emotion images as well as local binary pattern (LBP) facial images simultaneously. The weighted fusion techniques were used to improve the recognition ability. A partial VGG16 network is being constructed to extract features from the grey scale facial images and shallow CNN

architecture is constructed to extract features from LBP facial images. They have done their evaluation on three benchmark dataset to verify the effectiveness of their proposed approach. Their approach was able to give better state-of-the-art accuracy and was compared with other existing approaches in FER.

Even though the computational complexity was high in [4], they were able to recognize the facial expression in few seconds and have high accuracy. In this paper, we propose a shallow CNN with feature-based data as the input in order to reduce the computational complexity (in terms of number of learnable parameters) of the architecture. The proposed architecture is experimented using JAFFE dataset, which attains comparable performance with all the existing FER approaches which use deep learning.

## 2 Proposed CNN Architecture

Based on previous publications and results, we understood the need to reduce the number of learnable parameters of the architecture without compromising the accuracy. So, we decided to design our own shallow CNN architecture from the scratch. Our proposed architecture is a 6-layer CNN with 3 convolution layers, 2 pooling layers and 1 fully connected layer. The proposed CNN architecture is shown in figure 1 and the details about the architecture is given in table 1

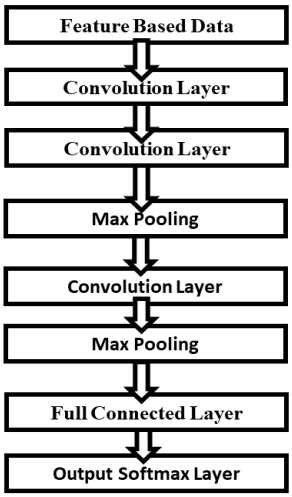


**Fig. 1.** Proposed shallow CNN architecture for Facial Emotional Recognition (FER)

**Table 1.** Details of The proposed shallow CNN architecture

| **Parameters** | **Conv2D** | **Conv2D** | **Max Pooling** | **Conv2D** | **Max Pooling** |
|---|---|---|---|---|---|
| No. of filters | 128 | 64 | - | 64 | - |
| Size of the kernel | 3x3 | 3x3 | 2x2 | 3x3 | 2x2 |
| Strides | 1 | 1 | 1 | 1 | 1 |
| Padding | Same(1) | Valid(0) | - | Same(1) | - |

## 3 Dataset Description

The dataset which we have used for facial emotion recognition is JAFFE dataset [2]. This dataset consists of 213 images of 10 different Japanese female posing for 7 different emotions such as happiness, sadness, surprise, anger, disgust, neutral and fear. The split of data is 70:30, where 70% of the data is used for training and 30% of the data is used for testing.

## 4 Results and Discussion

JAFFE dataset is given to our proposed shallow CNN architecture in three different ways: one as raw input without data augmentation, second as raw input with data augmentation and finally feature-based data is given, which is extracted using haar cascade package from OpenCV library. In this, only the features of the face are selected and the rest of the portions are deleted, thus helping in reducing the computational complexity. From table 2, it is clear that, feature based data when given to sample 1-layer CNN architecture ran for 1000 epochs, outperformed other ways of feeding the data.

**Table 2.** Test accuracy for different types of data input

| **Data** | **CNN layer** | **No of epochs** | **Test accuracy(%)** |
|---|---|---|---|
| Raw Input | 1 | 1000 | 22.27 |
| Raw Input (Data Augmentation) | 1 | 1000 | 50.32 |
| Feature based data | 1 | 1000 | 62.39 |

### 4.1 Hyper-Parameter Tuning

The reason for hyper-parameter tuning is to get the best combination of the number of filters in each layers, where the highest accuracy can be achieved. It is evident from the table 2 that, with feature-based data given to 1-layer CNN, an accuracy of 62.39% is achieved. From table 3, it is evident that the accuracy

**Table 3.** Hyper-Parameter tuning for No. of filters in each layer of convolution layer

| No of layers | No of filters (Each layers) | Batch size | Epochs | Test accuracy(%) |
|---|---|---|---|---|
| **3** | **128-64-64** | **20** | **3000** | **91.93** |
| 3 | 128-64-64 | 20 | 2000 | 90.30 |
| 3 | 64-32-32 | 20 | 2000 | 87.09 |
| 3 | 256-128-128 | 20 | 2000 | 88.70 |
| 3 | 512-256-256 | 20 | 2000 | 88.70 |

increases with the number of CNN layers (from 1 layer to 3 layers) and epochs (1000 to 3000).

