

# A HYBRID PREDICTIVE ARCHITECTURE FORMULATION USING DEEP LEARNING AND HISTOGRAM OF GRADIENTS FOR COMPOUND EMOTION RECOGNITION

Anjana Guru PRASAD<sup>1</sup> , Shruti KULKARNI<sup>1</sup> , Vaishnavi Suresh BHANGENAVAR<sup>1</sup> , Vineet BELAGOD<sup>1</sup> , Vijayalakshmi Gopasandra Venkateshappa MAHESH<sup>1</sup> , Alex Noel JOSEPH RAJ<sup>2</sup> 

<sup>1</sup>Department of Electronics and Communication, BMS Institute of Technology and Management, Bangalore – 560064, India

<sup>2</sup>Department of Electronic Engineering, College of Engineering, Shantou University, Shantou – 515063, China

anjana.lg01@gmail.com shrutikulkarni2701@gmail.com vaishnavib1601@gmail.com  
vineet.belagod4111@gmail.com vijayalakshmi@bmsit.in jalexnoel@stu.edu.cn

DOI: 10.15598/aece.v22i1.5467

Article history: Received Sep 29, 2023; Revised Jan 18, 2024; Accepted Mar 19, 2024; Published Mar 31, 2024.  
This is an open access article under the BY-CC license.

**Abstract.** Facial emotion recognition has gained attention of researchers all over the world in the past few decades. Initially, emotions were classified in the seven basic categories which included happy, sad, angry, etc. However, human emotions are rarely this simple. They are usually combinations of dominant and complimentary emotions and are known as Compound Emotions. Two different ways have been commonly adapted for the recognition of these emotions from facial images: firstly, by using handcrafted features, or by using deep learning networks. This research analyzes the performance of a much simpler designed deep learning model named as Sequential-Convolution Neural Network (S-CNN) and four predefined deep learning networks for the recognition of compound emotions from facial images. The objective of this paper is to replace sophisticated state-of-the-art prediction models with a straightforward but effective approach. Therefore, this research suggests a hybrid network that maintains the S-CNN model's design simplicity while boosting performance. The features extracted by the S-CNN model and the handcrafted features are combined in the hybrid S-CNN model. This process keeps the hybrid model's architecture simple while improving its metrics values and increasing its accuracy to 99.62% when compared to other state-of-the-art models.

The source code for this research can be found in our GitHub repository: SCNN\_Hybrid\_model

## Keywords

**Deep learning, Convolutional Neural Networks, Histogram of Oriented Gradients, Feature extraction, Compound emotion recognition, Hybrid model.**

## 1. Introduction

Emotions tell us about an individual's mental state. They are directly related to the experiences faced by people. A person's body language, facial expression, body movement, skin resistance, breathing level, tone of speech, etc. come into consideration for recognition of human emotions. Several techniques have been proposed for the detection of human emotions based on the above-mentioned factors. One of the most common and effective techniques is Facial Emotion Recognition (FER). FER has intrigued researchers over the world in the past decades. The initial FER systems were made to classify the seven basic emotions namely, surprise, happiness, sadness, anger, disgust, fear, and contempt. However, the emotions expressed by humans are rarely this simple. Human beings tend to display Compound Emotions, which are a combination of dominant and complimentary emotions, for example, happily surprised, sadly disgusted, fearfully happy, etc. In recent years, classifier models based on machine learn-

ing/deep learning to recognize such compound emotions in addition to basic emotions are gaining more attention across the globe.

According to a study performed by Byoung Chul Ko [1], automatic FER has been performed in two different ways, first, by using handcrafted features (referred to as conventional FER in this paper) and second, by using features generated by Deep learning networks. The flow in conventional FER is as depicted in Figure 2, the face is first detected from the image, features are extracted from the facial images and recognition results are produced using pre-trained models like artificial neural network, Support Vector Machine, random forests, k nearest neighbour, linear discriminant analysis, etc. The extraction of handcrafted features is no more deemed necessary since Deep learning emerged in contrast to conventional methods. Among the deep learning models available, Convolutional Neural Networks (CNN) is the most popular for object recognition through images. CNN's have multiple convolutional layers and filters as shown in Figure 1 which enables them to be a very apt and fit network for image processing and classifications.

Although, deep learning methods are seen to have much better performance results than the conventional FER technique, in most mobile applications, conventional FER is used due to its reduced complexity of implementation. The complexity of deep learning algorithms is increased owing to the many layers present in their structure. If the structure was a little simpler, that is if the number of layers in the architecture was to be reduced, it would be easier to understand and implement these networks better.

This proposed work focuses on presenting a simple Deep learning architecture as compared to existing models while maintaining the reliable performance and integrate it with hand crafted features to have effective prediction performance. The proposed work is approached in 2 stages:

- Transfer learning techniques are used to measure the performance of four pre-defined architectures of Resnet, Visual Geometry Group (VGG), MobileNet, Inception, and a fifth much simpler neural network on the RAF Database to recognize compound emotions. Comparative performance is carried to identify the architecture with better prediction,
- To increase the performance of emotion recognition, a hybrid model is developed using the result of (i) and hand-crafted features which is a combination of conventional FER and Deep Learning method.

This paper is organized into eight sections. In section two, the related work on the concept of deep learning

for recognition of compound emotions has been discussed. In section three, there are two subsections of which, the first subsection describes four pre-trained CNN models in addition to a deep learning network architecture experimented by the authors. The next subsection describes a methodology for compound emotion recognition which integrates the handcrafted features with deep learning algorithm features to achieve better outcomes. Section six includes a discussion on the results obtained. This paper is finally concluded in the last section.

## 2. Related Work

Compound emotion has been described by Du et al. [2] as a combination of two basic emotions. Their results indicate that the emotions usually shown by humans are better classified with the wider set of basic and compound emotions than by the smaller category of seven basic emotions. They have provided an in-depth analysis of the production of compound facial expressions. Appearance based methods such as Gabor Wavelets [3], Histogram of Oriented Gradients (HOG) [4], and Local Binary Patterns [5] are used with machine learning algorithms compound emotion recognition.

In recent times deep learning models have performed better than highly developed machine learning algorithms. Liu et al. [6] analyzed the facial expressions by applying a geometric model on facial regions using the concept of deep learning and Lu et al. [7] made use of CNN on facial appearance. The latest models designed focused on improving their performance by using different CNN classifiers [8, 9]. Although the previous models focused on static images, temporal information can be used for the analysis of facial expressions. Cohen et al. [10] used Hidden Markov models on video sequences.

The approach proposed by Pons et al. [11] is a novel multi-label loss function. This loss function can be incorporated into the CNN to enhance the training of the emotion recognition task by integrating complementary tasks and data from different sources. This method was developed to overcome the challenge of the lack of large publicly labeled databases. However, the accuracies obtained are considerably low.

