I. J. Engineering and Manufacturing, **2022, 6, 40-51**
Published Online on December 8, 2022 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijem.2022.06.05

Deep Convolution Neural Networks for Cross-Dataset Facial Expression Recognition System

Rohan Appasaheb Borgalli*
Department of Electronics Engineering, Fr. Conceicao Rodrigues College of Engineering, Bandra, Mumbai 400050, India
E-mail: rohanborgalli111@gmail.com
ORCID iD: https://orcid.org/0000-0003-4159-1938
*Corresponding Author

Sunil Surve
Department of Computer Engineering, Fr. Conceicao Rodrigues College of Engineering, Bandra, Mumbai 400050, India
E-mail: surve@fragnel.edu.in
ORCID iD: https://orcid.org/0000-0002-2568-9911

Received: 28 June, 2022; Revised: 29 July, 2022; Accepted: 24 September, 2022; Published: 08 December, 2022

Abstract: Facial Expressions are a true and obvious way to represent emotions in human beings. Understanding facial expression recognition (FER) is essential, and it is also useful in the area of Artificial Intelligence, Computing, Medical, Video games, e-Education, and many more. In the past, much research was conducted in the domain of FER using different approaches such as analysis through different sensor data, using machine learning and deep learning framework with static images and dynamic sequence. Researchers used machine learning-based techniques such as the Multi-layer Perceptron Model, k- Nearest Neighbors, and Support Vector Machines were used by researchers in solving the FER. These methods have extracted features such as Local Binary Patterns, Eigenfaces, Face-landmark features, and Texture features. Recently use of deep learning algorithms in FER has been considerable. State-of-the-art results show deep learning-based approaches are more potent than conventional FER approaches.

This paper focuses on implementing three different Custom CNN Architecture training them on FER13 Dataset and testing them on CK+ and JAFFE Dataset including FER13 after fine-tuning. The three pre-trained models' on FER2013 after fine-tuning have significantly improved the accuracy of the resulting CNN on the target test sets between 65.12 % to 79.07% on the JAFFE dataset and 50.96% to 68.81% on the CK+ dataset.

Index Terms: Facial expression recognition, ensemble learning, deep learning, convolution neural networks.

1. Introduction

Facial Expression Recognition (FER) is a vital research area as it is helpful in many applications. A FER system model mentioned in this paper detects the human facial expression and identifies the corresponding induced emotion for a static image or sequence of images. In FER though much advancement has been achieved, facial expression recognition with accuracy remains challenging due to the varieties of facial expressions and their interrelation.

According to Darwin & Prodger[1], human beings' facial expressions reflect their emotional states and intentions. Due to its importance in applications such as machine vision and machine learning, numerous attempts have been made in the past to implement a FER system practically. FER system is beneficial for people to communicate in a nonverbal way with each other. The use of FER systems depends on how accurately the system detects or extracts the facial expression from an image. The system is prevalent because this could be widely used in an application that depends on FER, such as lie detection, human-machine interaction, medical assessment, driver safety, and solving other complex real-world problems with FER.

In this paper, we are addressing problem of validation of models on cross dataset by implemented three Custom CNN architectures, and we used ensemble learning in the experiment, which showed that Fine-tuning of Pretrained CNNs is effective Instead of fully training the network, which leads to the best performance on both the source and the target dataset.

The rest of our paper is organized as follows. In Section 2, learning strategies for cross-dataset recognition is described. Detailed architecture of Convolution Neural Network (CNN) is mentioned in Section 3. Section 4 gives overview of dataset used for experiment. Section 5 presents the experiments and performance analysis. Data generator and augmentation is mentioned in Section 6. Various results obtained such as confusion matrix, model training, validation accuracy, loss, emotion analysis and accuracy are summaries in Section 7, 8, 9 and 10 respectively. Conclusions and future work are drawn in Section 11.

2. Learning Strategies for Cross-Dataset Recognition

In facial expression, recognition supervised machine learning, and deep learning approaches have shown promising results in the past. In this approach, many datasets are available for research. These datasets are generally divided into training, validation, and test data. To train the model, training data is used, and validation data is used to validate model performance after every training cycle to get an idea about how well the model is trained. Finally, once a model is adequately trained, it is tested on the unseen test dataset. For supervised learning, a large number of training data is required to get good accuracy, but the available dataset has a less number of labeled data. However, such datasets are inadequate for the effective training of the model. Also, mostly when we train a model on a particular dataset, we get good accuracy on the same kinds of test images. But, it should be performing better on all types of test images. To overcome the problem of lack of data and ineffective training of the model, cross-dataset learning strategies were used. Where there is a minimum of two datasets are required, one of which was used for model training called a source database and, The other called a target data set to which the model has to be tested to verify performance. Tannugi et al. [2] proposed deep learning networks to learn additional features that can be transferable for maintaining memory integrity to train CNNs using transfer learning or fine-tuning is not effective. Instead, training CNNs fully from scratch is better to achieve good performance on both source and target datasets.

HUA et al. [3] proposed a deep facial recognition algorithm HERO based on CNN and an ensemble learning algorithm to predict facial expression. They proposed an algorithm that combines neural networks and ensemble learning by connecting three sub-networks with different structures to get benefits of both to achieve algorithm stability, good test accuracy, and standardization.