It is evident from the table 3 that, 128-64-64 number of filter combinations in each layer gave the best result when it ran for 2000 epochs compared to all other combinations. Then all the combination was again ran for 1000 more epochs in which, only 128-64-64 combination was able to improve its accuracy from 90.30% to 91.93%. So, 3-layer CNN architecture with 128-64-64 number of the filter in each layer combination gave the best result. All the experiments were implemented by using keras[5] and scikit-learn[6] libraries.

### 4.2 Performance measure

The performance of the proposed architecture which gave better results compared to all other combinations is evaluated on test data set. The details of the result are shown in table 4.

**Table 4.** The performance measures for the proposed architecture

| Class | Precision | Recall | F1-score |
|---|---|---|---|
| Happiness | 0.89 | 1.00 | 0.94 |
| Sadness | 1.00 | 0.90 | 0.95 |
| Surprise | 0.67 | 1.00 | 0.80 |
| Anger | 1.00 | 1.00 | 1.00 |
| Disgust | 0.88 | 1.00 | 0.93 |
| Neutral | 1.00 | 0.69 | 0.82 |
| Fear | 1.00 | 1.00 | 1.00 |

It is evident from the table 4, four classes have achieved 100% in precision and five classes were able to achieve 100% in recall. The f1-score which is a weighted average of precision and recall, represents the accuracy of the classifier in classifying the images in that particular class compared to all other classes. Among the obtained result, two classes achieved 100% and the rest of the classes is above 80%.

### 4.3 Comparison with existing approaches

To determine the performance of the classifier, confusion matrix corresponding to our proposed architecture is being taken and it is compared with the confusion matrix [4]. This is shown in table 5 and table 6

**Table 5.** Confusion matrix of our proposed architecture

| True label \ Predicted label | Happiness | Sadness | Surprise | Anger | Disgust | Neutral | Fear |
|---|---|---|---|---|---|---|---|
| Happiness | **0.89** | 0.00 | 0.00 | 0.00 | 0.00 | 0.11 | 0.00 |
| Sadness | 0.00 | **1.00** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Surprise | 0.00 | 0.00 | **0.67** | 0.00 | 0.00 | 0.33 | 0.00 |
| Anger | 0.00 | 0.00 | 0.00 | **1.00** | 0.00 | 0.00 | 0.00 |
| Disgust | 0.00 | 0.12 | 0.00 | 0.00 | **0.88** | 0.00 | 0.00 |
| Neutral | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | **1.00** | 0.00 |
| Fear | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | **1.00** |

In [4], the authors proposed an architecture considered only six facial emotions such as anger, disgust, fear, happiness, sadness and surprise and was able to achieve an accuracy of 92.2%. In present work, we considered all the seven facial emotions including neutral emotion and we were able to achieve the comparable result of 91.93% with [4]. In our proposed architecture, we were able to get 100% classification accuracy in four classes such as sadness, anger, neutral and fear. Only in surprise class, misclassification of 33% with a neutral class occurs. Whereas in [4], the proposed architecture was able to obtain more than 89% classification accuracy in all the classes. So, in our proposed architecture, even though we considered one more extra facial expression we were able to achieve the comparable result (91.93%).

**Table 6.** Confusion matrix of benchmark paper [4]

| True label \ Predicted label | Anger | Disgust | Fear | Happiness | Sadness | Surprise |
|---|---|---|---|---|---|---|
| Anger | **0.95** | 0.02 | 0.01 | 0.00 | 0.02 | 0.00 |
| Disgust | 0.02 | **0.89** | 0.06 | 0.00 | 0.03 | 0.00 |
| Fear | 0.03 | 0.04 | **0.90** | 0.00 | 0.03 | 0.00 |
| Happiness | 0.01 | 0.00 | 0.02 | **0.94** | 0.00 | 0.03 |
| Sadness | 0.00 | 0.05 | 0.03 | 0.00 | **0.91** | 0.01 |
| Surprise | 0.02 | 0.00 | 0.02 | 0.03 | 0.00 | **0.93** |

The Table 7 shows the comparsion of the proposed architecture with some of the existing facial emotion recognition methods in deep learning.