Pendhari et al. [12] worked on InceptionResNet-v2 architecture for Image classification and feature extraction uses pre-trained CNNs. The purpose of this research is to detect 7 basic and 15 compound facial emotions using the Compound Facial Expression Emotions (CFEE) database. The accuracy of compound facial emotions detected from the proposed InceptionResNet-v2 architecture was 57.70%. The architecture gave an

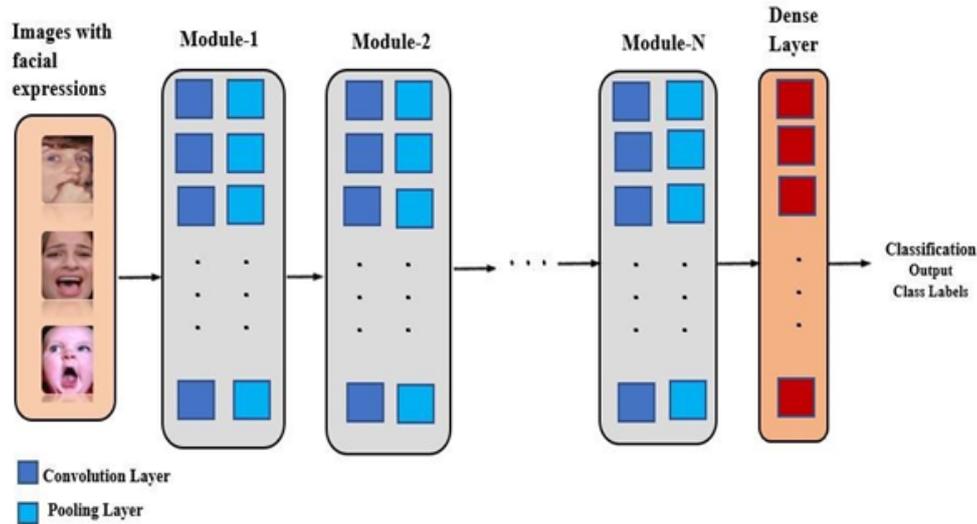


Fig. 1: Basic structure of CNN for image classification.

appalling performance for other emotions because the CFEE database consisted of fewer images which led to the prevention of the network to learn discriminant features to classify compound facial emotions accurately. This caused overfitting because of the small database. S. K. Jarraya et al. [13] performed a study on Compound Emotion Recognition (CER) on a group of autistic children during a meltdown crisis using deep spatio-temporal geometric features of micro-expressions. A comparison was made between the CER performance and diverse collections of micro-expressions features that selects the best features that differentiate autistic children CE in meltdown crisis from the normal state, and the best classifier performance. The performance obtained was 85.8%. But, this method fails to attain a higher accuracy.

Byoung Chul Ko [1], conducted a performance evaluation of the FER approaches presented by different researchers. The obtained results depict a direct comparison between the conventional (handcrafted-feature) based approach and the deep learning-based approach. It is shown that the average accuracy obtained by conventional techniques is 63.20% whereas the average accuracy by deep learning method is 72.65%. One of the issues mentioned in this article is the increasingly deep structure of the deep learning models which makes them unsuitable for implementation. A very simple architecture is used to resolve this issue for the proposed model at the same time achieving almost the same accuracy.

From the Literature survey, it is found that compound emotion is recognized or detected from the facial expression images using machine learning algorithms with handcrafted features and deep learning architectures. Each method has its strengths in recognizing emotions. Thus, in this work, the key strengths of

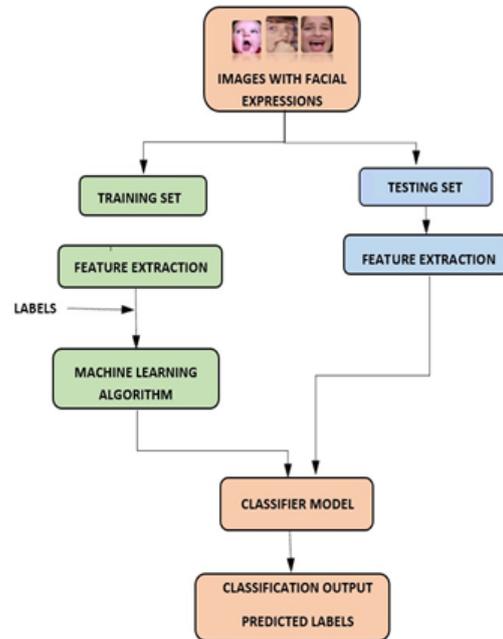


Fig. 2: Basic structure of CNN for image classification.

both methods are utilized to come up with a hybrid method that integrates the handcrafted features with deep learning features for better facial expression representation. Later the integrated features are deployed to train the classifier model for compound emotion recognition. Further, the performances of the existing deep learning models are compared with that of the Hybrid method in terms of the metrics: Accuracy, Precision, Recall, F1-Score, Area under ROC(Receiver operating characteristic)-curve(AUC) and loss function.

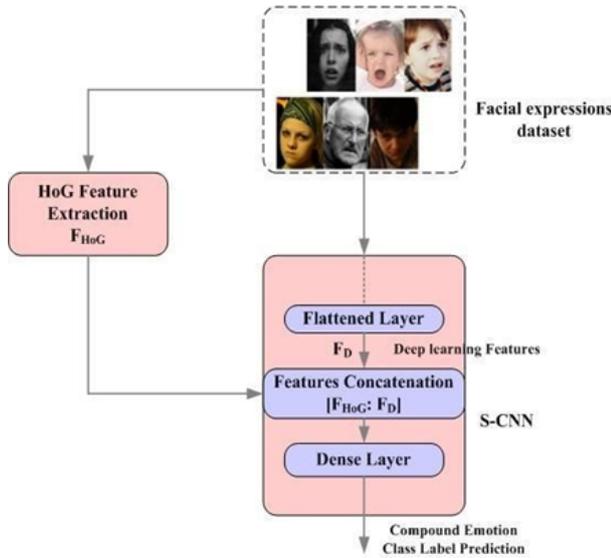


Fig. 3: Proposed Hybrid model.

### 3. Methodology

The proposed work was developed in two stages. (i) In the first stage, existing Deep learning models: ResNet50, MobileNet, VGG16, Inception-V3 were used based on transfer learning along with a CNN model (proposed by authors S-CNN) with three convolutional layers, three maxpool layers, one flatten-layer whose output is given to the dense layers for facial compound emotion recognition (FCER) as illustrated in Figure 2.

Further a comparison was done on the performance of different deep learning architectures in classifying eleven compound emotions as shown in the Table 1. The values of accuracy, precision, loss, f1\_score was compared. (ii) In the second stage, a hybrid method was developed with deep learning architecture (DLA), where the handcrafted Histogram of oriented gradients (HoG) features abstracted from the facial expression images were integrated/concatenated with the deep learning features obtained at the output of the flatten layer of DLA and was given to the fully connected layer for classification. The classification layer gives the class label of the image indicating the associated emotion. The process is depicted in Figure 3. The concatenation results in better performance and improves the class discrimination ability of the classifier.

### 4. Feature Extraction

Feature extraction involves describing the data for better representation. The proposed model utilizes the Histogram of Oriented Gradients (HoG) [14] feature descriptor to represent the facial expression images. The process of feature extraction is described in Figure

4 for this purpose. The HoG descriptor concentrates on the shape or structure of an object and since it is a local descriptor applied on the local regions of the image, it provides detailed representation required for better differentiation of the facial images in identifying the underlying emotion. This extraction technique uses magnitude as well as the angle of the gradient to compute the features which make it better than any edge descriptor.