XIA et al. [4] implemented an effective face recognition system based on standard CNN architecture Inception-v3 model on the TensorFlow platform. They used a transfer learning method on the Inceptionv3 model to retrain learned facial data, reducing training time as much as possible.

Corneanu et al. [5] surveyed FER systems based on RGB, 3D, thermal, and multiple modalities for Automatic Facial expressions recognition (AFER). Related work on a historical variation of FER methods and an in-depth discussion on the various components is analyzed with the latest trends. In addition, they provide an introduction to facial expression from an evolutionary point of view.

Wanget al. [6] mentioned features of a few key parts of the face, such as the eyes, nose, and mouth, are most commonly used to judge the final facial expression, while other elements play a minor role in the final result. To solve this problem, they proposed a novel framework based on CNN as a subregion auxiliary, which fully utilizes the three key regions and adjusts the learning outcomes of the main task by setting different parameters to improve the final accuracy level.

To implement FER, there are well-established methods with both handcrafted and automated features extraction through deep learning [7]. The CNN-based [8] approach has proved suitable for image classification-based applications. The use of standard CNN architectures gives a state of the art results in FER [7]. The main advantage of a CNN is that it enables end-to-end learning directly from an input source and helps to completely remove or highly reduce the dependency on models or other pre-processing techniques used for FER application.

Kim et al. [9] proposed a Deep Learning CNN-based model which utilizes Action Units (AUs) to demonstrate FER. They experimented with the CNN model trained on CK+ Dataset and classifies emotion based on extracted features. This CNN-based model classifies the multiple AUs with the help of extracted features and emotion classes, and the experiment shows that the CNN model predicted emotion classes from only features and generated AUs.

Minaee et al. [10] proposed deep learning based on visualization technique attentional convolutional network, which can focus on key parts of the face to find face regions for detecting different emotions, to achieve significant improvement in accuracy over other CNN based models on multiple datasets, including FER-2013, CK+, JAFFE, and FERG. They finally visualized the classifier output detecting different emotions based on important face regions.

In our approach for cross-dataset facial expression recognition, we used three custom CNN with different levels of complexity, trained on a dataset that belongs to a domain similar to a target domain. Considering that the source dataset FER-2013 [11] and the target datasets FER-2013, CK+ [12], and JAFFE [13] belong to the facial expression recognition domain, we trained three custom CNN models on the FER-2013 dataset, and accuracy is calculated in two cases, first when a model is trained from scratch and then when a model is trained on FER-2013 and tested on CK+ and JAFFE with fine-tuning.

3. Architecture of Convolution Neural Network (CNN)

A typical architecture of a CNN contains an input layer, some convolutional layers, some fully-connected layers, and an output layer. Fig. 1 shows the general Convolution Neural Network (CNN) Architecture for facial expression recognition.

The CNN architecture we are using takes a 48×48 gray-level image as input. Various convolutional layers interchanged with few max-pooling/average pooling layers and few fully connected layers. Rectified linear units (ReLU) and batch normalization was used after each convolutional layer in some models.

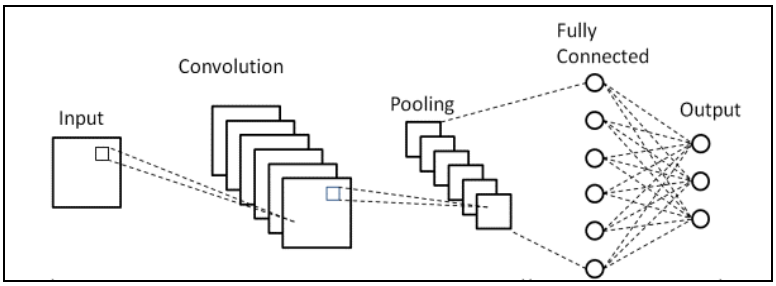


Fig. 1. Architecture of CNN

We used three different CNN models to compare their results and crossed different datasets for each approach. Details of them are mentioned below.

In this paper, we are using the following three models with different CNN architecture.

Model 1 is simple consist of 2 Convolution layer, 2 Maxpooling layers, and 3 Dense layers with RELU activation after each Convolution and Dense layer and Softmax activation at the end.

Model1		
Layer (type)	**Output Shape**	**Param #**
conv2d_1 (Conv2D)	(None, 48, 48, 64)	1664
max_pooling2d_1	(MaxPooling2 (None, 24, 24, 64)	0
conv2d_2 (Conv2D)	(None, 24, 24, 128)	73856
max_pooling2d_2	(MaxPooling2 (None, 12, 12, 128)	0
flatten_1 (Flatten)	(None, 18432)	0
dense_1 (Dense)	(None, 1024)	18875392
dropout_1 (Dropout)	(None, 1024)	0
dense_2 (Dense)	(None, 1024)	1049600
dropout_2 (Dropout)	(None, 1024)	0
dense_3 (Dense)	(None, 7)	7175
Total params: 20,007,687		
Trainable params: 20,007,687		
Non-trainable params:0		

