Received: May 8, 2020. Revised: June 23, 2020.

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

EMONET: A Cross Database Progressive Deep Network for Facial Expression Recognition

Michael Moses Thiruthuvanathan[1]* **Balachandran Krishnan[1]**

[1]*Department of Computer Science and Engineering, School of Engineering and Technology,*
CHRIST (Deemed to be University), Bangalore, India
* Corresponding author's Email: Michael.moses@christuniversity.in

Abstract: Recognizing facial features to detect emotions has always been an interesting topic for research in the field of Computer vision and cognitive emotional analysis. In this research a model to detect and classify emotions is explored, using Deep Convolutional Neural Networks (DCNN). This model intends to classify the primary emotions (Anger, Disgust, Fear, Happy, Sad, Surprise and Neutral) using progressive learning model for a Facial Expression Recognition (FER) System. The proposed model (EmoNet) is developed based on a linear growing-shrinking filter method that shows prominent extraction of robust features for learning and interprets emotional classification for an improved accuracy. EmoNet incorporates Progressive- Resizing (PR) of images to accommodate improved learning traits from emotional datasets by adding more image data for training and Validation which helped in improving the model's accuracy by 5%. Cross validations were carried out on the model, this enabled the model to be ready for testing on new data. EmoNet results signifies improved performance with respect to accuracy, precision and recall due to the incorporation of progressive learning Framework, Tuning Hyper parameters of the network, Image Augmentation and moderating generalization and Bias on the images. These parameters are compared with the existing models of Emotional analysis with the various datasets that are prominently available for research. The Methods, Image Data and the Fine-tuned model combinedly contributed in achieving 83.6%, 78.4%, 98.1% and 99.5% on FER2013, IMFDB, CK+ and JAFFE respectively. EmoNet has worked on four different datasets and achieved an overall accuracy of 90%.

Keywords: Facial expression recognition, Deep neural network, Emonet, Progressive resizing, Cross database.

1 Introduction

Human face is an index that portrays tangible information on the current emotional state of a person. Emotions are the superlative characteristics that a human being exhibits, to indicate the present state of mind. Facial expressions begin at an early age and evolves progressively to perfect the art of expressing emotions as they get old. Though, there are millions of faces we see over the years, every human tends to express facial emotions in the same manner and also it is quite obvious for every human to understand the emotions exhibited by a face. Facial expression makes it possible to understand instantly, if a person is happy, sad, angry, surprised, etc. Studies of assessment of physiognomy and facial expressions are very old and is found that people began to analyse facial movements during the Aristotelian era [1].

Face of a human-being expresses emotions based on inherited coordination of fixed, fragmentary movements of the face muscle that result in the emotional expressions. Research on human facial expression analysis, can be trailed to Pre-Darwinian period. While emotions are inherent characteristics, prone to change through peer influence and experience, it is a non-trivial process to map these emotional features into a computer-based system that can learn and interpret the expressions of a human. Recent advances in the field of technology and research has encouraged researchers, to develop machine learning algorithms to comprehend human's current emotional state. An automated expression identification system can link references to emotions from a face into a computationally intelligent system. Artificial intelligent systems have the capability to understand human feelings and respond accordingly. The current focus of researchers is to develop emotional intelligent computing model, that will help in the fields of behavioural science, psychology, telecommunications, instructional technology, automotive security and human computer interaction.

FER has attracted researchers in computer vision, due to the capability of interpreting images into emotional understanding. The aim of FER is to detect faces and predict the primary emotions such as Anger, Sadness, Happy, Neutral, Surprise, Disgust and Fear. Emotional Analysis plunges deep into Mental state identification, Safety, Automatic counselling, Face expression synthesis, Lie detection, Music Recommender systems, Automated Tutoring, Operator Fatigue Detection, Student Feedback, etc. Marechal et al. have used CNN for a facial expression recognition system and by incorporating Artificial Intelligence (AI) helped the authors in identifying facial expressions better than the conventional methods [2]. Feed-forward neural network or multilayer Perceptron with various hidden layers are generally regarded as Deep Neural Networks (DNN).

Facial Expression Recognition is a data-driven task, that makes training and detecting expressions among a group of people a challenging arena. FER must be in a position to handle challenges such as illumination changes, variations in size of images, occlusions, non-frontal image pose, identity bias, age range, cultural & demography differences and low intensity expression. The above-mentioned categories require a vast collection of accurately labelled data for training and testing process. Accurate classification is important though there are various factors affecting the precision of detection. The Proposed method is suitable to detect and classify facial images into 7 emotions using a supervised DNN model that can be used in emotional classifications that can help in identifying the emotional state of a human in various scenarios where an analysis can help in improving the satisfaction rate. This paper focusses on achieving,

(1) A comprehensive learning network that is trainable for facial expression analysis using Progressive Learning.

(2) A distinct recognition model for arbitrary expression and arbitrary pose for a cross database validation approach.

(3) A State-of-the-art recognition efficiency for facial expression on FER2013, IMFDB, CK+ and JAFFE dataset.

The rest of the paper is organised as follows, Section 2 describes the related work and Section 4 explains the Datasets. Section 4 explains the Proposed Work. The detailed discussions of Results are in Section 4. Finally, Section 5 concludes the paper.