#### 4.1. HoG Feature extraction steps

Step 1: The entire image is divided into cells of size  $M \times N$ . Consider an image matrix of size  $[4 \times 4]$  given by:

$$I = \begin{bmatrix} 12 & 15 & 16 & 18 \\ 20 & 30 & 50 & 40 \\ 100 & 150 & 10 & 13 \\ 23 & 35 & 45 & 22 \end{bmatrix}$$

Step 2: Using the following expressions, the values of vertical and horizontal gradients are calculated

$$\begin{aligned} G_x &= I(x+1, y) - I(x-1, y) \\ G_y &= I(x, y+1) - I(x, y-1) \end{aligned} \quad (1)$$

where

$$\begin{aligned} G_x &= \begin{bmatrix} 3 & 4 & 3 & 2 \\ 10 & 30 & 10 & -10 \\ 50 & -90 & -137 & 3 \\ 12 & 22 & -13 & -23 \end{bmatrix}, \\ G_y &= \begin{bmatrix} 8 & 15 & 34 & 22 \\ 88 & 138 & 11 & 11 \\ 77 & 117 & 46 & 24 \\ 12 & 22 & -13 & -23 \end{bmatrix}. \end{aligned}$$

Step 3: The magnitude and gradient are computed using the expressions indicated,

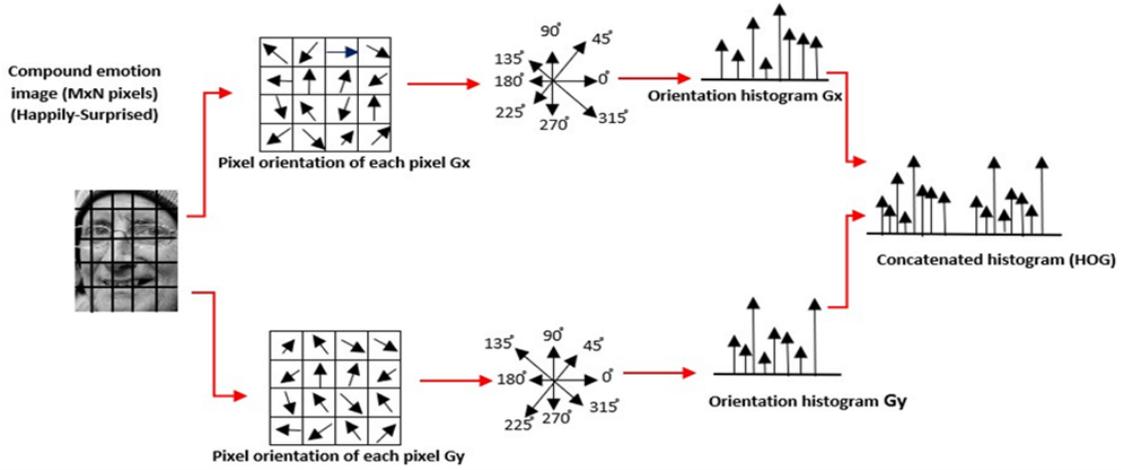
$$\text{Magnitude: } G = \sqrt{(G_x + G_y)}, \quad (2)$$

$$G = \begin{bmatrix} 8 & 15 & 34 & 22 \\ 88 & 135 & -6 & -5 \\ 3 & 5 & -5 & -18 \\ -77 & -115 & 35 & 9 \end{bmatrix},$$

$$\text{Angle: } \alpha = \tan^{-1} \frac{G_x}{G_y}, \quad (3)$$

$$\alpha = \begin{bmatrix} 69 & 75 & 84 & 84 \\ 83 & 77 & 149 & 26 \\ 3 & 177 & 2 & 100 \\ 99 & 101 & 111 & 159 \end{bmatrix}.$$

Step 4: Now, the pixels are classified into 8 evenly spaced bins using unsigned orientation from 0-360 degree. The value of the bin is given by the sum of the



**Fig. 4:** The figure depicts the method employed for the HOG feature extraction technique. The orientation of all  $M \times N$  pixels of the image cells is calculated and stored in an  $M$ -bins histogram of orientations. The final features vector is constructed by concatenating the cell histograms. The above figure depicts a cell size of 4 pixels and 8 orientation bins for the cell histograms Courtesy [15].

magnitudes of all the corresponding pixels having the angles lying in the same bin.

Step 5: In order to remove unwanted zeros present in the sample image, Normalization is performed and is done using equation (4),

$$L2 \text{ Norm: } |x|_2 = \sqrt{\sum_{i=1}^N |x_i|^2}. \tag{4}$$

In the end, all cell histograms are concatenated to construct the final feature vector HoG.

### 5. Performance Matrix

Performance metrics are important to evaluate a classifier model quantitatively. In this section, quantitative measurements of the model are taken to track its performance and quantify its quality of predictions. To do so six metrics functions have been made use of namely accuracy, precision, loss, F1 score, recall, and AUC.

**ACCURACY:** Accuracy is a metric that is used to measure the model performance over all classes. It is generally beneficial when all classes are of equally important. It is computed as the ratio between the number of correct predictions to the total number of predictions [16].

$$Accuracy = \frac{T_p + T_n}{T_p + T_n + F_p + F_n}. \tag{5}$$

**RECALL (R):** The recall is computed as the ratio between the number of Positive samples classified correctly as Positive to the total number of Positive samples. The recall measures the ability of the model to detect Positive samples. Higher recall denotes that more



**Fig. 5:** Images representing Compound emotions in the RAF-Database.

positive samples are detected [16].

$$Recall = \frac{T_p}{T_p + F_n}. \quad (6)$$

PRECISION (P): The precision is computed as the ratio between the number of positive samples classified correctly to the total number of samples classified as Positive (either correctly or incorrectly). The model's accuracy in classifying the sample as positive is given by precision metrics [16]

$$Precision = \frac{T_p}{T_p + F_p}. \quad (7)$$

F1-score: Precision and accuracy are the building blocks of this metrics. The two metrics are combined into a single metrics, it is the harmonic mean of precision and accuracy. F1 score is high if both precision and accuracy are high and low if both are low [16].

$$F1 - score = 2 * \frac{P * R}{P + R}. \quad (8)$$

AUC: "Area under curve" measures the entire area underneath a ROC which is graph where true positive is plotted against false positive. AUC provides an aggregate measure of performance across all possible classification thresholds where performance across all possible classification is the ROC [17].

LOSS: A loss function computes the distance between the current output from that of the expected output. This distance being less represents that the model algorithm is learning well. This function is used as feedback to the model algorithm to vary the weights in turn improve the learning of the model [18].

Where  $T_p$  is True positive,  $T_n$  is True negative,  $F_p$  is False positive, and  $F_n$  is False negative.