Model 2		
Layer (type)	**Output Shape**	**Param #**
conv2d_1 (Conv2D)	(None, 44, 44, 64)	1664
max_pooling2d_1(MaxPooling2	(None, 20, 20, 64)	0
conv2d_2 (Conv2D)	(None, 18, 18, 64)	36928
conv2d_3(Conv2D)	(None, 16, 16, 64)	36928
average_pooling2d_1(Average)	(None, 7, 7, 64)	0
conv2d_4 (Conv2D)	(None, 5, 5, 128)	73856
conv2d_5 (Conv2D)	(None, 3, 3, 128)	147584
average_pooling2d_2(Average)	(None, 1, 1, 128)	0
flatten_1 (Flatten)	(None, 128)	0
dense_1 (Dense)	(None, 1024)	132096
dropout_1 (Dropout)	(None, 1024)	0
dense_2 (Dense)	(None, 1024)	1049600
dropout_2 (Dropout)	(None, 1024)	0
dense_3 (Dense)	(None, 7)	7175
Total params: 1,485,831		
Trainable params: 1,485,831		
Non-trainable params: 0		

Model 2 is a moderately complex model consisting of 5 Convolution layers, 1 Maxpooling layer, 2 Averagepooling layers, 3 Dense layers, and 2 Dropouts after dense layer with RELU activation after each Convolution and Dense layer and Softmax activation at the end.

Model 3 is a complex model consisting of 6 Convolution layers, 4 Maxpooling layers, 4 Dense layers, 7 Dropouts after each Convolution layer and dense layer, 7 Batchnormalization after each Convolution layer with RELU activation after each Convolution and Dense layer, and Softmax activation at the end.

Model 3		
Layer (type)	**Output Shape**	**Param #**
conv2d_1 (Conv2D)	(None, 46, 46, 64)	640
conv2d_2 (Conv2D)	(None, 46, 46, 64)	36928
batch_normalization_1 (Batch)	(None, 46, 46, 64)	256
max_pooling2d_1 (MaxPooling2)	(None, 23, 23, 64)	0
dropout_1 (Dropout)	(None, 23, 23, 64)	0
conv2d_3 (Conv2D)	(None, 23, 23, 128)	73856
batch_normalization_2 (Batch)	(None, 23, 23, 128)	512
conv2d_4 (Conv2D)	(None, 23, 23, 128)	147584
batch_normalization_3 (Batch)	(None, 23, 23, 128)	512
max_pooling2d_2 (MaxPooling2)	(None, 11, 11, 128)	0
dropout_2 (Dropout)	(None, 11, 11, 128)	0
conv2d_5 (Conv2D)	(None, 11, 11, 256)	295168
batch_normalization_4 (Batch)	(None, 11, 11, 256)	1024
conv2d_6 (Conv2D)	(None, 11, 11, 256)	590080
batch_normalization_5 (Batch)	(None, 11, 11, 256)	1024
max_pooling2d_3 (MaxPooling2)	(None, 5, 5, 256)	0
dropout_3 (Dropout)	(None, 5, 5, 256)	0
conv2d_7 (Conv2D)	(None, 5, 5, 512)	1180160
batch_normalization_6 (Batch)	(None, 5, 5, 512)	2048
conv2d_8 (Conv2D)	(None, 5, 5, 512)	2359808
batch_normalization_7 (Batch)	(None, 5, 5, 512)	2048
max_pooling2d_4 (MaxPooling2)	(None, 2, 2, 512)	0
dropout_4 (Dropout)	(None, 2, 2, 512)	0
flatten_1 (Flatten)	(None, 2048)	0
dense_1 (Dense)	(None, 512)	1049088
dropout_5 (Dropout)	(None, 512)	0
dense_2 (Dense)	(None, 256)	131328
dropout_6 (Dropout)	(None, 256)	0
dense_3 (Dense)	(None, 128)	32896
dropout_7 (Dropout)	(None, 128)	0
dense_4 (Dense)	(None, 7)	903
Total params: 5,905,863		
Trainable params: 5,902,151		
Non-trainable params: 3,712		

4. Dataset Overview

We used the FER2013, CK+, and JAFFE datasets for our experiment. These datasets have seven basic classes: anger, disgust, fear, happiness, sadness, surprise, and neutral. But, the CK+ dataset also contains an extra contempt class and seven basic emotions. The number of image samples of every class in each dataset is given in Table 1.

Table 1. The Distributions of Every Class in Each Dataset

	FER-2013	CK+(last frame)	JAFFE
Angry	4,593	45	30
Disgust	547	59	30
Fear	5,121	25	31
Happy	8,989	69	31
Sad	6,077	28	31
Surprise	4,002	83	30
Neutral	6,198	327	30
Contempt	0	18	0

In our experiment, to compare results using three Custom CNN architectures. Initially, pre-trained CNN on the dataset that belongs to the FER domain. These custom CNN is, therefore, trained and optimized on a source dataset (FER dataset) as it has large number of images that is good to train CNN model from scratch as illustrated in Fig. 2 (a) & (b). Accuracy was calculated for the pre-trained model on the source dataset and the target dataset with fine-tuning.

Fig. 2. (c) approach requires a pre-trained CNN on the FER database. The CNN model was fine-tuned on a target dataset (JAFFE or CK+). Target dataset are popular dataset in FER domain and as they have less number of images hence selected as target dataset. However, the CNN was initialized with the parameter values learned previously on the source dataset (FER). Also, we are comparing results with fully trained and tested CNN from scratch on the same dataset.