2 Related work

The works carried out by authors in [3-6] used histogram equalization method to normalize the values between 0 and 255 to enhance the contrast among training dataset. This process is in-efficient if the brightness of the foreground and background have similar intensity values. However, using these methods the value in the data is linearized. Brightness and contrast in the image dataset are symmetrical leading to drastic degradation of the outcome and Histogram equalization often produces unrealistic effects when applied to images with low color depth. The weighted summation approach to combine histogram equalization and linear mapping to address issues, in how the symmetry of data needs to be used for higher recognition was addressed by the authors in [7], however the issues with weighted summation model is that a small change in in weights may results in big changes in the objective vectors. The model used in [4], works on Worldwide standardization of comparison, Localization and Equalization of histograms to achieve the better accuracy during training and testing phase where Global Contrast Normalization (GCN) played a major role in handling inaccuracies. This Paper deals with methods of pre-processing and how to help in improving image pre-processing, emphasize was not given on feature extraction and classification.

“AdaBoost” and Viola-Jones approaches for face detection are commonly used, to detect and identify face, eyes, nose, mouth, etc., Comparison of faces were based on Euclidean distances, the major drawback being normalization [8]. In the model of facial expression recognition, the authors in [9] used the neutral face of a person, extracting features and CNN is used to train a facial expression system with ease instead of using hand-crafted features. Facial features using landmarks can determine eyes, nose, ears, mouth play a significant role in facial expression analysis. This work was limited to few data sets and also has not considered real time data. Two models BKVGG12 and BKVGG10 were used by authors for Facial Expression recognition using deep CNN however the model was able to perform at 71% accuracy. The inaccuracies may cause the model to perform even lesser when introduced to newer data [10]. Gerard and David assembled CNNs, to improve the capability of detecting facial expression. This

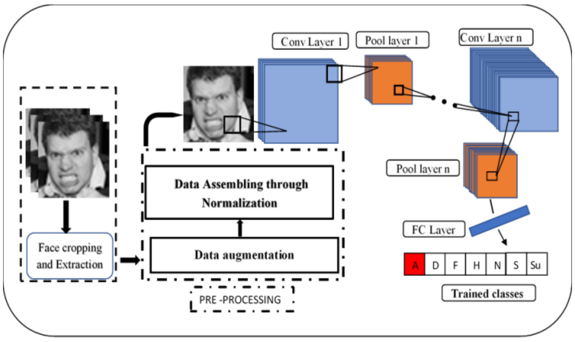

Figure. 1 The general pipeline of deep facial expression recognition systems

involved fine-tuning hyper-parameters of CNN assembled network for a Four-layer trainingarchitecture comprising of 72 CNN's with distinct primary parameters. Though optimization was carried out in the networks the model was able to achieve 42% accuracy [11].

Xie and Hu, have proposed a model based on Feature redundancy-reduced convolutional neural network, which results in producing less redundant features to get a dense illustration of an image. In the case of FER, classes disgust, fear and Neutral are differentiated based on very subtle changes in expressions. By using less features and increasing the CNN architecture will affect the classification accuracy with respect to change detection and time [12]. Siamese networks were used for managed pooling along with CNN for discriminative image depiction. Facial expression reflects physical weakness, condition of fitness and health. Moving the muscles on a face relates to an emotion. This model has worked with detecting fatigue and has used Supervised Descent Method (SDM). SDM's fails to achieve results when the faces have large head poses and extreme expressions in unparametrized situations. [13].

A model to track faces that are closer to the camera and the faces away from the camera using ResNet model was introduced by authors by incorporating a 3d-inception ResNet layers followed by LSTM models to extract features such as spatial relationship and neighbourhood representation in each frame. This technique was found to be a way more agile than traditional methods. ResNet model uses bigger sized filters and is prone to ignore smaller faces in an image [14]. A survey is available for the methods and datasets that are used on FER. These methods are based on visible facial expressions. Iqbal et al., proposed a local descriptor named Neighbourhood-aware Edge Directional Pattern (NEDP) was introduced to enable feature descriptions on weak and distorted edges using Sobel operators. Sobel detectors are generally weak performers when noise is introduced. [19].

This work focusses on three important phases training, model development and testing. In the training phase inclusive adaptations through progressive learning framework is used that utilizes data from various datasets that comprises of different sizes, illumination, variations in face registrations and increased number of subjects. In the model development a Deep CNN is designed based on the parameters listed in Table 1. The parameters are tuned to and tested with unknown data to achieve best accuracy. This work contributes to enhance the efficiency of detecting emotions on a face by extracting required features for learning on varied face images so as to enable the model to be strong in classifying while exposed to a newer data. This work also takes into consideration the facial data classification. Since majority of the facial data are generally sized between 32x32 and 128x128, a model is required to handle the facial data with the sizes mentioned. Majority of the prevalent models are designed to work with sizes over 128x128 which causes the data handling insufficiency.

3 Dataset

Kaggle's FER2013 Dataset [15] comprises of 48x48 pixel gray-scale facial expression images. The faces were automatically registered so that the face is more or less centred in each image, it occupies about the same amount of space. The job is to categorize each face in one of seven categories based on the emotion shown in the facial expression (Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral). The training set consists of 31,759 examples. The Validation data consists of 3,589 images and the test set consists of 3,813 images. Indian Movie Face database (IMFDB) is a large free facial database comprising of 34,512 pictures of 100 Indian artist collected from more than 100 videos. All the images are physically selected and cropped from the video frames resulting in a huge variety of pose, scale and expressions [18].

The JAFFE database contains 213 images for 7 facial expressions (6 basic facial expressions + 1 neutral) postured by 10 Japanese female models [16]. The CK+ dataset consists of 717 images from 123 subjects. [17]. Recent advancements, frequently used databases are listed in a Deep facial Expression Survey by Li in 2018[26].