## 6. Results and Discussion

The proposed work on compound emotion recognition is carried out using facial expression images. Images were considered as compared to other modalities as the variations in the face during emotion are clearly visible such as raised eyebrow, opened mouth, broadened cheeks, wrinkled nose and formation of fine lines on forehead and chin etc. These variations in the face are captured using images and can be described efficiently using shape descriptors. The experiment on CER was conducted in two stages. (i) In first stage, available deep learning models: Resnet-50, InceptionV3, VGG-16, mobile-net and a CNN model (S-CNN proposed by authors) were used that combines both feature extraction and classification. The facial expression images with class labels were applied to deep learning models for training and testing. Later the performance of each

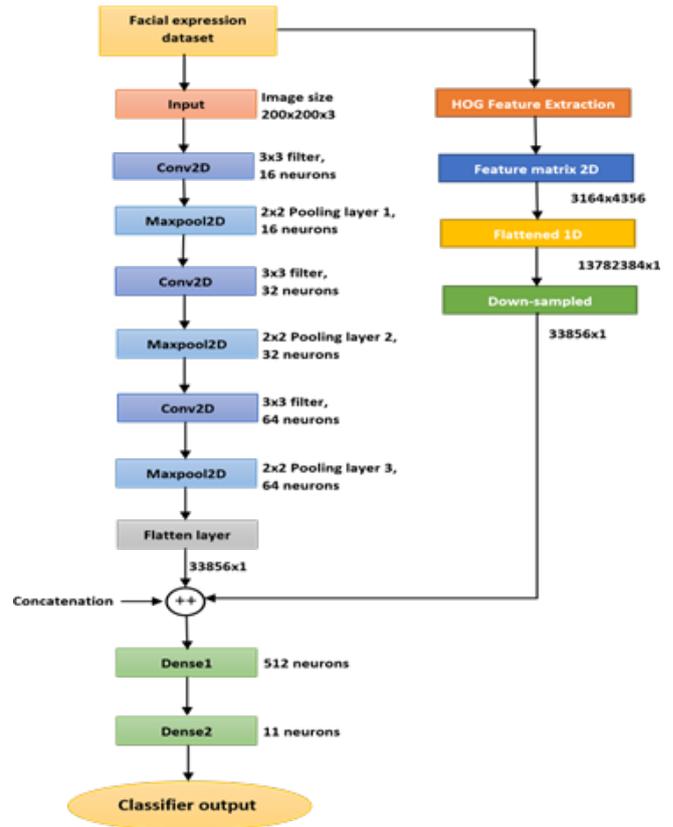
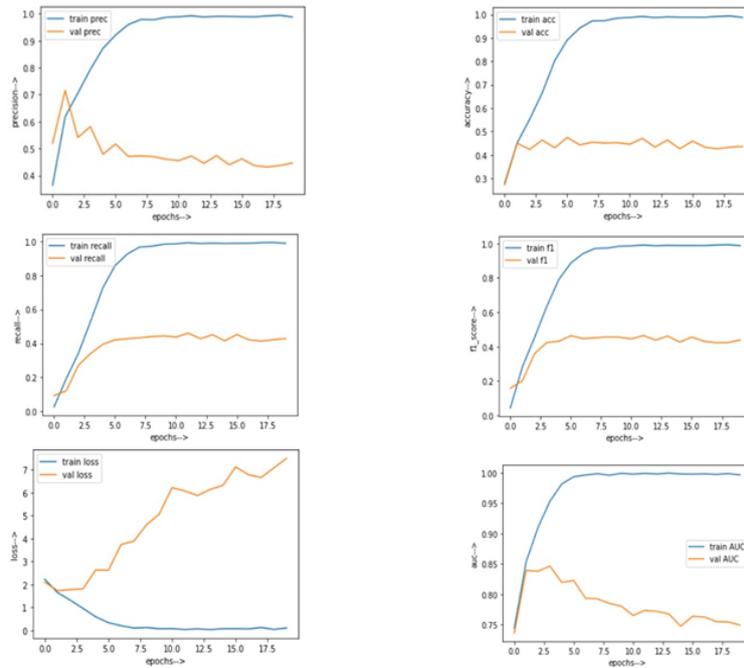


Fig. 6: Flowchart of architecture of the proposed hybrid model.

model is assessed and compared. (ii) In the second stage, to improve the performance of CER, a hybrid methodology was framed. In this method, the deep learning features were concatenated with the hand-crafted HoG features extracted from the facial expression images. The integration was done after flattened layer. The integrated features were provided to the last layer for classification to identify the associated emotion and the performance of the hybrid model is evaluated.

### 6.1. Datasets

The primary step in the experimentation is the assemblage of good datasets. For training and testing various models considered in the proposed work, Real-world Affective Faces Database (RAF-DB) [19] is used. Authors collected Flickr image URLs and downloaded them in bulk using an open-source downloader. They used the well-structured Extensible markup language output from Flickr's Application program Interface to read and filter images with neutral expressions and basic emotions. It is a large-scale database with around 30K heterogeneous collection of facial expression images. The database has images of individuals of different age, gender, captured under different lightning conditions, various filters and effects. The ages of in-



**Fig. 7:** Plots of hybrid model performance metrics- precision, accuracy, recall, f1-score, loss, auc.

dividuals included in this database range from 0 to 70. Of these, 43% are men, 52% are women, and 5% are not sure. There are 77% Caucasian, 8% African-American, and 15% Asian in terms of racial distribution. The images of the dataset are annotated with eleven compound emotion class labels. The samples of the dataset are provided in the Figure 5. In the images shown, the variations in the face during an emotion can be observed. These variations form key patterns to identify the emotion.

To measure the performance of the models, the database is divided into training and test datasets where the size of training set is five times more than test set with uniform distribution in both sets. From these 30K images of size 100X100 each, 3889 images were taken and divided into 19% (725) as test dataset and remaining 81% (3164) as training dataset. The images of the dataset were reorganized based on 11 distinct compound emotions for both training as well as for testing. Before training the models, the aligned images were resized and then given to the input layer. As mentioned, the experimentation is done in two stages, the results of both the stages are presented and discussed here in the same order.

- Inception V3

InceptionV3 is a superior version of the basic model inceptionV1 that employs the concept of factorizing convolutions, efficient grid size reduction, and utility of auxiliary classifiers to burn down the number of processing parameters and hence resulting in a computationally inexpensive model and faster training [21]. As

seen from Figure 6(b) there are three convolution layers (3x3,5x5 and 1x1) and one pooling layer are all placed on the same level hence framing a wider model than a deeper one. The model consists of 48 such levels using factorized filters (nxn is nx1 and nx1) and the ReLU activation function.

## 6.2. Stage 1

In this stage, a comparative study between five CNN models namely Resnet-50, InceptionV3, VGG-16, mobile-net, and the proposed models(S-CNN) has been done to assess the performance of each architecture/model.

- Residual Network-50

This model was immensely successful and developed by a Microsoft research Asia team in 2015. A residual learning framework is applied in its architecture to ease the training of the network [20] to resolve the gradient descent issue. The concept of “skip connections” lies at the core of the residual blocks which strengthens the neural network as viewed in Figure 6(a). The model is 50 layers deep comprising 48 convolution layers, one max pool layer, and one average pool layer. The convolution layers mostly have 3x3 filters whilst the skip connections use a 1x1 filter with the Rectified Linear unit (ReLU) as its activation function.

