J. Inst. Eng. India Ser. B (April 2022) 103(2):439–448
https://doi.org/10.1007/s40031-021-00681-8

ORIGINAL CONTRIBUTION

FERNet: A Deep CNN Architecture for Facial Expression Recognition in the Wild

Jyostna Devi Bodapati[1] · **U. Srilakshmi**[1] · **N. Veeranjaneyulu**[2]

Received: 6 April 2021 / Accepted: 11 September 2021 / Published online: 7 October 2021

Abstract Facial expression recognition is an intriguing and demanding subject in the realm of computer vision. In this paper, we propose a novel deep learning-based strategy to address the challenges of facial expression recognition from images. Our model is developed in such a manner that it learns hidden nonlinearity from the input facial images, which is critical for discriminating the type of emotion a person is expressing. We developed a deep convolutional neural network model composed of a sequence of blocks, each consists of multiple convolutional layers and subsampling layers. Investigations on the benchmark FER2013 dataset indicate that the proposed facial expression recognition network (FERNet) surpasses existing approaches in terms of performance and model complexity. We trained our model on the FER2013 dataset, which is the most challenging of all the available datasets for this task, and achieve an accuracy of around 69.57%. Furthermore, we investigate the effects of dropout, batch normalization, and augmentation, as well as how they aid in the reduction of over-fitting and improved performance.

Keywords Facial emotion recognition · Convolution neural network (CNN) · Data augmentation · Human–computer interaction (HCI) · Image classification · FER2013

✉ Jyostna Devi Bodapati
jyostna.bodapati82@gmail.com

[1] Department of Computer Science and Engineering, Vignan's Foundation for Science Technology and Research, Vadlamudi, Andhra Pradesh, India

[2] Department of Information Technology, Vignan's Foundation for Science Technology and Research, Vadlamudi, Andhra Pradesh, India

Introduction

An expression conveys more information than actual actual spoken words. Since a long time non-verbal communication was considered equally if not more important than verbal communication. The most visible and essential type of nonverbal communication is facial expression. According to Ekman and Friesen [1], primary expressions used by people to communicate their emotions include anger, disgust, fear, happiness, sorrow, and surprise. Emotion recognition has sparked a lot of interest in the scientific community, with substantial advancements in the field of computer vision. This is an important application and provides baseline for many other applications in the field of human–computer interaction and marketing.

Machine has become successfully applied for a variety of real-time applications like weather prediction, image and object recognition and medical image analysis applications [2–5]. Following the success of machine learning, research community started utilizing them for facial emotion recognition and significant improvement was observed [6]. However, success of these traditional machine learning models is subject to the quality of the features that are provided during training. In the recent past, several conventional feature learning techniques are applied on facial images to obtain feature descriptors which are used to identify facial emotions [7]. A support vector machine (SVM) classifier is trained using histogram of oriented gradients (HOG) feature descriptors extracted from facial images are used to detect the kind of emotion present in facial images [8, 9]. Local binary patterns features are employed for automated emotion recognition from facial images [10]. The success of these traditional machine learning models is limited dual to the manual setting required for the hand-crafted feature extraction [11].

Deep neural networks (DNNs) were introduced addressing the challenges of traditional models and these models were much appreciated by both industry and academia as they are capable of learning hidden patterns from the input data at multiple levels [12]. Convolutional neural networks (CNNs), a variant of DNNs, led to major breakthroughs in many computer vision related tasks like object detection [13], medical image analysis [14]. CNNs are very efficient and are excellent feature extractors that can extract features from images on a deeper level. They basically have two major operations: convolution and pooling. The input picture is sent to the convolution layer, where filters are applied to extract the features from the supplied image. Multiple filters can be used to extract different kinds of features that are required for the task under consideration. Convolutional layers are followed by pooling layers to minimize duplication and, to ultimately, to reduce computational cost and improve network depth. Deep features have become popular and contributed to boost performance of the models developed for facial expression recognition [15].

Making use of the existing research and careful analysis of various CNN architectures, we have come up with a novel and relatively simple and straight forward CNN architecture. The images were sent to the CNN to extract features from the provided input face at several layers, which were then fed into the output softmax layer for facial image categorization into one of the seven emotion classes. FER2013, most challenging dataset for facial expression recognition, is used to train the FERNet model and an accuracy of around 69.57% is achieved. This is greater than the human accuracy on this dataset, which was about 65%. Various augmentation approaches have been applied to the facial images to enhance the dataset to compensate for the limited size dataset. We have made many findings about how different data augmentation approaches and their settings impact the outcomes. This work includes the following major contributions:

- Various augmentation approaches have been applied to the facial images to enhance the dataset and compensate for the limited size dataset
- Developed a novel CNN architecture & train it using FER 2013 benchmark dataset
- Study the effect of varying kernel sizes, Dropout, and Batch-normalization in CNN

Related Work

For humans, understanding the emotions and feelings of other individuals based on his/her facial expressions is a trivial task, but for a machine, doing the same task is quite challenging. Machines are now capable of recognizing human emotions through facial expressions and are nearly as accurate as humans, thanks to recent advances in research, particularly in the fields of computer vision and machine learning. Numerous models and approaches have been proposed for this task over the years. This section examines current machine learning models, with an emphasis towards deep learning frameworks developed for the FER 2013 dataset. For real-time facial emotion recognition, this dataset is known to be one of the most challenging as it includes the images of cartoons and animals.