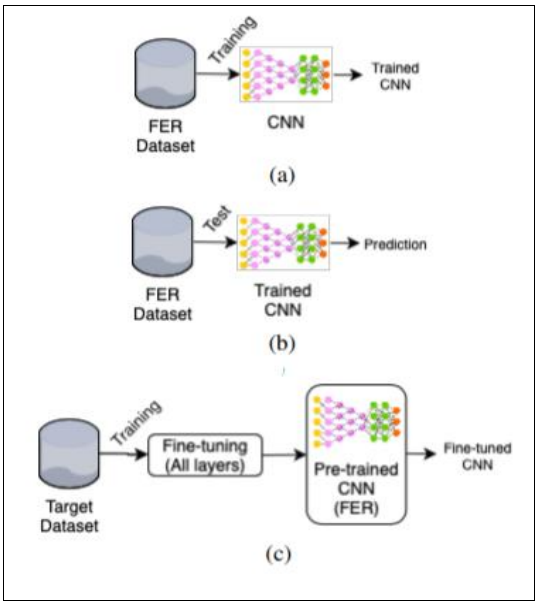


Fig. 2. The baseline CNN is: (a) trained on the training set of the FER dataset (source dataset) and; (b) evaluated at the test set of the same source dataset. (c) fine-tuned all layers on the Target dataset.

Therefore, our experiment has two parts. First, we are training our custom CNN model twice, with two different datasets, Initially on the source dataset and then fine-tuning on the target dataset to preserve the parameters learned from the source dataset [1] In the second part, we are training and testing three Custom CNN models on the same dataset. This learning strategy was used to carry out experiment for cross-dataset recognition to test it validity and efficiency. As model trained on source dataset, how well work on target dataset after fine-tuning instead of training them from scratch.

5. Experiments and Performance Analysis

In this section, the performance of our three custom CNN models for the facial expression recognition task was analyzed. Keras framework was chosen to build a model and carry out experiments.

A. FER2013 dataset [11]

The FER2013 dataset is available on Kaggle for the FER challenge. The database contains gray faces for women of all ages. Each image in a dataset has a size of 48x48 pixels. The database was divided into seven categories (anger, disgust, fear, happiness, sadness, surprise, and neutrality) contains 28,709 training, 3,589 validation, and 3,589 test images. The database contains face images of samples from all ages and a posed variety of directions, including a cartoon face.

FER2013 is an excellent dataset to be used for FER. A data generator (given in section 6) is used on the FER2013 dataset to balance the data distribution and generate more training data. First, we trained three custom CNN Models, which are stated previously, independently on the FER2013 dataset with training epoch was set to 75, and the Adam optimizer is used to reduce the cross-entropy loss function.

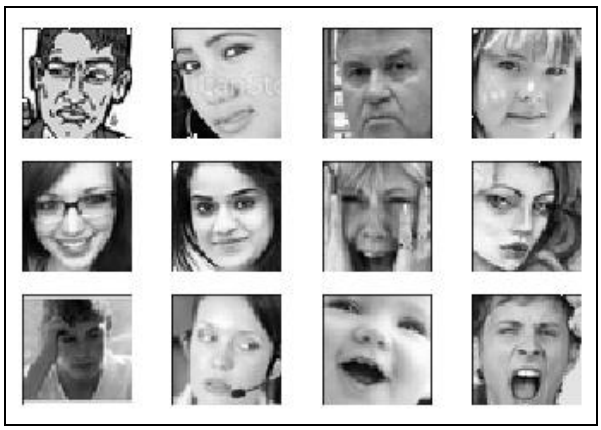


Fig. 3. FER13 Dataset Sample Images

We observed that by training these custom CNN models from scratch on the FER2013 dataset, the accuracy of each model is close to 62%. That is a good result for custom CNN-based networks. After fine-tuning, these models have been tested on the CK+ and JAFFE (test) datasets. Comparision shows that we achieved good accuracy on the test set after fine-tuning the model compared with the models trained from scratch. The recognition accuracy increased significantly. Table 5 compares the performance between models on the FER dataset.

B. CK+ dataset [12]

The Extended Cohn-Kanade Dataset (CK +), released in 2010, extends the Cohn-Kanade (CK) database, increasing the sequence and number of courses by 22% and 27%, respectively. The database includes 593 video sequences of 640 × 480 frames of digital images from neutral emotions to peak emotions labeled with FACS-coded emotion for the peak emotion frames. These image frames have both posed and non-posed (spontaneous) expressions—the 123 subjects in this database range. From 18~ to 50 years of age (81% Euro-American, 13% Afro-American, 6% other races), 69% of whom are women. In addition to the seven basic facial expressions, another emotion class, contempt, is labeled in a database to give eight facial expressions. But, as we focus on seven basic emotions like feelings of contempt are close to disgust, we have turned contempt labels into disgust. And considered only seven basic emotions of the CK+ Dataset. Fig. 4 Shows CK+ dataset sample images.

CNN model input layer size is 48x48x1. We first standardized the size of all images to 48x48x1. Then, to balance the data distribution and generate more training data we used a data generator (in section 6). We trained and fine-tuned the three models using the generated train dataset. Finally, we tested the model's performance on the test dataset of CK+ using the fine-tuned model on the FER13 dataset and from scratch. The recognition accuracy of each model is significantly better for a fine-tuned model. The comparison between fine-tuned model on the FER13 dataset and from scratch is shown in Table 7.