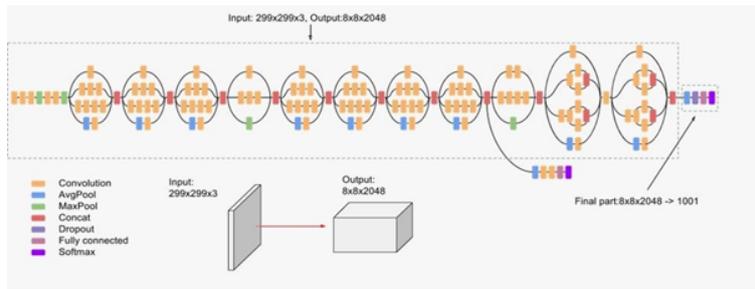


Fig. 8: (a) Inception V3 model with reduced number of parameters for faster computation, courtesy [24]

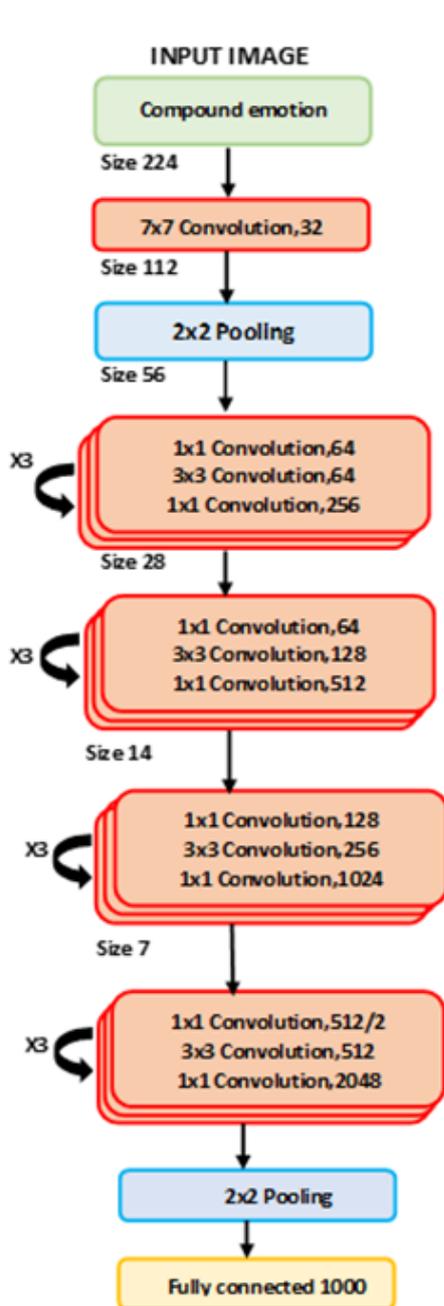


Fig. 8: (b) ResNet architecture with 34 layers

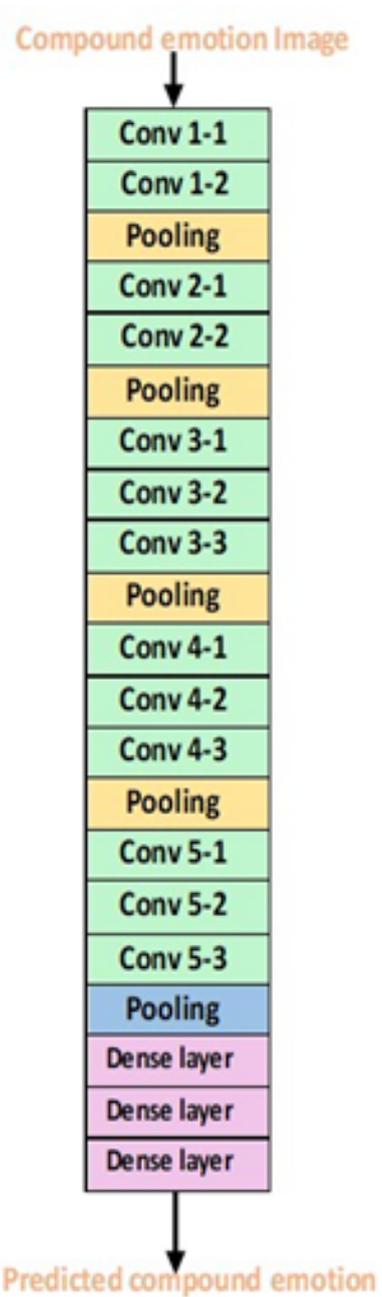


Fig. 8: (c) VGG-16 architecture

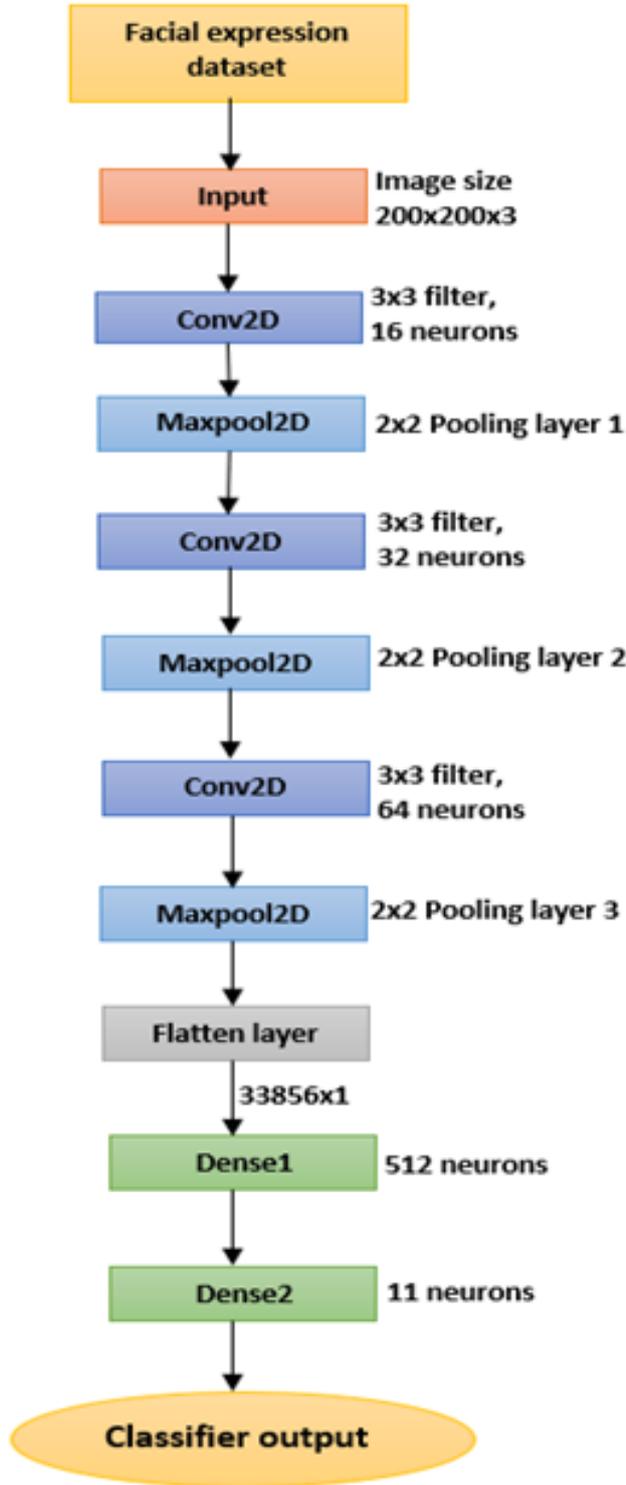


Fig. 8: (d) Simplified feed forward convolution neural network with 11.7 million processing parameters.