Liu et al. [16] proposed an ensemble of three CNNs that are trained separately at first and are then assembled together into a single model and trained as a single model. They claimed that by focusing on individual CNNs instead of a single one and then ensemble of those models leads to better performance. In their study, Mollahosseini et al. [17] developed a deep neural network with two convolution blocks, each comprising convolutional and sub sampling operations, followed by four Inception layers. The network is developed as a single-component architecture that accepts recorded facial images and categorizes them into one of six basic or neutral emotions. The output from the inception layers was given to two dense layers to make predictions. A hybrid method was proposed to get the advantages from both scale invariant feature transformations (SIFT) and CNN features [18]. They experimented using both SIFT and dense SIFT features which are merged at the final stage of CNN.

Gan et al. [19] tried fine-tuning of deep CNN architectures such as AlexNet, GoogleNet, VGGNet and ResNet on the FER2013 and their experimental results indicate that AlexNet gives better accuracy compared to the other deep CNN models and ResNet is the next better model for FER. Agarwal et al. [20] proposed two different CNN architectures by evaluating the influence of kernel sizes and filter counts on the model. Their work was inspired from the pre-trained deep CNN architectures such as VGGNet and Inception and resulted in improved performance for facial expression recognition.

Giannopoulos et al. [21] carried out extensive set of experiments on FER2013 and they examined and realized that the performance of GoogLenet is far better than AlexNet on FER2013. They approached the challenge as a two-class classification problem, grouping all non-neutral faces into a single class. In their initial experiments, they developed a model to check the existence of emotion in the given image. They concluded that both GoogLenet and AlexNet were almost identical in their performance which may be due to the fact that it is a binary classification task. As a second experiment, they trained the networks with all the classes except the neutral class. In this task, their

objective is to recognize the emotion type and their experimental studies claim that GoogLenet performs better but as iterations complete AlexNet catch up and achieves results on-par with GoogLenet. In the final task, the authors trained the networks on the entire dataset and conclude that because of its shallow architecture AlexNet trains faster and converges faster when compared to GoogLeNet; however, after a large number of iterations, GoogLenet gives better results because of its deep architecture while AlexNet suffers from over training.

The introduction of generative adversarial networks (GAN), new options for image generation, such as generating faces for facial expression recognition, have emerged [22]. They increased the the number of classes in the dataset from N to 2N (during the learning phase) by taking into account actual and fake emotions. A deep CNN was suggested, consisting of two blocks of convolutional layers with 3072 filters and a sub-sampling layer, followed by a convolutional layer and a dense layer for emotion classification [23]. Their experiments infer that setting proper parameters for regularizations, dropout and batch normalization using grid search lead to better accuracy by reducing over-fitting. It has been widely accepted that CNNs are good at capturing details from the input images at various levels. The activation function used at different layers has to be wisely chosen, because the nonlinearity of the deep neural network comes with the type of activation functions used and these activation functions make the deep neural network powerful. The influence of existing activation functions in the CNN models developed for FER is studied and realized the need for the new piece wise activation function [24]. The outcome from their experiments demonstrates that the CNN with the new enhanced activation function outperforms existing activation functions.

An attempt was made to improve accuracy of the FER models by using pre-processing techniques such as face detection, key point extraction and illumination correction. A CNN architecture with six layers was introduced [25] which takes detected portions of the face images after applying illumination correction. Further studies explored into the effectiveness of network structure and data preprocessing approaches for developing a baseline structure for face expression recognition [26]. Through their experiments, they discovered that the histogram equalization (Hist-eq) is the most dependable approach for preprocessing the input facial images, as it performed better with a variety of machine learning and deep learning models. Moreover, they found that Tang's [27] network achieved reasonably high accuracy with Hist-eq images even with less network complexity compared to the other networks such as Caffe-ImageNet. In an effort to apply VGGNet and AlexNet, dropout was effectively used to reduced architecture [28]. Authors carefully picked layers from the pre-trained VGGNet architecture and came up with a novel model for the classification of expressions from facial images. They were able to decrease model over-fitting and obtain improved performance on the FER-2013 dataset by using dropout and other regularization techniques.

Proposed Method

Data Preprocessing

The original images of the dataset are gray scale and are resized to the shape of $48 \times 48 \times 1$. Data augmentation is much beneficial when working with limited datasets since it prevents the model from over-fitting. In case of few expressions, the number of images are low and hence are augmented by using flipping and rotating them. The resized images are flipped horizontally and rotated randomly to generate different versions of the original images. In addition, the data are normalized to bring it to the same scale.