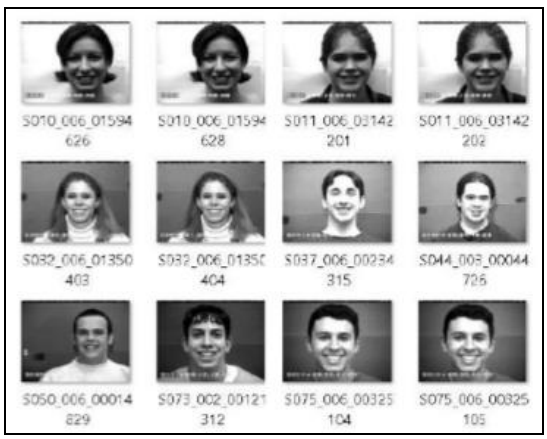


Fig. 4. CK+ Dataset Sample Images

C. JAFFE dataset [13]

The JAFFE dataset consists of seven Japanese female models 213 grayscale posed facial images with seven facial expressions (anger, disgust, neutrality, sadness, surprise, happiness, and fear). Each image size is 256x256 pixels. The JAFFE dataset is divided into three sections: the training set, the validation set, and the test set. contains 129 images (60%), 42 images (20%), and the 42 images (20%) respectively. Michael Lyons, Jiro Gyoba, and Miyuki Kamachi of Kyushu University created a dataset. It is a facial database useful for many facial recognition activities. CNN model input layer size is 48x48x1. We convert the input image size to 48x48x1. Then, we used the Data Generator (provided in section 6) to measure data distribution and generate additional training data. We have trained and optimized the three models using the generated train database. Fig. 5 shows JAFFE database sample database images.

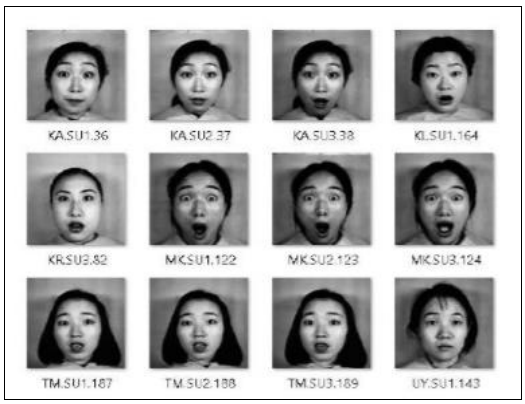


Fig. 5. JAFFE Dataset Sample Images

Finally, we tested the performance of the model on the test dataset of JAFFE using the fine-tuned model on the FER13 dataset and from scratch. The recognition accuracy of each model is significantly better for fine-tuned Models. Table 6 compares between fine-tuned model on FER13 Dataset and from scratch.

6. Data Augmentation and Data Generator

Our experiments use three custom CNN models with the FER2013 dataset, the JAFFE dataset, and the CK+ dataset. These datasets are viral for FER and widely used by related researchers. The training datasets of these three datasets consist of anger, disgust, fear, happy, sad, surprise, and neutral as seven basic facial expressions categories. However, the number of samples is usually uneven, as shown in Table 1. For example, in the FER2013 dataset, the number of images of disgust is 547, while the other class's number of images ranges from 4,000-to 9,000. The imbalance of data distribution will harm network performance while training the model. To eliminate this adverse effect, we performed data augmentation, as shown in Fig. 6 that performs a different operation on a given image, such as flipping, rotation, cropping, blurring, and adding noise to generate more images.

The algorithm used the DataGeneration operation to generate more training data and balance the data distribution.

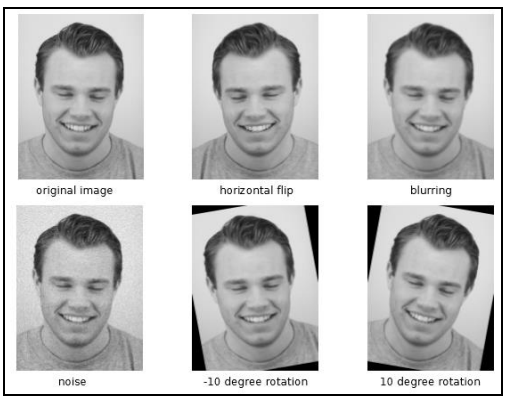


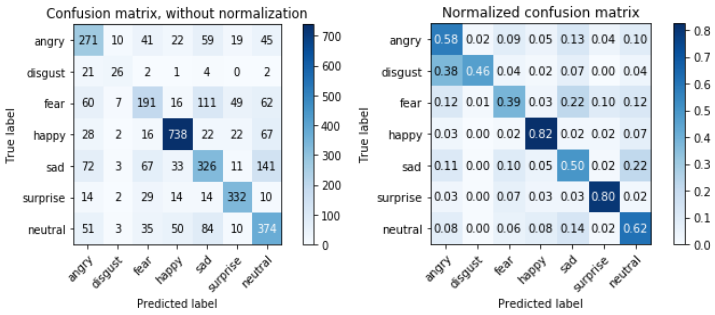
Fig. 6. Illustrating the different augmentation techniques used to increase the number of training images.