- Visual Geometrics Group

VGG, short for Visual Geometry Group supports 16 convolution layers. Mainly achieved high accuracy with small 3x3 filters. It is comprised of two sets of 2 convolution layers and three sets of 3 convolution layers. Each of these convolution layers is followed by a

Tab. 1: Labels indicating each class during training the models.

| Description | Label | Emotion               |
|-------------|-------|-----------------------|
| c1          | 0     | Happily Surprised     |
| c2          | 1     | Happily Disgusted     |
| c3          | 2     | Sadly Fearful         |
| c4          | 3     | Sadly angry           |
| c5          | 4     | Sadly Surprised       |
| c6          | 5     | Sadly Disgusted       |
| c7          | 6     | Fearfully Angry       |
| c8          | 7     | Fearfully Surprised   |
| c9          | 8     | Angrily Surprised     |
| c10         | 9     | Angrily Disgusted     |
| c11         | 10    | Disgustedly Surprised |

pooling layer lastly the flattened output vector is fed to a fully connected layer shown in Figure 6(c). The hidden layers used the ReLu activation function whereas the last dense layer used softmax [22].

- Mobilenet

MobileNet was developed to effectively execute mobile and embedded applications. It has a linear architecture that is 28 layers deep. These layers make use of depth-wise separable convolutions to build a lightweight deep convolutional neural [23]. Figure 6(d) is the outline of its architecture. The number of parameters processed is largely reduced when compared to the network of the same depth in the nets and the same convolutions. A depth-wise separable convolution is made of two operations, that is, depth-wise convolution(3x3) and pointwise convolution(1x1) essentially using the ReLu6 activation function.

- S-CNN

This is a feed-forward convolution network with 10 layers deep. The architecture of the model Figure 6(e) comprises an input layer followed by three sets of 2D convolution and 2D max-pool layers. Each 2D convolution layer having 16,32, and 64 neurons respectively. Finally, the output from the last 2D max-pool layer is flattened resulting in a [33856x1] vector. This 1D Keras tensor type vector is fed to the fully connected dense layers. The output layer having 11 neurons is a [11x1] vector of probabilities p1, p2, p3, ... p11. Each probability represents how likely the input image belongs to a class label c1, c2, c3, c4, ... c11. The predicted class for the given input image is the one which has highest probability in the class label vector. The hidden layers make use of the ReLu activation function with a filter size of 3x3. The final dense layer uses a softmax function. Each max-pool layer makes use of a 2x2 window with a stride set to one. The metrics values obtained are consolidated as shown in Table 2.

**Tab. 2:** illustrates the performance/efficiency of the respective models based on the metrics equations 1,2,3 and 4. Each of these models has been trained for 20 utilizing Adam (Mobile net), RMSprop (Resnet-50 and inceptionv3), or SDG(VGG-16) optimizers.

| Metrics   | Resnet-50 | Inceptionv3 | VGG-16  | Mobilenet | S-CNN  |
|-----------|-----------|-------------|---------|-----------|--------|
| Accuracy  | 0.9829    | 0.8521      | 0.4172  | 0.8489    | 0.9867 |
| Precision | 0.9857    | 0.9048      | 0.4173  | 0.8500    | 0.9867 |
| Loss      | 0.0603    | 0.4640      | 48.2884 | 1.9430    | 0.1507 |
| F1-Score  | 0.9831    | 0.8407      | 0.4171  | 0.8500    | 0.9867 |
| Recall    | 0.9801    | 0.7870      | 0.4172  | 0.8486    | 0.9867 |
| Auc       | 0.9999    | 0.9903      | 0.6883  | 0.9411    | 0.9955 |

**Tab. 3:** Confusion matrix of proposed model. {co1, co2, ... co11} represent the obtained classifier outputs. {c1, c2, ... c11} represent the true class labels.

|     | co1 | co2 | co3 | co4 | co5 | co6 | co7 | co8 | co9 | co10 | co11 |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|------|------|
| c1  | 76  | 113 | 18  | 37  | 23  | 22  | 14  | 120 | 24  | 83   | 22   |
| c2  | 125 | 140 | 20  | 37  | 21  | 31  | 15  | 136 | 22  | 87   | 34   |
| c3  | 21  | 21  | 4   | 8   | 2   | 7   | 4   | 24  | 5   | 13   | 5    |
| c4  | 42  | 52  | 10  | 17  | 7   | 13  | 2   | 38  | 8   | 22   | 6    |
| c5  | 22  | 27  | 2   | 8   | 1   | 2   | 3   | 17  | 3   | 14   | 7    |
| c6  | 21  | 26  | 3   | 13  | 4   | 7   | 2   | 30  | 5   | 21   | 5    |
| c7  | 14  | 17  | 8   | 3   | 2   | 2   | 2   | 9   | 1   | 11   | 0    |
| c8  | 108 | 135 | 23  | 44  | 15  | 18  | 9   | 121 | 23  | 81   | 29   |
| c9  | 23  | 22  | 6   | 9   | 7   | 10  | 2   | 13  | 8   | 10   | 7    |
| c10 | 85  | 83  | 18  | 30  | 17  | 16  | 13  | 73  | 12  | 76   | 17   |
| c11 | 18  | 32  | 2   | 12  | 7   | 9   | 3   | 24  | 6   | 19   | 6    |

### 6.3. Stage 2

Analyzing the results obtained from stage 1, it is observed that the performance of Resnet-50, Inception, and Mobilenet is better compared to VGG-16 and S-CNN. At the same time, these transfer learning models have skip-connections, several layers and sub-layers, etc. which bring in non-linearity in their architectures making the computations complex. On the other hand, S-CNN has a simpler architecture that is straightforward, easier to comprehend, analyze, and implement for various applications. Hence, in this sub-section, S-CNN has been modified to obtain similar performance as the transfer learning models while retaining its simple architecture. This has been achieved by concatenating features obtained from S-CNN with handcrafted HoG features resulting in a proposed hybrid model.

In this, the input layer accepts images of size 200 x 200 and has a depth of 3 (RGB). The first, second and third convolution layers and maxpool layers has 16, 32 and 64 neurons respectively and uses ReLu activation function. The flatten layer which gives output as 1-Dimensional array has 33856 neurons. The first dense layer, which has been given the concatenated output of flatten layer and flattened handcrafted features has 512 neurons with activation function ReLu. The output layer, also a dense layer, has 11 neurons which is equal to the number of outputs, uses SoftMax activation function. A characteristic of ReLu activation function is to not activate all the neurons of that layer at once. In addition to it, the good computational efficiency obtained makes it a good choice. Also, SoftMax is used in the output layer to classify images into var-

ious classes. The architecture of the proposed hybrid model is depicted in Figure 6.

The HOG feature matrix of [3164x4356] extracted from the compound emotion dataset is flattened[13782384x1]. To concatenate this matrix with the flattened output of the neural network two conditions must be satisfied. Firstly, the order of the two matrices must be equal. Secondly, the datatype of the two matrices must match. Hence the one-dimension HOG feature matrix is down-sampled to the order[33856x1] and explicitly type converted from CSV to Keras tensor type. This resulting matrix is concatenated with the flattened output ensuring a prominent increase in the prediction efficiencies.