Model Architecture

Convolutional neural network models, one of the successful and the most widely used neural network models for variety of vision tasks. CNN models contributed to achieve record scores for variety of real-time applications. Training deeper CNN models often require large computational power and take longer time to train. We designed a comparatively simple CNN architecture from scratch that is able to produce substantially good results.

The architecture is made up of five convolution blocks, as illustrated in Fig. 1. The first block is made up of two convolution and batch normalization layer stacks, followed by max pooling and drop out layers. Except for the number of convolution and batch normalization layers stacked, the remaining blocks have the same structure. The third block uses 3 stacks, while the fourth and fifth blocks employ four stacks of convolution and batch normalization layers. Finally, the output of convolution base is flattened and processed through a dense layer to classify the given images into one of the seven classes representing emotions. After every max pooling, a dropout layer is used to avoid over-fitting. A kernel size of 4×4 is used with all the convolutional layers while a 2×2 non-overlapping window size is used for the max-pooling layers. The ReLU activation functions are known to improve model convergence when compared to other activation functions and we use these ReLU activation functions with all the convolutional layers. The output from the last max-pooling layer is fed into a dense layer with soft-max activation, which is the

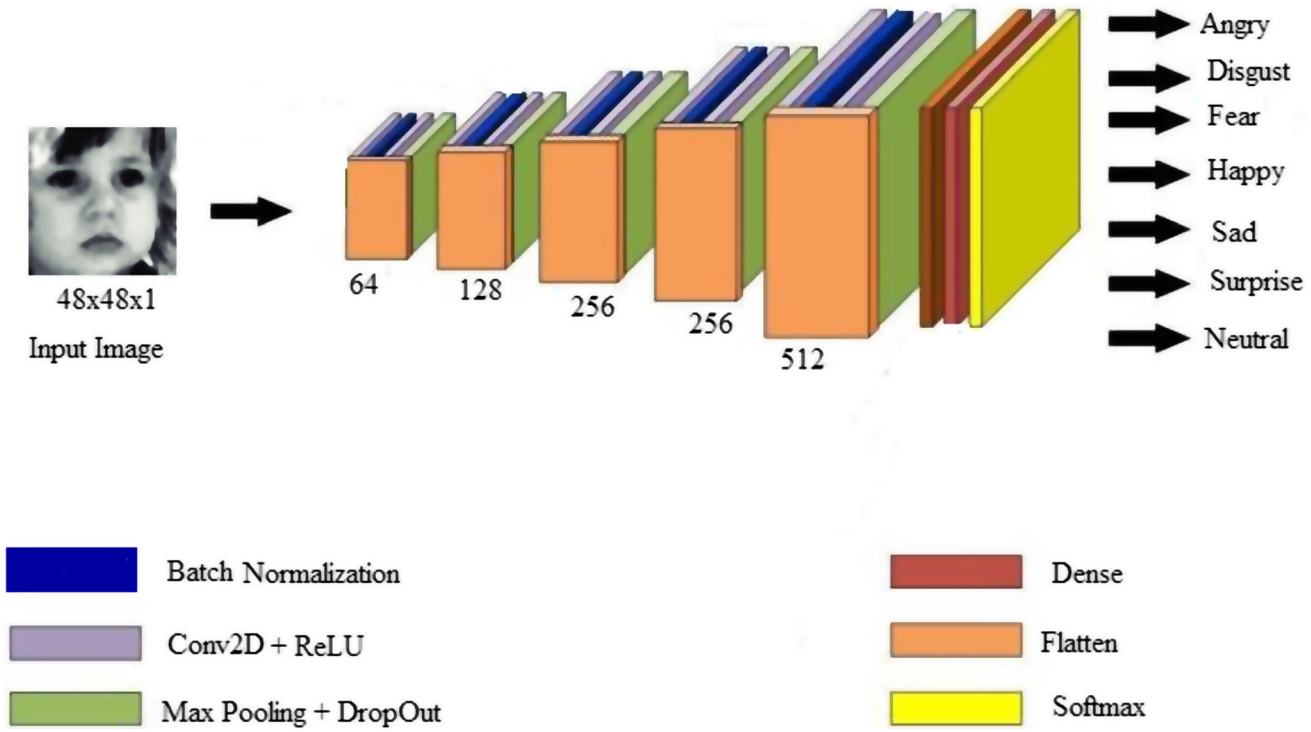

Fig. 1 Proposed deep convolutional neural network architecture for emotion recognition

most frequent activation for multi-way classification tasks. The mathematical expressions for different operations are listed below. The expression for the ReLU nonlinearity is:

$$f(s) = max\{0, s\} \tag{1}$$

where s is the aggregated output from the neuron. The expression for softmax function:

$$f(s_i) = \frac{e^{s_i}}{\sum_{j=1}^{k} e^{s_j}} \tag{2}$$

In Eq. 2, s_i represents the aggregated value from the ith neuron, while k refers to the number of neurons at a layer.