4 Proposed network

In general, there are two steps in the DNN (Deep Neural Network) framework. The first step is to extract the features *f* of the image and map it into a Feature Map (F_m). The input for each feature is a neuron and it is connected to the local receptive fields of the previous layer. Upon extraction of local features, the positional link between them and other features are determined. The second step is to compile all the feature maps. Each Feature map is a plane, the size of the plane cell is equal. The feature map is constructed based on a sigmoid function and serves to be an activation function for the network. Furthermore, the neurons share weight in the same mapping plane, the amount of free network parameter are decreased. Each layer of convolution in the neural network is followed by a computing layer that calculates the local average and the secondary derivative. This distinctive two-function extraction technique decreases the dimensional complexity.

Improving the architecture of Convolutional Neural Network is done by increasing the number of neurons and depth of the layers. Often doing this leads to increased requirement of computation power. With the advent of GPU based systems it is possible to carry out higher computation requirements. However, learning complexity with reference to training and validation tends to either over-fit or under-fit. To overcome this problem, a sparse network should be designed with the derived parameters that replicates human biological networking system. The proposed system is developed based on Convolutional Neural Networks that primarily classifies face into emotions. To highlight the work done for Facial Emotion Recognitions, the process is divided into three modules, namely i) Pre-processing. ii) Developing the EmoNet Architecture and iii) Training, Testing and Validation.

Images are prepared for training and testing phases through Pre-processing operations. Histogram equalization is carried out on the images to reduce variations in illumination, lighting, contrast and brightness. Face registration is an important process, as the emotions are derived from the face. All images in the dataset has a single face and are introduced to affine transformation, in which. angle of variations range between -45° to +45°. In this model, a Progressive Resizing (PR) method is used to handle different face sizes. Using this method, it is very convenient to reduce the time taken for resizing all images to a particular fixed size. Fig. 1, illustrates the general procedural pipeline involved in FER.

Face Registration varies with every dataset due to the large distribution of facial subjects. Extensive researches on image handling and robust comparative scheme is essential. Learning semantic features through every pixel is a challenging arena and would create vast feature maps that are redundant and cause the network to overfit. Large volume of data from each of the dataset are used for Learning and Validation. However, a network should be feasible to perform well on external data to cross boundaries of dependency and enhance cross database efficiency.

Architecture: CNN is one of the most widely used technique for computer vision application, and has been tested by various researchers for its robustness and performance. It is a system that can detect, acquire and interpret features acquired from an image, to classify emotions accurately. The core requirement of the application is to improve its efficiency. The architecture is designed with a fully connected network having to distinguish seven classes. The CNN network architecture was built to extract maximum neurons which consists of 10 convolutional layers, 4 Max Pooling Layers and 2 Average Pooling Layers. Along with these is a fully connected layer after all the convolutions, and a SoftMax layer ahead of the classification layer to classify the emotions. The Window size (Ws) chosen for the network is 3x3 for the initial layers and 5x5 for the later layers in the network. This has helped to increase scaling within a single training routine. The CNN network is introduced to Non-linearity using the leaky ReLU (Rectified Linear Unit). The number of convolutional filters used are 8, 16, 32, 64, 128 respectively. Max-Pooling supports in translation invariance and reduces the computation complexity in the deeper layers. Since, the model uses multiple filters on the same image to extract features, convolution for each of them is carried out discretely and the results are loaded one above the other.

The illustrations of individual layers are given in Fig. 2. Eq. (1), explains the above-mentioned process, in which: n is the size of the image, p is the padding size, f is the size of the filter, nc is the number of channels, nf is the number of filters and s is the stride.

$$[n.n, nc] * [f, f, nc] = \left[\left[\frac{n+2p-f}{s}+1\right]\left[\frac{n+2p-f}{s}+1\right], nf\right] \tag{1}$$

Based on Eq. (1) convolution is carried out on each image based on *n, nc* and *f,* preserving the relation between pixels and creating a matrix of feature maps. *s* is used to shift the filter over the image pixels and *p.*

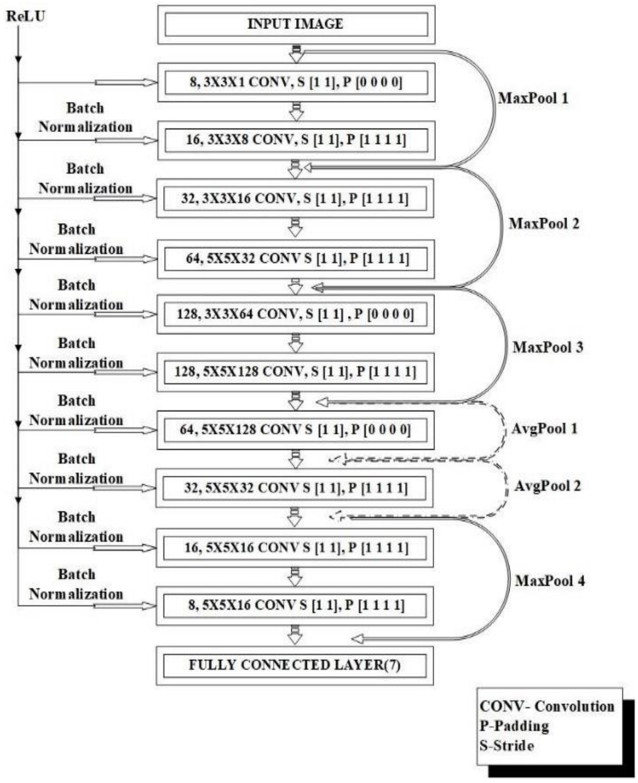
ReLU
INPUT IMAGE
8, 3X3X1 CONV, S [1 1], P [0 0 0 0]
Batch Normalization
16, 3X3X8 CONV, S [1 1], P [1 1 1 1]
MaxPool 1
Batch Normalization
32, 3X3X16 CONV, S [1 1], P [1 1 1 1]
Batch Normalization
64, 5X5X32 CONV S [1 1], P [1 1 1 1]
MaxPool 2
Batch Normalization
128, 3X3X64 CONV, S [1 1] , P [0 0 0 0]
Batch Normalization
128, 5X5X128 CONV, S [1 1], P [1 1 1 1]
MaxPool 3
Batch Normalization
64, 5X5X128 CONV S [1 1], P [0 0 0 0]
AvgPool 1
Batch Normalization
32, 5X5X32 CONV S [1 1], P [1 1 1 1]
AvgPool 2
Batch Normalization
16, 5X5X16 CONV S [1 1], P [1 1 1 1]
Batch Normalization
8, 5X5X16 CONV S [1 1], P [1 1 1 1]
MaxPool 4
FULLY CONNECTED LAYER(7)
CONV- Convolution
P-Padding
S-Stride