7. Confusion Matrix

The confusion matrix, a tabular representation also known as an error matrix, is a specific table layout that indicates the performance of an algorithm in the classification task. Each row of the confusion matrix represents the predicted class,

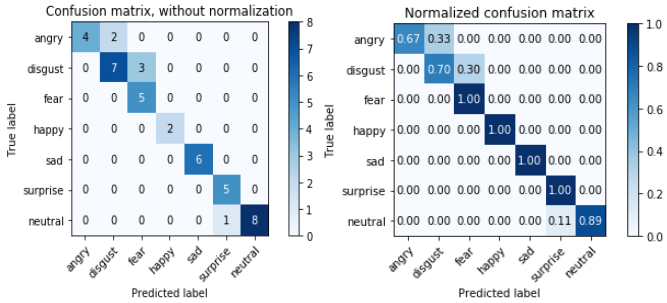
while each column represents the actual class (or vice versa). As the name suggests, this shows whether a FER system is confused between recognizing two classes (i.e., commonly predicting one class as another).confusion matrix of the recognition accuracy for seven different facial expressions is shown in Table 2 for the FER13 dataset, Table III for JAFFE dataset, and Table 4 for CK+ dataset. It has two parts, one without normalization on the left and normalized on the right. FER2013 shows that the recognition accuracy of anger, disgust, fear, and sadness is lower than the final recognition accuracy for the particular model tested on the test dataset. For the FER13 dataset, for training, validation, and test set, we use around 28,709, 3,589, and 3,589 images, respectively. Table 2 shows the confusion matrix indicating that one facial expression class is easily confused with other facial expression classes. For example, disgust is usually confused with anger, and fear is confused with sadness. The reason for the same is the training data is not enough. The training data for disgust is less than others on the training set. The method achieves high recognition accuracy on happy, surprise, and neutral because the three facial expressions have easily differentiable, and the training data is more.

Table 2. Confusion Matrix for FER13 Dataset



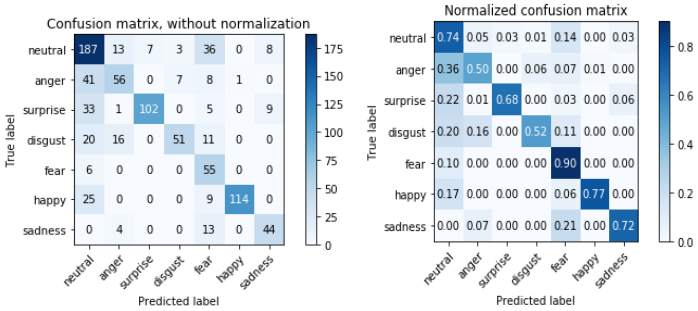
For the JAFFE dataset, Table 3 shows the confusion matrix of the predicted results with training, validation, and testing set, 130, 41, and 42 images, respectively.

Table 3. Confusion Matrix for JAFFE Dataset



For CK+, Table 4 shows the confusion matrix of the predicted results with training, validation, and test set has 70%, 10%, and 20% of the images, respectively.

Table 4. Confusion Matrix for CK+ Dataset



8. Model Training, Validation Accuracy, and Loss

We have built the model but would like to validate it by inducing different datasets. Usually, we train the model on training samples (seen samples) in terms of the amount of accurate data we have for training. But the validating model is also necessary to rely on the model based on its evaluation through validation samples (unseen Samples). We evaluate the

trained model on the validation dataset before testing it on the test dataset.

A learning curve shown below is the training and validation accuracy and loss of a learned model for training and validation samples of a particular dataset. This graph is handy for validating the model's performance while training and can be decided until we train a model. If both the training and validation loss converge to a too low value without increasing the validation set's accuracy, we will stop training the model. The following Fig. 7, 8, and 9 shows model training, validation accuracy, and loss for FER13, JAFFE, and CK+ datasets for a particular model, respectively. These curves are smooth for FER2013 as they have more samples to train than others.

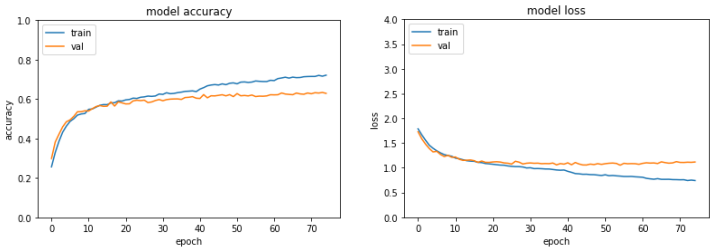


Fig. 7. Model Training, Validation Accuracy and Loss for FER13 Dataset

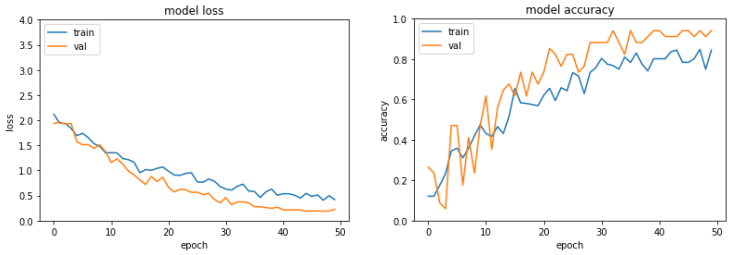


Fig. 8. Model Training, Validation Accuracy and Loss for JAFFE Dataset

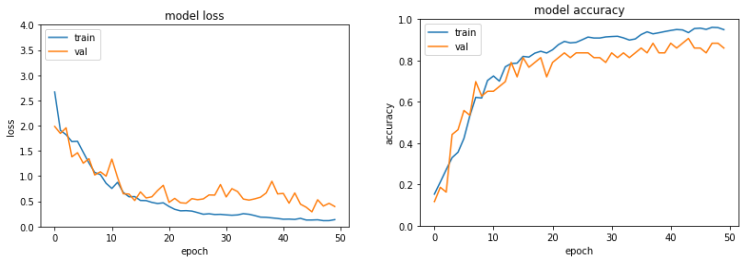


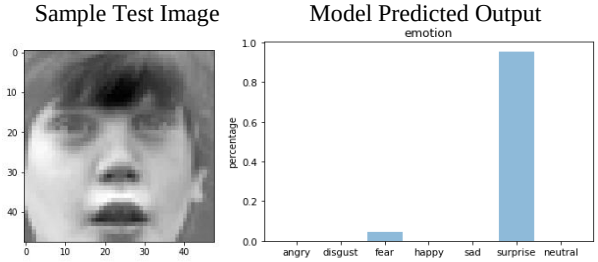
Fig. 9. Model Training, Validation Accuracy and Loss for CK+ Dataset