The output of layer $l-1$ is convolved by the filters at layer l during the forward propagation stage of model training, and the output of layer l may be derived using the following equations:

$$u_j^l = \sum_{i \in M_j} s_j^{l-1} k_{ij}^l + b_j^l \tag{3}$$

where s_j^{l-1} refers to the feature maps produced by the i^{th} channel in layer $l-1$; $k^l ij$ is the filter $i^t h$ filter applied on $j^t h$ channel; u_j^l refers to the output of the channel j at layer l; At each dense layer, l, the aggregated value of the neuron is passed to suitable activation function to get required nonlinearity:

$$s_j^l = f(s_j^l) \tag{4}$$

the output s_j^l of the layer l at neuron j is passed through the activation function f, ReLU at the hidden layers and softmax at the output layer.

The error function used is the categorical entropy loss function and the loss for a single instance can be expressed as:

$$Loss(x_i|\theta) = \sum_{i=1}^{C} y_i log\hat{y}_i \tag{5}$$

where y_i and $\hat{y}_i$ refer to the ground truth and predicted label associated with the sample x_i respectively; θ refers to the model parameters and C refers to the number of classes involved. Following the forward propagation of the input, the loss is computed and the weights are updated using the back propagation method, which employs the gradient descent approach to update the network weights.

Experimental Results

This section provides a detailed discussion on the experiments carried out to validate the proposed CNN architecture for recognizing the type of facial emotion. We start with a detailed summary of the datasets used for the

experimental research is provided, followed by the experimental setup and evaluation metrics used to assess the effectiveness of the proposed model for facial emotion recognition. Finally, we offer a comprehensive analysis of the outcomes of each of the experiments performed on the FER2013 benchmark dataset, as well as a comparison of the model to prior research in the associated disciplines.

Dataset

The benchmark FER2013 dataset, consisting of a total 35,887 images belonging to 7 different classes, is used for experimental studies. This dataset is developed by collecting images available on the internet extracted using the Google Image Search API. Unlike the CK+ dataset, where iamges were captured in a controlled environment, the images in this FER2013 dataset were captured in an uncontrolled environment, since all of the images were obtained from the internet. As a result, obtaining good results on this dataset is more challenging than compared to CK+ and Jaffe datasets. Another challenge in this FER2013 dataset is that the presence of cartoon images and emojis along with human facial images and hence the model has to be well generalized to make good predictions. Another observation on FER2013 dataset is that the data are rather unbalanced with the smallest class 'Disgust' containing only 547 images while the largest class 'happy' containing 8989 images. This huge unbalance also leads to a degradation of results in certain cases. The distribution of images across each emotion in the all the classes is shown in Table 1.

Performance Evaluation Measures

Different classification metrics such as Accuracy, Precision, Recall, and F_1 Score are used to determine the efficacy of the proposed approach for the emotion recognition from facial images. Accuracy is defined as the proportion of properly recognized expressions to the total number of expressions analyzed. Precision indicates the percentage of accurately anticipated emotions. The percentage of anticipated emotions that really belong to certain emotion classes is referred to as recall. The harmonic mean of precision and recall is used to get the $F1$ Score. The following are the mathematical expressions for computing each of these metrics:

$$\text{Accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})} \tag{6}$$

$$\text{Precision} = \frac{\text{TP}}{(\text{TP} + \text{FP})} \tag{7}$$

$$\text{Recall} = \frac{\text{TP}}{(\text{TP} + \text{FN})} \tag{8}$$

$$F_1\text{Score} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{9}$$

In Eqs. 6–9, TP and FP correspond to the number of true positives and false positives, respectively, while TN and FN represent the number of true negatives and false negatives, respectively.

Table 1 Emotion-wise distribution of image samples in FER2013 dataset

Emotion-type	Number of images
Disgust	547
Happy	8989
Angry	4953
Fear	5121
Sad	6077
Surprise	4002
Neutral	6198
Total	35887

Experimental Setup and Hyper Parameters

All our experiments are implemented in keras deep learning framework and run on Intel Xeon processor with NVIDIA GeForce RTX5000 GPU. The proposed CNN model for which we claim the best accuracy is trained on FER2013 and the results can be reproduced by setting the following hyper parameters. The Adam optimization technique was used to update the parameters after calculating the loss during back propagation, with a batch size of 64 and a momentum of 0.9. A fixed learning rate of 0.001 is set and was not modified during the iterations. There had been no fine-tuning, and the weight decay was set to 0.000001. We included a dropout of 0.5, to avoid over-fitting, after each conclusion layer in each convolutional block. Instead of using the random weight initialization, we use Xavier weight initialization to set the starting weight. The output dense layer is attached with a softmax activation as the model need to recognize one of the multiple emotions. Early stopping and checkpoints were employed to avoid model over-fitting. Categorical cross entropy loss, which is the most commonly used loss function when solving multi-class classification, problem was used for computing model loss. During model training, at the end of every epoch, hold-out validation strategy is followed to minimize model over-fitting. Out of the 35,887 total facial images, 28,709 are utilized for training, 3589 are used for validation, and another 3589 are used for testing.