Figure. 2 Emoet architecture

$$E_R(\theta) = E(\theta) + \lambda\Omega(w) \qquad (2)$$

$E_R(\theta)$, is the regularization loss function. where w is the weight vector, λ is the factor for regularization (coefficient), and $\Omega(w)$ Regularization Function is given by,

$$\Omega(w) = l_2 w^T w \qquad (3)$$

The l_2 Regularization is used in this model to enhance stability and produce one single solution. Since the l_2-norm squares the error, it is sensitive and adjusts the model to minimize error. The network used Stochastic Gradient Descent. Stochastic gradient descent algorithm helps in improving the learning rate to oscillate along the steepest downward route to the optimum value. This may lead to underfitting issues. However, a momentum controller is introduced to reduce this oscillation and enhance smoothness in the learning curve. The Stochastic Gradient Descent with Momentum SGDM can be written as,

$$\theta_{l+1} = \theta_l - \alpha\Delta E\ (\theta_l) + \gamma\ (\theta_l - (\theta_{l-1})) \qquad (4)$$

Where, l is the iteration number, α is the learning rate, θ is the parameter vector, E(θ) is the loss function and γ determines the contribution of θ at the present iteration to the prior gradient phase. $\Delta E(\theta)$ is evaluated using the entire dataset. The underfitting issues are controlled by the momentum of SGDM and Overfitting issues is regularized by the Error function. After each epoch weights are updated, the weight updating factor for each epoch is given by,

$$w_j = w_j - l_r \frac{\partial x}{\partial w_j} \qquad (5)$$

where, l_r is the learning rate and w_j is the Weight updating factor. $w_{j,}$ is updated every epoch and the learning rate also gets updated based on the networks learning update. The objective of a pooling layer is to provide a spatial variance to enhance detection rate of the object being inspected. The Pooling layer functions on a down sampling method along the spatial dimensions (width, height). Drop out layers have been used only when $E_R(\theta)$ is not converging towards zero.

A fully connected layer helps in reducing the last convolution layer's output and connect each of the node in the present layer to the next layer node. Neurons in a fully linked layer consists of broad links ahead of all layers' activation functions. There are algorithms that can optimise functions however, the derivative in the function should be able to change itself rapidly in the positive direction. The process is cumbersome and inefficient if it is to be performed on each iteration and the effect of the learning gradient is minimal. Hence, the weight upgradation is done using Mini-Batch (B_m). The approximation is performed on a minibatch data and the derivative acquired is used to update the weights. The Learning Rate (L_r) hyper parameter was set to 0.001 and the size of the mini-batch (B_s) is 64. Pooling layers help

Table 1. Parameter description

Parameter's Name	**Value**
Networks Layers	40
InitialLearnRate(Lr)	1.00E-03
Regularization Function	l_2Regularization
GradientThreshold	'l2norm'
MaxEpochs	40
MiniBatchSize(B_s)	64
Verbose Frequency	50
Validation Frequency (V_f)	500
Shuffle	'every-epoch'
Padding Direction	'right' (1,1)
Filter Size	(3 x 3), (5 x 5)
Stride	(1,1)
Number of Filters(layer)	8,16,32,64,128
Optimiser	SGDM

to provide spatial variance, which merely implies that the system will be able to recognize an object as an object, even when the object appears to be variant at a certain angle. The pooling layer will conduct a down-sampling procedure along spatial measurements (width, height), to a sampled output [16 x 16] for a pooling size (2,2).