**Table 7.** Comparison of proposed architecture with the existing approaches in FER

| **Existing Approaches** | **Testing Accuracy (%)** |
|---|---|
| *Aly et al[7]* | 87.32 |
| *Rivera et al[8]* | 88.75 |
| *Prudhvi [2]* | 86.38 |
| *Zhang et al[9]* | 91.48 |
| *Yank et al[4]* | 92.21 |
| **Proposed Architecture** | **91.93** |

It is evident from the Table 7, our proposed architecture was able to outperform many existing approaches in FER.

The Table 8, shows the number of learnable parameters of our proposed architecture compared with learnable parameters of the existing approaches.

**Table 8.** Comparison of learn-able parameter of proposed architecture with the existing approaches

| **Approaches** | **Learn-able Parameters** |
|---|---|
| *Yang et al[4]* | 58,482,062 (approx.) |
| *Prudhvi [2]* | 440,167 |
| **Proposed Architecture** | **18,990,471** |

The work proposed by Yank et al., is considered as the benchmark paper and the work proposed by Prudhvi, is considered as the base paper. It is evident from the table 7, our architecture was able to give better accuracy than the results obtained in [2], and was able to achieve comparable accuracy with Yang et al. From table 8, it is clear that the number of learnable parameters in [2] is very less and our proposed architecture have more learnable parameters than of [2], but we were able to increase the accuracy. When compared with, the learnable parameter in that architecture proposed in [4] is about 58,482,062 and ours is 18,990,471. We were successful in reducing the number of learnable parameters three times of the benchmark paper and still we were able to achieve the comparable accuracy with [4]. Hence, we have reduced the complexity (in terms of number of learnable parameters) to a larger extent without compromise on accuracy by using the feature based approach on JAFFE dataset [2].

## 5 Conclusion

This study proposes a method which uses shallow CNN architecture with feature-based data of JAFFE dataset. The proposed method is able to achieve state-of-the-art-accuracy with reduced number of learnable parameters compared to existing approaches based on deep learning for Facial Emotion Recognition.The performance of the proposed shallow CNN for Facial Emotional Recognition is mainly due to the haar cascade features extracted from the images are given as input to the proposed architecture. The future scope on the present work is to extend the analysis of feature based shallow CNN for the classification problems in other domains such as bio-medical and image based cyber security applications

## References

1. Andre Teixeira Lopes, Edilson De Aguiar, and Thiago Oliveira-Santos. A facial expression recognition system using convolutional networks. In *2015 28th SIBGRAPI Conference on Graphics, Patterns and Images*, pages 273–280. IEEE, 2015.
2. Prudhvi Raj Dachapally. Facial emotion detection using convolutional neural networks and representational autoencoder units. *arXiv preprint arXiv:1706.01509*, 2017.
3. R Vinayakumar, KP Soman, and Prabaharan Poornachandran. Applying convolutional neural network for network intrusion detection. In *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 1222–1228. IEEE, 2017.
4. Biao Yang, Jinmeng Cao, Rongrong Ni, and Yuyu Zhang. Facial expression recognition using weighted mixture deep neural network based on double-channel facial images. *IEEE Access*, 6:4630–4640, 2018.
5. Antonio Gulli and Sujit Pal. *Deep Learning with Keras*. Packt Publishing Ltd, 2017.
6. Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.
7. Sherin Aly, A Lynn Abbott, and Marwan Torki. A multi-modal feature fusion framework for kinect-based facial expression recognition using dual kernel discriminant analysis (dkda). In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–10. IEEE, 2016.
8. Adin Ramirez Rivera, Jorge Rojas Castillo, and Oksam Oksam Chae. Local directional number pattern for face analysis: Face and expression recognition. *IEEE transactions on image processing*, 22(5):1740–1752, 2013.
9. Wei Zhang, Youmei Zhang, Lin Ma, Jingwei Guan, and Shijie Gong. Multimodal learning for facial expression recognition. *Pattern Recognition*, 48(10):3191–3202, 2015.