Table 2 Model hyper parameters

Hyper parameter	Value
Convolutional blocks	5
Convolutional block	3–5 stacks of convolution and batch normalization layers + max pooling + dropout
Dropout	0.5
Epochs	Converged at 50 with early stopping
Loss	Category cross entropy
Optimizer	ADAM
Learning rate	0.000001
Filter size	4×4
Padding	Same

Result Analysis

This subsection provides a thorough discussion of the outcomes from the experimental studies carried out. This discussion helps to understand how we arrived at the proposed model for facial emotion recognition. We have experimented with different hyper parameters and different augmentation techniques, the results of which are discussed in the following subsections.

Affect of Model Configuration

This section will justify the proposed model configuration. Each convolution block of the model contains consecutive convolution operations which allows it to learn more nonlinearity hidden in the data. This convention is followed by the standard deep CNN architectures in the literature like ResNet. Then, we experimented varying the number of convoultional layers and blocks in the model and we arrived at the proposed model by using cross-validation approach. In our initial experiments, we design the model and vary the number of convolution blocks. Our empirical studies on benchmark FER2013 dataset indicate that the model with five blocks is more robust compared to the other models. In this section, we show the influence of kernel size, dense, batch normalization and dropout layers on the model. We analyze the following 5 models in this subsection in detail:

- Baseline: The baseline model, is the one designed with 5 convolution blocks, comprises of 3, 4, 5, 5, 5 convolution layers, respectively, in each of the consecutive blocks. Each convolution layer uses a kernel of size 3×3. For the experiments in this section, all the models are variants of this baseline model.
- Model 1: The baseline added with 2 dense layers at the end of convolution base with ReLU activation function.
- Model 2: The baseline model after removing the dense layers after the convolution base.
- Model 3: Model2 with kernel size of 4×4 unlike model1 and model2.
- Model 4: Model3 after including batch normalization layer after every convolution layer and dropout at the end every block, and the complete configuration of this model is shown in Table 2.

We have some interesting findings about how different configurations influence the model performance based on the outcomes of our tests and are tabulated below in Table 3. All the models shown in Table 3 follow the same configuration, except few changes in addition of the dense layers and kernel sizes.

The results shown in Table 3 indicate the performance of the models of different configurations. From the results, we can understand that removing the dense layers significantly improved model performance. This is because the model with dense layers suffered from over-fitting and was reduced after removing the dense layers and enhances the model performance. After looking at model3, we can observe that bigger kernel size helps to improve the performance further. In model4, the regularization methods introduced, helps to extend the performance further to a greater extent and also allows the model to converge faster.

Figure 2 shows the model loss over the epochs and accuracy over the epoch for both training and validation data. After looking into the loss plot, we realize that the model still suffers from over-fitting. We believe that this is due to the scarcity of number of images in certain emotion classes and hence the model could not generalize well. The next set of experiments are carried out addressing these issues of over-fitting.

Table 3 Comparison of results with different data augmentation techniques

Model	Accuracy	Precision	Recall	F-Score
Model 1	63.03	63	63	58
Model 2	65.25	65	65	65
Model 3	65.42	66	65	63
Model 4	66.37	67	66	66

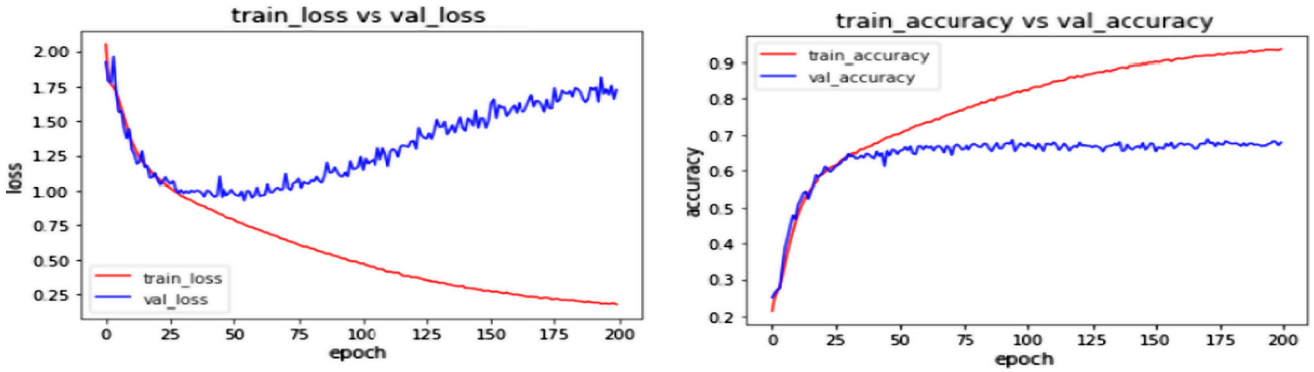

Fig. 2 Visualization of loss vs epoch and accuracy vs epoch of the baseline model