Algorithm 1: EmoNet Training Algorithm

```
Data: Training data ($(X_i,Y_i)$), i = 1, 2, · · · , n;
$X_i$ is the $i^{th}$ image and $Y_i$ is the corresponding class label.
Pre-Processing: Histogram Equalization on all $X_i$
    Result: Classified Image;

Declare: $f_i$, $E_R(\theta)$, $\Omega(w)$, $w_j$, $L_r$, $B_s$, $V_i$, $F_{mi}$, Vf, Lr
    For $X_i$ in $Y_i$ do
        Introduce padding p for all $X_i$;
        While $X_i$ in $Y_i$ do
            j=i;
            Perform Convolution on $X_i$ based on $f_i$;
            Use Max pooling to capture features;
            Construct $F_{mi}$;
            Calculate $E_R(\theta)$;
                If $E_R(\theta) \rightarrow 0$, then
                    Associate trained $F_{mi}$ into $Y_i$;
                    Update $V_i$ and $T_i$;
                Else
                    Initiate dropout;
                End
                At $V_{fi}$ Perform Validation;
                Update $w_j$;
                Increment i for X;
        End
        Increment i for Y;
    End
```

The network is equipped with 40 layers consisting of input layer, Convolutional layers, max pool layers, ReLu Layers, Batch Normalization layers, SoftMax layer and a classification layer. Every time an image is fed into the network, features are extracted and the process of learning begins. Since there are several convolutional layers in the network, the feature acquired over an epoch is huge due to which there is a probability that the system can lead into an over fitting problem by memorizing objects and may not yield good results when the data is unknown. Also, the validation loss usually reduces after a number of epochs and then starts to reduce further based on the learning rate of the network. In order to overcome such issues precisely tuning the loss function is important, the number of training data, and the number of epochs are important. The network is designed to train for 40 epochs while many of the existing models are trained for over 100 epochs and yield lesser results. The developed Network is configured to use 70%, 20 % and 10% of the images in each dataset for Training, Validation and Testing respectively

The training samples are shuffled and taken at random after every epoch to reduce over-fitting or under-fitting issues. The Training Accuracy(T) and Validation Accuracy(V) is plotted and visualized for the efficiency index of the model after every iteration. The Validation frequency (V_f) is an important parameter as it tests the model's training efficiency in regular intervals. The value for V_f is set for every 500 iterations for this model. A weighted average of the scores independently derived from individual layers using the posterior class probabilities is cumulated to improve Learning rate and reduce Validation Loss. These weights are then trained for face image on cross dataset, which helps in reducing the blindness of the model to newer data. This helps in breaching into cross dataset borders by effectively incorporating progressive resizing of image during training and testing. However, while comparing with the other methods as in the Table 2, Emonet performs with an accuracy higher than the majority of the work carried out earlier.

In many of the dataset's individual classifications, there have been instances where emotions have been wrongly classified, however individualistic and holistic approach will enhance the model's performance. To shed light on this classification model the following metrics Accuracy, Precision, Recall, Specificity and F1-score is calculated for EmoNet on the dataset and the results are furnished in Table 2. The number of epochs used to train the network was set to 40. The network has undergone numerous trial and error methods to fix on the hyperparameters. Moreover, decisive conclusions in the number of iterations were based on the Loss function.

5 Results and discussions

All the experiments were carried out using an Intel Core i5 8250U 1.8 GHz with 8GB of RAM and a NVIDIA GeForce MX150 with 4GB of memory. Images from the testing data is drawn at random and fed into the network, these images are tested for True Positive, false positive, True negative or false negative.

Based on the confusion matrix calculations, for each dataset the following parameters Accuracy, precision, sensitivity, Specificity, and F1-score are calculated for each class. Table 2, enumerates the results of the

Table 2. Metric table

Data		***Acc%***	***Pre%***	***Sen%***	***Spe%***	***F1%***
FER2013	AR	93.3	75.4	80.7	95.4	77.9
	DT	99.5	98.4	80.3	100	88.4
	FR	91.8	70.9	76.0	94.5	73.4
	HY	94.7	92.9	89.3	97.1	91.1
	NL	91.9	82.6	77.3	95.7	79.8
	SD	91.2	74.5	77.6	94.2	76.1
	SE	96.7	87.0	87.4	98.1	87.2
IMFDB	AR	94.6	69.5	79.4	96.2	74.1
	DT	92.0	71.5	80.6	94.0	75.8
	FR	99.0	99.2	92.3	99.9	95.6
	HY	89.5	83.0	76.8	94.2	79.8
	NL	86.0	81.3	75.0	91.4	78.0
	SD	92.3	67.2	75.6	94.7	71.2
	SE	96.9	71.3	79.7	98.0	75.3
CK+	AR	98.1	92.2	95.0	98.6	93.6
	DT	97.8	91.3	92.3	98.6	91.8
	FR	97.2	88.2	91.1	98.1	89.6
	HY	99.0	97.6	94.2	99.7	95.9
	NL	97.2	88.2	91.1	98.1	89.6
	SD	97.5	89.5	92.4	98.3	90.9
	SE	98.2	94.8	92.9	99.1	93.8
JAFFE	AR	100.0	100.0	100.0	100.0	100.0
	DT	99.4	96.0	100.0	99.4	98.0
	FR	98.6	93.1	96.4	98.7	94.7
	HY	99.4	100	96.2	100.0	98.0
	NL	100.0	100.0	100.0	100.0	100.0
	SD	99.4	100.0	96.0	100.0	98.0
	SE	99.4	96.0	100.0	99.4	98.0

[0] *AR: Anger, DT: Disgust, FR: Fear, HY: Happy, NL: Neutral, SD: Sad, SE: Surprise. Acc: Accuracy, Pre: Precision, Sen: Sensitivity, Spe: Specificity, F1: F1-Score.*

network performance on various dataset. The class Neutral and Fear are classified incorrectly in majority of the dataset. Distinction in the training and testing face for the class Fear, Neutral and Sad requires a clear distinction. Table. 3, gives a detailed report on the various methods used on these datasets and the result obtained by them. Though facial occlusion, change in pose, and closeness of data were significantly present in the dataset, EmoNet is able to classify new test data at a substantial percentage. In CK+ dataset it is observed that model used by Zhang et al. is performs 0.8% more than Emonet. Though Bias reduction and generalization of the images was taken care while developing EmoNet, the ambiguity of facial expressions in the images are the reason for the lesser detection rate.