The flattened features extracted by the convolution and maxpooling layers in CNN model were then concatenated with hand crafted features obtained using HOG descriptor before giving it to the dense layers of the network. The extraction of handcrafted features was done with the help of the online tool, MATLAB. The extracted features were found to be in the format of a 2-dimentional matrix of size 3164 x 4356. Before concatenation, the obtained 2D matrix was flattened to 13782384 x 1 and then down-sampled to 33856 x 1. Down-sampling is done to match the size of the flatten layer of the network. Concatenation was done using the predefined function concatenate(). The concatenated output of the flatten layer and flattened feature matrix is then given to dense layers. Concatenation of features in CNN model has made the model more efficient and gave good performance compared to regular

**Tab. 4:** Comparison of different CER models.

| Author                   | Dataset Used                          | Method  | Accuracy        |
|--------------------------|---------------------------------------|---|-----------------|
| Pons Et al. [11]         | Extended Cohn-Kanade, SFEW2.0database | Multilabel loss function (Deep learning)          | 0.5940          |
| Pendhari Etal. [12]      | CFEE                                  | Deep learning                                     | 0.5770          |
| S. K. JarrayaEt al. [13] | Captured using Kinect camera          | Deep spatio-temporal geometric features           | 0.8550          |
| ByoungChul Ko [1]        | -                                     | Conventional (handcrafted-feature) FER approaches | 0.6320(average) |
| ByoungChul Ko [1]        | -                                     | Deep-learning-based FER approaches                | 0.7265(average) |
| Proposed method          | RAF-DB                                | Combination of CNN and handcrafted HoG            | 0.9962          |

**Tab. 5:** Performance measure of proposed hybrid S-CNN model.

| Metrics   | Values |
|-----------|--------|
| Accuracy  | 0.9962 |
| Precision | 0.9964 |
| Loss      | 0.0204 |
| F1-score  | 0.9964 |
| Recall    | 0.9961 |
| AUC       | 0.9962 |

CNN model. The accuracy of hybrid model is increased by an average of 22.17%.

For examining the classifier's performance, different metrics have been obtained. The confusion matrix for the hybrid model is given in Table 3. Other metrics values obtained are shown in Table 5. Graphical Representation of each metric of hybrid model is as shown in Figure 7.

It is found that although while S-CNN processes less training parameters than transfer learning models, the resulting performance is on par with sophisticated architectures, indicating that a basic design might be adequate for CER applications.

#### 6.4. Limitations and Future Scope

It should also be noted that the proposed approach also has certain areas of improvements. In today's world, wearing face-masks has become a relatively common thing to do, but this results in covering half of the face features. Emotion recognition is still possible by observing the human eye, however, the proposed model has been trained to detect emotions based on the entire face, due to which, the results in such cases might not be up to the mark.

The model was trained only on a very small number of images (3164). Due to this reason, the model might not perform optimally in special situations like when people are wearing face masks.

**Tab. 6:** Performance measure of proposed hybrid S-CNN model.

| Metrics            | Values        |            |
|--------------------|---------------|------------|
| Type/Stride        | Filter Shape  | Input Size |
| Convolution /S2    | 3x3x3x32      | 224x224x3  |
| Convolution DW/S1  | 3x3x32 DW     | 112x112x32 |
| Convolution /S1    | 1x1x32x64     | 112x112x32 |
| Convolution DW/S2  | 3x3x64 DW     | 112x112x64 |
| Convolution /S1    | 1x1x64x128    | 56x56x64   |
| Convolution DW/S1  | 3x3x128 DW    | 56x56x128  |
| Convolution /S1    | 1x1x128x128   | 56x56x128  |
| Convolution DW/S2  | 3x3x128 DW    | 56x56x128  |
| Convolution /S1    | 1x1x128x256   | 56x56x128  |
| Convolution DW/S1  | 3x3x256 DW    | 56x56x128  |
| Convolution /S1    | 1x1x256x256   | 56x56x128  |
| Convolution DW/S2  | 3x3x256 DW    | 56x56x128  |
| Convolution /S1    | 1x1x256x512   | 14x14x256  |
| Convolution DW/S1  | 3x3x512 DW    | 14x14x512  |
| 5x Convolution /S1 | 1x1x512x512   | 14x14x512  |
| Convolution DW/S2  | 3x3x512 DW    | 7x7x512    |
| Convolution /S1    | 1x1x512x1024  | 7x7x1024   |
| Convolution Dw/S2  | 3x3x1024 DW   | 7x7x1024   |
| Convolution /S1    | 1x1x1024x1024 | 7x7x1024   |
| Average Pool/S1    | Pool 7x7      | 7x7x1024   |
| FC/S1              | 1024x1000     | 1x1x1024   |
| Softmax/S1         | Classifier    | 1x1x1000   |

The next step would be to implement this model in real time. Currently, the code only performs on images, however, this model would be more useful if it could work on videos. Another area of improvement could be the detection of intensity of emotions, i.e., the model could be improvised so as to be able to detect the dominant of the compound emotion they are displaying.

## 7. Conclusion

In this work, 3889 images, collected from the RAF-DB, were classified into 11 compound emotions. All models were trained and a comparative study was performed. For 20 epochs, it can be inferred that the hybrid model that was proposed gave 99.62% accuracy, almost matching with 98.29% of ResNet50, even

though it has only 10 layers and a very simple architecture when compared to the complex architecture of Resnet model. This model can be used in various domains for various purposes. It can detect sudden emotional changes and can help to take preventive measures to deal with meltdown crisis. They can be used in measuring effectiveness of a lecturer for real time responses in online classes. They could also be used in the field of robotics, where interactions with robots is done through facial expressions.

## References

- [1] KO, B. A Brief Review of Facial Emotion Recognition Based on Visual Information. *Sensors* [online]. 2018, 18(2), 401 [viewed 27 September 2023]. ISSN 1424-8220. DOI: 10.3390/s18020401.
- [2] DU, S., Y. TAO, and A. M. MARTINEZ. Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences* [online]. 2014, 111(15), E1454–E1462 [viewed 27 September 2023]. ISSN 1091-6490. DOI: 10.1073/pnas.1322355111.
- [3] BARTLETT, M. S., G. LITTLEWORT, I. FASEL, and J. R. MOVELLAN. Real Time Face Detection and Facial Expression Recognition: Development and Applications to Human Computer Interaction. Online. In: 2003 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW). Madison, Wisconsin, USA, 2003-06-16 – 2003-06-22. IEEE, 2003. ISBN 0-7695-1900-8. DOI: 10.1109/cvprw.2003.10057 [viewed 2023-09-27].
- [4] LI, Z., J.-I. IMAI, and M. KANEKO. Facial-component-based bag of words and PHOG descriptor for facial expression recognition. In: 2009 IEEE International Conference on Systems, Man and Cybernetics - SMC [online]. IEEE, 2009 [viewed 27 September 2023]. ISBN 9781424427932. DOI: 10.1109/ic-smc.2009.5346254.
- [5] ZHAO, G., and M. PIETIKAINEN. Dynamic Texture Recognition Using Local Binary Patterns with an Application to Facial Expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence* [online]. 2007, 29(6), 915–928 [viewed 27 September 2023]. ISSN 2160-9292. DOI: 10.1109/tpami.2007.1110.
- [6] LIU, M., S. LI, S. SHAN, R. WANG, and X. CHEN. Deeply Learning Deformable Facial Action Parts Model for Dynamic Expression Analysis. Online. In: *Computer Vision – ACCV 2014*, pp. 143–157. Cham: Springer International Publishing, 2015. ISBN 9783319168166. DOI: 10.1007/978-3-319-16817-3\_10. [viewed 2023-09-27].
- [7] LU, G., J. HE, J. YAN, and H. LI. Convolutional neural network for facial expression recognition. Online. *Journal of Nanjing University of Posts and Telecommunications*, vol. 36 (2016), pp. 16–22. ISSN 1673-5439. DOI: 10.48550/arXiv.1704.06756. [viewed 2023-09-28].
- [8] KIM, B.-K., H. LEE, J. ROH, and S.-Y. LEE. Hierarchical Committee of Deep CNNs with Exponentially-Weighted Decision Fusion for Static Facial Expression Recognition. Online. In: *ICMI '15: INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION*. Seattle Washington USA. New York, NY, USA: ACM, 2015. ISBN 9781450339124. DOI: 10.1145/2818346.2830590. [viewed 2023-09-27].
- [9] PONS, G., and D. MASIP. Supervised Committee of Convolutional Neural Networks in Automated Facial Expression Analysis. *IEEE Transactions on Affective Computing* [online]. 2018, 9(3), 343–350 [viewed 27 September 2023]. ISSN 2371-9850. DOI: 10.1109/taffc.2017.2753235.
- [10] COHEN, I., N. SEBE, A. GARG, M. S. LEW, and T. S. HUANG. Facial expression recognition from video sequences. Online. In: *IEEE International Conference on Multimedia and Expo (ICME)*. Lausanne, Switzerland. IEEE, [n.d.]. ISBN 0780373049. DOI: 10.1109/icme.2002.1035527. [viewed 2023-09-27].
- [11] PONS, G., and D. MASIP. Multi-task, multi-label and multi-domain learning with residual convolutional networks for emotion recognition. Preprint; online. *IJCV*, 2018. DOI: 10.48550/arXiv.1802.06664. [viewed 2023-09-27].
- [12] PENDHARI, H., S. NAGDEOTI, S. RATHOD, L. KHAN, and S. VISHWAKARMA. Compound Emotions: A Mixed emotion detection. Online. *SSRN Electronic Journal*, 2022. ISSN 1556-5068. DOI: 10.2139/ssrn.4120265. [viewed 2023-09-27].
- [13] JARRAYA, S. K., M. MASMOUDI, and M. HAMMAMI. Compound Emotion Recognition of Autistic Children During Meltdown Crisis Based on Deep Spatio-Temporal Analysis of Facial Geometric Features. *IEEE Access* [online]. 2020, 8, 69311–69326 [viewed 27 September 2023]. ISSN 2169-3536. DOI: 10.1109/access.2020.2986654.