9. Emotion Analysis

Our Facial Expression Recognition (FER) System is trained on a different dataset and tells whether the underlying emotion behind a facial image is: anger, disgust, neutral, sad, surprise, happiness, or fear. It has two sections on the left sample image from a dataset and on the right emotion type of emotion present on an object.

Facial expression (angry, disgust, neutral, sad, surprise, happy, and fear) with probabilities that indicate the recognition score of each of the emotions indicates confidence score. It lies between 0 to 1. A higher score states the model is predicting particular emotion with a higher confidence score.

Fig 10 shows Emotion Analysis of Sample test Images from FER13, JAFFE, and CK+ datasets. And it gives a good probability value for correct emotion, indicating the accurate classification of emotion through analysis.



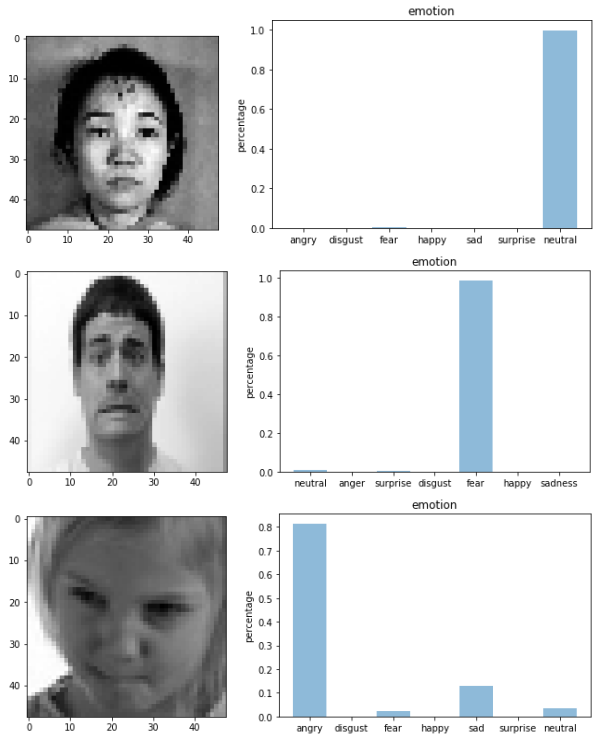


Fig. 10. Emotion Analysis of Sample test Images from FER13, JAFFE, and CK+ Dataset

10. Experimental Results

We have evaluated the three custom CNN Models training from scratch and with the pre-trained model on FER2013 as the source dataset and after fine-tuning it on JAFFE and CK+ target dataset. Table 5 shows the results in terms of accuracy on the test set of FER13 for all three models, and it shows each model gives nearly the same accuracy of 62%; this indicates that the accuracy is dependent on the size of the dataset used for training as FER2013 has a large size. The CNN model trained more precisely and achieved good classification on test data in this dataset. However, by looking at the accuracy trained from scratch and cross-dataset results using FER as a source of data and fine-tuning it in targeted data sets: JAFFE and CK +.

(1) The performance of models trained and tested on the FER13 dataset achieves nearly the same accuracy of 62% on each model in Table 5.
(2) Training CNN models on the target dataset from scratch gives accuracy between 60.47% to 74.42% on the JAFFE dataset and 46.89 % to 60.90% on the CK+ dataset.
(3) Performance on the target test sets after fine-tuning the CNN: The three models' fine-tuning pre-trained model on FER2013 have significantly improved the accuracy of the resulting CNN on the target test sets between 65.12 % to 79.07% on the JAFFE dataset and 50.96% to 68.81% on the CK+ dataset.

Therefore, fined tuned CNN models are most effective and give the best performances on the target datasets instead of training them from scratch.

Table 5. Results on FER13 Dataset

MODEL	ACCURACY
Model 1	61.19 %
Model 2	61.58 %
Model 3	61.62 %

Table 6. Results on JAFFE Dataset

MODEL	ACCURACY trained from scratch	ACCURACY With Fine Tuning
Model 1	74.02%	76.74%
Model 2	74.42%	79.07%
Model 3	60.47%	65.12 %

Table 7. Results on CK+ Dataset

MODEL	ACCURACY trained from scratch	ACCURACY with Fine Tuning
Model 1	46.89 %	50.96 %
Model 2	50.73%	68.81 %
Model 3	60.90%	61.45%

11. Conclusion and Future Work

In this paper, an ensemble-based deep recognition algorithm was proposed for which three CNNs-based models with different structures were trained independently. The structures of the three models we adopted to experiment on the FER13, JAFFE, and CK+ datasets were simple to complex. Because of the huge data scale of the FER13 dataset, deeper neutral models were effective in making up the ensemble model of the FER13 dataset. Mentioned accuracy indicates Test accuracy. Each model was trained using the ensemble method. The test sample's emotion is predicted and compared with a true label during the test phase. The results of the experiments on the FER2013 are summarised in Table 5. JAFFE, and CK+ datasets are not ideal for training from scratch as they have fewer samples. Hence we tested them of fine-tuned CNN model and model trained from scratch, indicating that our proposed fine-tuned CNN models have relatively good performance compared with models trained from scratch. However, there is room for improvement.