The model has used small sized Filters and Kernels, by doing so there is a significant improvement in capturing of smaller and complex local features from the images. For an image which has 40 x 40 pixels the Fm created are (n-1) * (n-1) features which is 1521 features using a 3x3 filter and (n-3) * (n-3) which is 1369 features. The above-mentioned feature extraction is for one CNN layer, as the number of filter size increases the layers tends to miss on acquiring smaller and important features. This process has resulted in improved weight sharing process by 9% and acquiring dense features for learning and testing modules. The Loss Regularization and automated dropout layer association helps regulate the learning of features and improves the model's classification accuracy by 2%.

The trained feature-map extracted are massive, resulting in the higher accuracy of classification. The network training is stopped when there is no significant change in the variation of the loss function. This max epochs for the model was fixed as 40 due to the multiple early stopping method used for training. The Early stopping of training was selected to ideally stop the training at an epoch where the L_r has no changes. The model was designed from the scratch by choosing the parameters individually and adding them into the network based on fine tuning and hyper parameter optimization. The derived model was initially started with 68 layers and gradually brought down to 40 layers, which produced the most appropriate and optimized results.

The reduction of layers was done through trial and error methods to achieve the objective of creating a framework solely for the purpose of detecting emotions with a smaller network compared to the existing models. VGG16 has 41 layers, VGG19 comprises of 47, resnet50 contains 177 layers, resnet18 is created with 72 layers, resnet101 has 372 layers and densenet201 has 709 layers. Since, the other models encompass large number of layers, and the size of image inputs for individual networks are different. EmoNet is designed to overcome these issues by achieving significant results when compared with the existing models by using smaller number of layers and requiring smaller sized facial images. In addition to this, introduction to cross database made the network viable to learn new features and hence improved the prediction capability of the network even in imperceptible environments. Emonet is a compact model that performs well with known and unknown environments.

6 Conclusions

This paper focus on creating a model that can detect emotions accurately for a Facial Emotion Recognition system. This model is compatible for facial emotional recognition systems. Deep Neural Networks is the crux of the model and drives the classification system to a good accuracy rate. The model is able to perform better when it is introduced to cross database, this is evaluated during training and classification. The model is also able to adapt and recognize various sizes of the faces, illumination and angle of face registration. Progressive resizing helps in assimilating data with different sizes and reduces the need for resizing the data after each epoch. Progressive resizing has helped the network to improve the networks ability to learn from 63% to 68% during the validation process in FER2013 dataset. Though there are State of the art Networks that can be used through transfer learning, image size, image orientation, cross datasets and computer resource utilization are some factors that needs to be tackled. EmoNet is able to work well with 40 layers, when compared with other models that have above 100 layers as mentioned in Section 5. Small sized filters (3x3,5x5) and training, validation on cross datasets, have made viable to improve the efficiency of classification by 8% on FER2013 and 0.2% in JAFFE. This model performed classification on unknown 3589 images and classified them into seven classes at 2.77 seconds. EmoNet was created to be dedicatedly used for Facial Emotional Classification. Hence making it a network that works at an average of 90%. Further, this model can be used in real time applications for emotional analysis for videos and various applications pertaining to Facial Emotional recognitions.

Table 3. Performance summary of static methods for facial expression recognition

Dataset	*Author*	*Accuracy%*
FER2013	Tang et.al 13[20]	71.2
	Devries et al. 14[21]	67.2
	Zhang et al. 15[22]	75.1
	Guo et al. 16[23]	71.3
	Kim et al.16[24]	73.7
	Pramerdorfer et al.16[25]	75.2
	EmoNet	83.6
JAFFE	Liu et al.14[27]	91.8
	Hamester et al. 15[28]	95.8
	Siyue Xie18[16]	99.3
	EmoNet	99.5
IMFDB	EmoNet	78.8
CK+	Ouellet14 [29]	94.4
	Li et al. 15 [30]	96.8
	P. Liu et al. 14 [27]	96.7
	M. Liu et al. 13 [31]	92.1
	M. Liu et al. 15 [32]	93.7
	SiyueXie 18[16]	83.9
	Khorramiet al.15 [33]	95.1
	Ding et al.17 [34]	96.8
	Zeng et al. 18[35]	93.7
	Cai et al. 17 [36]	90.6
	Meng et al. 17 [37]	95.4
	Liu et al. 17 [38]	96.1
	Yang et al. 18[39]	96.5
	Zhang et al. 18[40]	98.9
	EmoNet	98.1

Conflicts of Interest

The authors declare no conflict of interest.

Author Contributions

The contributions by the authors for this research article are as follows: “conceptualization, methodology, Formal analysis, data curation, visualization and writing—original draft preparation Michael Moses Thiruthuvanathan; Result validation, data curation, resources, formal analysis writing—review and editing and supervision Balachandran Krishnan;"

Acknowledgments

Authors wish to acknowledge the technical and infrastructural help rendered by the faculty members from the department of Computer Science and Engineering, CHRIST (Deemed to be University), Bangalore, India.

References

[1] Ooommen and T. Oomen, "Physiognomy: A Critical Review". *Journal of Anatomical Society of India*, Vol. 52, No. 2, pp. 189–191, 2003.