Table 4 Comparison of results with different data augmentation techniques

Model	Accuracy	Precision	Recall	F-Score
Model 1	69.57	70	70	70
Model 2	69.10	69	69	69
Model 3	66.37	67	66	66
Model 4	69.27	70	69	69

Affect of Augmentation Approaches

The models presented in the above experiments suffered from over-fitting as the number of images in few cases is scarce and hence could not generalize well. Even when we supply a sufficient quantity of data for training, the deep learning models we build face the under-fitting problem when the data for some categories are insufficient. Sometimes the data we feed have to be more diverse; otherwise, the model suffers from over-fitting even if we are feeding a large quantity of data. In this work, we use different data augmentation approaches to generate data that is much similar to the original data and this enhanced data are used for training the proposed CNN model.

We have some interesting findings about how different augmentation approaches affect the model performance based on the outcomes of our tests and tabulated below in Table 4. All the models shown in Table 4 follow the same configuration, except the augmentation techniques used on data on which the models were trained.

In $Model-1$, we randomly flipped the images horizontally and rotated them between the ranges of 0° and 10°. With these criteria, we were able to get the superior results. We then also flipped the images vertically in $Model-2$ such that the data now consist of images that are both flipped both horizontally and vertically along with the actual images and as can be seen from the above table the results dipped slightly. This maybe due to the fact that the model getting confused with the emotions displayed in the images when they are flipped vertically. In $Model-3$, we tried to set the random rotation range for the images to be between 0° and 90°. The results have fallen through a greater extent which we conclude must be because the images were getting randomly rotated over a large interval, because of which some images may get rotated to an angle of 20 while some to 40 others 80 and so on, the model has a difficult time to understand and learn the emotions from the rotated images. Finally, in $Model-4$, we didn't rotate the images at all; however, the results still dipped slightly because the data haven't been augmented in relation to the

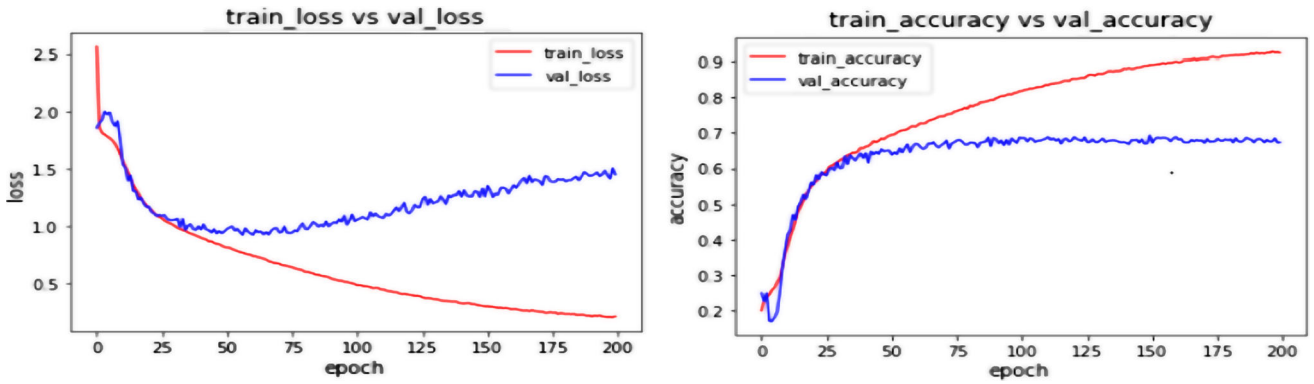

Fig. 3 Visualization of loss and accuracy of the proposed FERNet model over epochs