Note that cross-database testing in the JAFFE database is more challenging than on the CK + database due to the nature itself. It is also challenging to use CNN to classify a separate database, especially if the interclass variation is not dispersed as a source database that is used for training. This is one of the limitations of using CNN, defined as "data sensitivity" in an image classification task.

In the future, For FER, we are considering an unsupervised pre-training strategy from transfer learning, which may further reduce the level of classification error. Pre-processing and Feature Extraction techniques can be applied before training models [14]. Transfer Learning [15] can also be tried on these datasets CK+ and JAFFE Dataset have very few images as they are not balanced, so making them balanced will give better testing results as provided in the survey [14]. FER based on Facial Action Units (AUs) has shown promising results in the state of the art that can also be tried [16]. Some Advanced models such as Deep Belief Network (DBN) and Generative Adversarial Networks (GANs) can be applied to improve accuracy [17,18].

References

[1] Darwin, C., & Prodger, P. (1998). The expression of the emotions in man and animals. Oxford University Press, USA.
[2] Tannugi, D.C., Britto Jr, A.S. and Koerich, A.L., 2019. Memory Integrity of CNNs for Cross-Dataset Facial Expression Recognition. arXiv preprint arXiv:1905.12082.
[3] Hua, W., Dai, F., Huang, L., Xiong, J. and Gui, G., 2019. HERO: Human emotions recognition for realizing intelligent Internet of Things. IEEE Access, 7, pp.24321-24332.
[4] Xia, X.L., Xu, C. and Nan, B., 2017. Facial expression recognition based on tensorflow platform. In ITM Web of Conferences (Vol. 12, p. 01005). EDP Sciences.
[5] Corneanu, C.A., Simón, M.O., Cohn, J.F. and Guerrero, S.E., 2016. Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications. IEEE transactions on pattern analysis and machine intelligence, 38(8), pp.1548-1568.
[6] Wang, Y., Li, Y., Song, Y. and Rong, X., 2019. Facial Expression Recognition Based on Auxiliary Models. Algorithms, 12(11), p.227.
[7] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25, 1097-1105.
[8] Caramihale, T., Popescu, D. and Ichim, L., 2018. Emotion classification using a tensorflow generative adversarial network implementation. Symmetry, 10(9), p.414.

[9] Kim, S. and Kim, H., 2019, April. Deep Explanation Model for Facial Expression Recognition Through Facial Action Coding Unit. In 2019 IEEE International Conference on Big Data and Smart Computing (BigComp) (pp. 1-4). IEEE.
[10] Minaee, S. and Abdolrashidi, A., 2019. Deep-Emotion: Facial Expression Recognition Using Attentional Convolutional Network. arXiv preprint arXiv:1902.01019
[11] Wolfram Research, "FER-2013" from the Wolfram Data Repository (2018)
[12] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression," 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, 2010, pp. 94-101.
[13] Lyons, Michael, Kamachi, Miyuki, & Gyoba, Jiro. (1998). The Japanese Female Facial Expression (JAFFE) Dataset [Data set]. Zenodo.
[14] S. Li, W. Deng, Deep facial expression recognition: a survey, arXiv preprint.(2018) arXiv:1804.08348.
[15] Ramalingam, S. and Garzia, F., 2018, October. Facial Expression Recognition using Transfer Learning. In 2018 International Carnahan Conference on Security Technology (ICCST) (pp. 1-5). IEEE.
[16] Kim, S., & Kim, H. (2019, February). Deep explanation model for facial expression recognition through facial action coding unit. In 2019 IEEE International Conference on Big Data and Smart Computing (BigComp) (pp. 1-4). IEEE.
[17] Liu, P., Han, S., Meng, Z., & Tong, Y. (2014). Facial expression recognition via a boosted deep belief network. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1805-1812).

Authors' Profiles

Mr. Rohan Appasaheb Borgalli, currently pursuing Ph.D. under Fr. Conceicao Rodrigues College of Engineering, Mumbai University. He received his M.Tech degree in Digital Systems from Motilal Nehru National Institute of Technology (MNNIT), Allahabad, in 2013. Working as an Asst. Prof. in Shah & Anchor Kutchhi Engineering College, Mumbai. His research interests include artificial intelligence, machine learning, human computer interaction and deep learning.

Dr. Sunil Surve, completed his PhD from VJTI, Mumbai in robotics and artificial intelligence. Working as a Professor in Fr. Conceicao Rodrigues College of Engineering, Mumbai. He published more than 30 research papers in various journals and conferences. His research interests include machine learning, robotics and computer architecture.

How to cite this paper: Rohan Appasaheb Borgalli, Sunil Surve, "Deep Convolution Neural Networks for Cross-Dataset Facial Expression Recognition System", International Journal of Engineering and Manufacturing (IJEM), Vol.12, No.6, pp. 40-51, 2022. DOI:10.5815/ijem.2022.06.05