[2] Marechal, D. Miko, K. Tyburek, P. Prokopowicz, L. Bougueroua, and C. Ancourt, "Survey on AI-Based Multimodal Methods for Emotion Detection", *High-Performance Modelling and Simulation for Big Data Applications,* Springer International Publications, Vol. 11400, pp. 307-324, 2019.

[3] Z. Yu and C. Zhang, "Image Based Static Facial Expression Recognition with Multiple Deep Network Learning", In: *Proc. of the 2015 ACM on International Conf. on Multimodal Interaction,* Seattle, USA, pp.435–442, 2015.

[4] A. Pitaloka, A. Wulandari, T. Basaruddin and D. Y. Liliana, "Enhancing CNN with Pre-processing Stage in Automatic Emotion Recognition", In: *Proc. of the 2nd International Conf. on Computer Science and Computational Intelligence*, Bali, Indonesia, pp. 523-529, 2017.

[5] S. E. Kahou, V. Michalski, K. Konda, R. Memisevic, and C. Pal, "Recurrent Neural Networks for Emotion Recognition in Video", In: *Proc. of the 2015 ACM on International Conf. on Multimodal Interaction,* Seattle, USA, pp. 467–474, 2015.

[6] S. A. Bargal, E. Barsoum, C. C. Ferrer, and C. Zhang, "Emotion Recognition in the Wild from Videos using Images", In: *Proc. of the ACM International Conf. on Multimodal Interaction,* Tokyo, Japan, pp. 433–436, 2016.

[7] C.-M. Kuo, S.-H. Lai, and M. Sarkis, "A Compact Deep Learning Model for Robust Facial Expression Recognition", In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition Workshops*, Salt Lake City, USA, pp. 2121–2129, 2018.

[8] S. Shakya, S. Sharma and A. Basnet, "Human Behaviour Prediction using Facial Expression Analysis", In: *Proc. of International Conf. on Computing, Communication and Automation*, Noida, India, pp. 399-404, 2016.

[9] Bandrabur, L. Florea, C. Florea and M. Mancas" Emotion Identification by Facial Landmarks Dynamics Analysis", In: *Proc. of IEEE International Conf. on Intelligent Computer Communication and Processing*, Cluj-Napoca, Romania, pp. 379-382, 2015.

[10] Sang, N. Dat & D. Thuan, "Facial Expression Recognition using Deep Convolutional Neural Networks", In: *Proc. 9th International Conference on Knowledge and Systems Engineering (KSE)*, pp. *130-135*, 2017.

[11] G. Pons and D. Masip, "Supervised Committee of Convolutional Neural Networks in Automated Facial Expression Analysis", *IEEE Transactions on Affective Computing*, Vol. 9, No. 3, pp. 343-350, 2018.

[12] S. Xie and H. Hu, "Facial Expression Recognition with FRR-CNN", *Electronics Letters*, Vol. 53, No. 4, pp. 235-237, 2017.

[13] M. A. Haque, R. Irani, K. Nasrollahi and T. B. Moeslund, "Facial Video-Based Detection of Physical Fatigue for Maximal Muscle Activity", *IET Computer Vision*, Vol. 10, No. 4, pp. 323-329, 2016.

[14] B. Hasani and M. H. Mahoor, "Facial Expression Recognition Using Enhanced Deep 3D Convolutional Neural Networks", In: *Proc. IEEE Conf. on Computer Vision and Pattern Recognition Workshops,* Honolulu, Hawaii, pp. 2278-2288, 2017.

[15] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D-H. Lee, Y. Zhou, C. Ramaiah, F. Feng, R. Li, X. Wang, D. Athanasakis, J. S-Taylor, M. Milakov, J. Park, R. Ionescu, M. Popescu, C. Grozea, J. Bergstra, J. Xie, L. Romaszko, B. Xu, Z. Chuang, and Y. Bengio "Challenges in Representation Learning: A Report on Three Machine Learning Contests", *Neural Networks*, Vol. 64, No. 1, pp. 59-63, 2015.

[16] M. J. Lyons, S. Akamatsu, M. Kamachi and J. Gyoba, "Coding Facial Expressions with Gabor Wavelets", In: *Proc. of the 3rd IEEE International Conf. on Automatic Face and Gesture Recognition,* Nara, Japan, pp. 200-205, 1998.

[17] 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", In: *Proc. of IEEE Computer*

Society Conf. on Computer Vision and Pattern, San Francisco, California, pp. 94-101, 2010.

[18] S. Setty, M. Husain, P. Beham, J. Gudavalli, M. Kandasamy, R. Vaddi, V. Hemadri, J. C. Karure, R. Raju, B. Rajan, V. Kumar and C. V. Jawahar, "Indian Movie Face Database: A Benchmark for Face Recognition Under Wide Variations", *In: Proc. of the 4th National Conf. on Computer Vision, Pattern Recognition, Image Processing and Graphics,* Jodhpur, India, pp. 1-5, 2013.

[19] M. T. B. Iqbal, M. Abdullah-Al-Wadud, B. Ryu, F. Makhmudkhujaev and O. Chae, "Facial Expression Recognition with Neighborhood-Aware Edge Directional Pattern (NEDP)", *IEEE Transactions on Affective Computing*, Vol. 11, No. 1, pp. 25-137, 2020.

[20] Y. Tang, "Deep Learning Using Linear Support Vector Machines", In: *Proc. of International Conf. on Machine Learning: Challenges in Representation Learning Workshop,* Atlanta, USA, pp. 1-6, 2013.