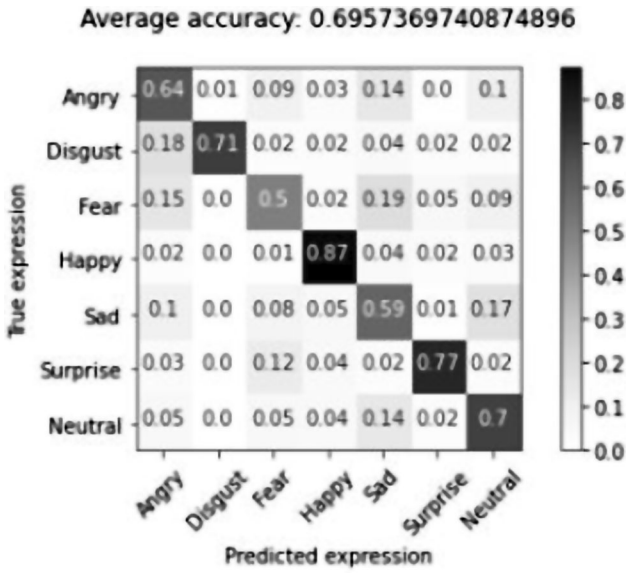

Fig. 4 Confusion matrix

rotation thus less data for the model to train on. From all these experiments, we conclude that for this task, data augmentation has to be applied as the data sets for emotion recognition task are considerably smaller in size than when compared to other vision related tasks. However, care must be taken while applying the augmentation techniques as specific techniques are suited for specific tasks. In case of emotion recognition, when we rotated the images over a large interval the results obtained were bad but while the interval was reduced to a 180° difference we were able to achieve better results.

The following image, Fig. 4 is the depiction of confusion matrix for the best model we arrived.

As can be seen from the confusion matrix, the emotions fear and sadness were being wrongly classified, while the emotions happy and surprise are being classified correctly in most instances. We believe that the reason for fear accuracy being so low has to to do with the images of the dataset under fear class. We observed the images and even as humans we sometimes wrongly classified them into other emotions. The dataset has considerable noise in some classes and the no. of images in each class is also not uniform which might lead to bias and inaccurate results.

Figure 3 depicts the loss incurred by the model and accuracy achieved over epochs for both training and validation data. After comparing the plots shown in Figs. 2 and 3, one can realize that the model trained using augmented data is more stable, smooth and leads to decrease in the loss and enhances performance. We believe that augmentation reduces model training and allows it to converge faster and results in a more robust model.

Comparison Study

Proving the efficiency of the proposed CNN model architecture, we compare it's performance to that of current state-of-the-art models for face recognition task on FER2013. In addition to the model performance, we compare model parameters to prove the simplicity of the model compared to the rest of the models.

The results and parameters provided in Table 5 reveal that our proposed model outperforms several of the current models created for the FER2013 dataset. From the results, we can understand that deep neural architectures are far better compared to the shallow models like SVM. Our proposed architecture outperforms many of the existing models that follow the architectures of VGGNet and AlexNet as baselines. Our model is not only better in performance but also have less number of parameters. This all due to the way we designed the architecture. The consecutive convolutional layers placed in the network allow the model to learn hidden nonlinearity from the input face

Table 5 Comparison of results with different data augmentation techniques

Model	Baseline	Accuracy	Parameters
Mishra [23]	SVM	63.03	–
Gan et al. [19]	VGGNet	64.24	–
Manual	–	65.00	–
Liu et al. [16]	VGGNet	65.03	84M
Wan et al. [28]	AlexNet + VGG	65.34	14M
Agarwal et al. [20]	–	65.77	0.93M
Mollahosseini et al. [17]	AlexNet	66.4	25M
Tang [27]	AlexNet	69.30	7.17M
FERNet (Proposed)	Original	69.57	2.3M

emotion images. Furthermore, dropout, batch normalization, and augmentation prevent over-fitting and improve performance, while the removal of dense layers allows to reduce the model complexity. The accuracy of the model on FER2013 dataset is slightly higher than that of human accuracy achieved which is approximately 65%.

Conclusion

In this study, we provide an unique and simple CNN architecture for identifying emotions from face images captured in the real-time with uncontrolled situations. In this paper, we introduce a novel and simple CNN architecture for recognizing emotions from the facial images taken in the unconstrained environments. Based on our experimental results, we may conclude that the model outperforms various models on the benchmark FER2013 dataset. The way the layers are placed in the model, batch normalization, dropout, weight decay allows the model to learn and improve accuracy. We have also shown the effect of augmentation and how it helps in improving results while dealing with small datasets. Moreover, we observed how critical the hyper parameters tuning is and at the same time how they can degrade the result if not chosen properly for the given task.

Funding No Funding.

Declarations

Conflict of interest The Authors declare that there is no conflict of interest.