- [14] DALAL, N., and B. TRIGGS. Histograms of Oriented Gradients for Human Detection. Online. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). San Diego, CA, USA. IEEE, [n.d.]. ISBN 0769523722. DOI: 10.1109/cvpr.2005.177. [viewed 2023-09-27].
- [15] CARCAGNÌ, Pierluigi, et al. Facial expression recognition and histograms of oriented gradients: a comprehensive study. SpringerPlus [online]. 2015, 4(1) [viewed 15 March 2024]. ISSN 2193-1801. DOI: 10.1186/s40064-015-1427-3.
- [16] POWERS D.M.W. Evaluation: from Precision, Recall and F-measure to Roc, Informedness, Markedness & Correlation. Journal of Machine Learning Technologies [online]. 2011, 2(1), 37–63. ISSN 2229-3981. DOI: 10.48550/arXiv.2010.16061.
- [17] FAWCETT, T. An introduction to ROC analysis. Pattern Recognition Letters [online]. 2006, 27(8), 861–874 [viewed 27 September 2023]. ISSN 0167-8655. DOI: 10.1016/j.patrec.2005.10.010
- [18] KUMAR, S. *Neural Netwrok*. 2nd ed. McGraw Hill Education, 2017. ISBN 1259006166.
- [19] LI, S., and W. DENG. Reliable Crowdsourcing and Deep Locality-Preserving Learning for Unconstrained Facial Expression Recognition. IEEE Transactions on Image Processing [online]. 2019, 28(1), 356–370 [viewed 27 September 2023]. ISSN 1941-0042. DOI: 10.1109/tip.2018.2868382
- [20] HE, K., X. ZHANG, S. REN, and J. SUN. Deep Residual Learning for Image Recognition. Online. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA, 2016-06-27 – 2016-06-30. IEEE, 2016. ISBN 9781467388511. DOI: 10.1109/cvpr.2016.90. [viewed 2023-09-27].
- [21] SZEGEDY, C., V. VANHOUCKE, S. IOFFE, J. SHLENS, and Z. WOJNA. Rethinking the Inception Architecture for Computer Vision. Online. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA, 2016-06-27 – 2016-06-30. IEEE, 2016. ISBN 9781467388511. DOI: 10.1109/cvpr.2016.308. [viewed 2023-09-27].
- [22] SIMONYAN, K., and A. ZISSERMAN. Very Deep Convolutional Networks for Large-Scale Image Recognition. Online. In: International Conference on Learning Representations. Visual Geometry Group, Department of Engineering Science, University of Oxford, 2015. arXiv 1409.1556 Available at: <https://doi.org/10.48550/arXiv.1409.1556>. [viewed 2023-09-27].
- [23] HOWARD, A. G., M. ZHU, B. CHEN, D. KALENICHENKO, W. WANG, T. WEYAND, M. ANDREETTO, H. ADAM. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. Preprint; online. Google Inc., 2017. arXiv:1704.04861 Available at: <https://doi.org/10.48550/arXiv.1704.04861>. [viewed 2023-09-27].
- [24] Advanced Guide to Inception v3 | Cloud TPU | Google Cloud. In: Google Cloud [online]. [no date] [viewed 12 March 2024]. Available at: <https://cloud.google.com/tpu/docs/inception-v3-advanced>.

## About Authors

**Anjana Guru PRASAD** received her BE in Electronics and Communication Engineering from BMS Institute of Technology and Management, Bangalore, India. Her research interests include Machine learning, Signal processing and Deep learning.

**Shruti KULKARNI** received her BE in Electronics and Communication Engineering from BMS Institute of Technology and Management, Bangalore, India. Her research interests include Image Processing, Deep learning, and Artificial Intelligence.

**Vaishnavi Suresh BHANGENAVAR** received her BE in Electronics and Communication Engineering from BMS Institute of Technology and Management, Bangalore, India. Her research interests include Image Processing, Signal Processing, Communication and Deep learning.

**Vineet N BELAGOD** received his BE in Electronics and communication Engineering from BMS Institute of Technology and Management, Bangalore, India. His research interests include Image processing, Data processing and Machine learning.

**Vijayalakshmi G.V. MAHESH** (corresponding author) received her BE in Electronics and Communication Engineering from Bangalore University, India in 1999, and M.Tech in Digital Communication and Networking from Visvesvaraya Technological University in 2005 and the Ph.D. degree from the Vellore Institute of Technology, Vellore, India. Currently she is working as an Associate Professor at BMS Institute of Technology and Management, Bangalore, India. Her research interests include Machine Learning, Image Processing, Pattern Recognition and Deep learning.

**Alex Noel Joseph RAJ** (Member, IEEE) re-

ceived the B.E. degree in electrical engineering from Madras University, India, in 2001, the M.E. degree in applied electronics from Anna University, in 2005, and the Ph.D. degree in engineering from University of Warwick, in 2009. From October 2009 to September 2011, he was a Design Engineer with Valeport Ltd., Totnes, U.K. From March 2013 to December 2016, he was a Professor with the Department of Embedded Technology, Vellore, India. Since January 2017, he has been with the Department of Electronic engineering, College of Engineering, Shantou University, China. His research interests include deep learning, signal learning, signal and image processor and FPGA implementations.