[21] T. Devries, K. Biswaranjan, and G. W. Taylor, "Multitask Learning of Facial Landmarks and Expression", In*: Proc. of IEEE Canadian Conf. on Computer and Robot Vision,* Regina, Canada, pp. 98–103, 2014.

[22] Z. Zhang, P. Luo, C.-C. Loy, and X. Tang, "Learning Social Relation Traits from Face Images", In: *Proc. of the IEEE International Conf. on Computer Vision*, Santiago, Chile, pp. 3631–3639, 2015.

[23] Y. Guo, D. Tao, J. Yu, H. Xiong, Y. Li, and D. Tao, "Deep Neural Networks with Relativity Learning for Facial Expression Recognition", In: *Proc. of IEEE International Conf. on Multimedia and Expo Workshops,* Seattle, USA, pp. 1-6, 2016.

[24] B. Kim, S. Dong, J. Roh, G. Kim, and S. Lee, "Fusing Aligned and Non-aligned Face Information for Automatic Affect Recognition in the Wild: A Deep Learning Approach", In: *Proc. of IEEE* Conf. *on Computer Vision and Pattern Recognition Workshops,* Las Vegas, Nevada, pp. 1499-1508, 2016.

[25] V. Sang, N. V. Dat and D. P. Thuan, "Facial Expression Recognition using Deep Convolutional Neural Networks", In: *Proc. of 9th International Conf. on Knowledge and Systems Engineering,* Hue, Vietnam, pp. 130-135, 2017.

[26] S. Li and W. Deng, "Deep Facial Expression Recognition: A Survey", *IEEE Transactions on Affective Computing,* 2020.

[27] P. Liu, S. Han, Z. Meng, and Y. Tong, "Facial Expression Recognition via a Boosted Deep Belief Network", In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition,* Columbus, USA, pp. 1805–1812, 2014.

[28] D. Hamester, P. Barros and S. Wermter, "Face Expression Recognition with a 2-Channel Convolutional Neural Network", In: *Proc. 2015 International Joint Conf. on Neural Networks,* Killarney, Ireland, pp. 1-8, 2015.

[29] T. Carvalhais and L. Magalhães, "Recognition and Use of Emotions in Games", In: *Proc. of International Conf. on Graphics and Interaction*, Lisbon, Portugal, pp. 1-8, 2018.

[30] J. Li and E. Y. Lam, "Facial Expression Recognition using Deep Neural Networks", In: *Proc. of IEEE International Conf. on Imaging Systems and Techniques,* Macau, China, pp. 1-6, 2015.

[31] M. Liu, S. Li, S. Shan, and X. Chen, "AU-Aware Deep Networks for Facial Expression Recognition", In: *Proc. of 10th IEEE International Conf. and Workshops on Automatic Face and Gesture Recognition,* Shanghai, China, pp. 1-6, 2013.

[32] M. Liu, S. Li and S. Shan, and X. Chen, "Facial Expression Recognition, AU-Inspired Deep Networks (AUDN), Micro-Action-Pattern, Receptive field, Group-wise Sub-Network Learning", *Neuro computing,* Vol. 159, No. 1, pp. 126-136, 2015.

[33] P. Khorrami, T. L. Paine, and T. S. Huang, "Do Deep Neural Networks Learn Facial Action Units When Doing Expression Recognition?", In*: Proc. of IEEE International Conf. on Computer Vision Workshop,* Santiago, Chile, pp. 19-27, 2015.

[34] H. Ding, S. K. Zhou, and R. Chellappa, "FaceNet2ExpNet: Regularizing a Deep Face Recognition Net for Expression Recognition", In*: Proc. of 12th IEEE International Conf. on Automatic Face and Gesture Recognition,* Washington, USA, pp. 118-126, 2017.

[35] N. Zeng, H. Zhang, B. Song, W. Liu, Y. Li, and A.M. Dobaie, "Facial Expression Recognition via Learning Deep Sparse Autoencoders", *Neuro Computing*, Vol. 273, No. 1, pp. 643–649, 2018.

[36] J. Cai, Z. Meng, A. S. Khan, Z. Li, J. O'Reilly, and Y. Tong, "Island Loss for Learning Discriminative Features in Facial Expression Recognition", In*: Proc. of 13th IEEE International Conf. on Automatic Face and Gesture Recognition,* Xi'an, China, pp. 302-309, 2018.

[37] Z. Meng, P. Liu, J. Cai, S. Han, and Y. Tong, "Identity-Aware Convolutional Neural Network for Facial Expression Recognition", In: *Proc. of*

12th IEEE International Conf. on Automatic Face and Gesture Recognition, Washington, USA, pp. 558-565, 2017.

[38] X. Liu, B. V. K. V. Kumar, J. You, and P. Jia, "Adaptive Deep Metric Learning for Identity-Aware Facial Expression Recognition", *In: Proc. of IEEE Conf. on Computer Vision and Pattern Recognition Workshops,* Honolulu, Hawaii, pp. 522-531, 2017.

[39] H. Yang, U. Ciftci, and L. Yin, "Facial Expression Recognition by De-Expression Residue Learning", In: *Proc. of IEEE/CVF Conf. on Computer Vision and Pattern Recognition,* Salt Lake City, USA, pp. 2168-2177, 2018.

[40] Z. Zhang, P. Luo, C. L. Chen, and X. Tang, "From Facial Expression Recognition to Interpersonal Relation Prediction", *International Journal of Computer Vision,* Vol.126, No. 5, pp. 1–20, 2018