References

1. P. Ekman, W.V. Friesen, Constants across cultures in the face and emotion. J. Personal. Soc. Psychol. **17**(2), 124 (1971)
2. J.D. Bodapati, N. Veeranjaneyulu, Abnormal network traffic detection using support vector data description. in Proceedings of the 5th international conference on frontiers in intelligent computing: theory and applications. Springer. pp 497–506 (2017)
3. K. Polisetty, K.K. Paidipati, J.D. Bodapati. Modelling of monthly rainfall patterns in the North-West India using SVM. in Ingénierie des Systèmes d'Information 24.4 (2019)
4. D. Kancherla, J.D. Bodapati, N. Veeranjaneyulu, Effect of different kernels on the performance of an SVM based classification. Int. J. Recent Technol. Eng. **5**, 1–6 (2019)
5. J.D. Bodapati et al., Joint training of twochannel deep neural network for brain tumor classification. Signal Image Video Process. **15**(4), 753–760 (2021)
6. J.D. Bodapati and N. Veeranjaneyulu, Facial emotion recognition using deep CNN based features. Int. J. Innov. Technol. Explor. Eng. (IJITEE). **8**(7), 2278–3075 (2019)
7. N. Dalal, B. Triggs (2005) Histograms of oriented gradients for human detection. in IEEE computer society conference on computer vision and pattern recognition (CVPR'05). Vol. 1. IEEE. **2005**, 886–893 (2005)
8. S. Saeed et al., Empirical evaluation of svm for facial expression recognition. Int. J. Adv. Comput. Sci. Appl. **9**(11), 670–673 (2018)
9. M. Quinn, G. Sivesind, G. Reis, *Realtime Emotion Recognition From Facial Expressions* (Standford University, Standford, 2017)
10. X. Wang et al., A new facial expression recognition method based on geometric alignment and lbp features. in 2014 IEEE 17th international conference on computational science and engineering. IEEE. 2014, pp. 1734–1737
11. K. Deepika, J.D. Bodapati, R.K. Srihitha, An efficient automatic brain tumor classification using LBP features and SVM-based classifier. In: Proceedings of International Conference on Computational Intelligence and Data Engineering. Springer, pp. 163–170 (2019)
12. J.D. Bodapati, B. Suvarna, Role of deep neural features vs hand crafted features for hand written digit recognition. Int. J. Recent Technol. Eng. (IJRTE) **7**, 147–152 (2019)
13. R. Girshick et al., Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 580–587 (2014)

14. V. Dondeti et al., Deep convolution features in non-linear embedding space for fundus image classification. Rev. d'Intelligence Artif. **34**(3), 307–313 (2020)
15. M.-I. Georgescu, R.T. Ionescu, M. Popescu, Local learning with deep and handcrafted features for facial expression recognition. IEEE Access **7**, 64827–64836 (2019)
16. K. Liu, M. Zhang, Z. Pan, Facial expression recognition with CNN ensemble. in International conference on cyberworlds (CW). IEEE. **2016**, 163–166 (2016)
17. A. Mollahosseini, D. Chan, H. Mohammad, Mahoor. Going deeper in facial expression recognition using deep neural networks. in IEEE Winter conference on applications of computer vision (WACV). IEEE. **2016**, 1–10 (2016)
18. T Connie et al. Facial expression recognition using a hybrid CNN-SIFT aggregator. In: International Workshop on Multi-disciplinary Trends in Artificial Intelligence. Springer, pp. 139–149 (2017)
19. Y. Gan, Facial expression recognition using convolutional neural network. in Proceedings of the 2nd international conference on vision, image and signal processing. (2018), pp. 1–5
20. A. Agrawal, N. Mittal, Using CNN for facial expression recognition: a study of the effects of kernel size and number of filters on accuracy. Vis. Comput. **36**(2), 405–412 (2020)
21. P. Giannopoulos, I. Perikos, I. Hatzilygeroudis. Deep learning approaches for facial emotion recognition: a case study on FER-2013. in Advances in hybridization of intelligent methods. Springer, (2018), pp. 1–16
22. T. Caramihale, D. Popescu, L. Ichim, Emotion classification using a tensorflow generative adversarial network implementation. Symmetry **10**(9), 414 (2018)
23. S. Mishra et al. Emotion recognition through facial gestures-a deep learning approach. in International Conference on Mining Intelligence and Knowledge Exploration. Springer. (2017), pp. 11–21
24. Y. Wang et al., The influence of the activation function in a convolution neural network model of facial expression recognition. Appl. Sci. **10**(5), 1897 (2020)
25. S. Singh, F. Nasoz, Facial expression recognition with convolutional neural networks. in 10th Annual Computing and Communication Workshop and Conference (CCWC). IEEE. **2020**, 0324–0328 (2020)
26. M. Shin, M. Kim, D.-S. Kwon, Baseline CNN structure analysis for facial expression recognition. in 25th IEEE international symposium on robot and human interactive communication (RO-MAN). IEEE. **2016**, 724–729 (2016)
27. Y. Tang, Deep learning using linear support vector machines. arXiv:1306.0239 (2013)
28. W. Wan, C. Yang, Y. Li "Facial expression recognition using convolutional neural network. A case study of the relationship between dataset characteristics and network performance. In: (2016)
29. J.D. Bodapati, N. Veeranjaneyulu, S. Shaik. "Sentiment analysis from movie reviews using LSTMs. in Ingenierie des Systemes d'Information 24.1 (2019)
30. J.D. Bodapati et al., Blended multi-modal deep ConvNet features for diabetic retinopathy severity prediction. Electronics **9**(6), 914 (2020)
31. P. Carcagnì et al., Facial expression recognition and histograms of oriented gradients: a comprehensive study. SpringerPlus **4**(1), 645 (2015)