

Linear fusion approach to convolutional neural networks for facial emotion recognition

Usen Dudekula, Purnachand N

School of Electronics Engineering, VIT-AP University, Amaravati, Andhra Pradesh, India

Article Info

Article history:

Received Aug 12, 2021

Revised Dec 31, 2021

Accepted Jan 21, 2022

Keywords:

Convolutional neural networks

Deep learning

Machine learning

Transfer learning

Visual geometry group

ABSTRACT

Facial expression recognition is a challenging problem in the scientific field of computer vision. Several face expression recognition (FER) algorithms are proposed in the field of machine learning, and deep learning to extract expression knowledge from facial representations. Even though numerous algorithms have been examined, several issues like lighting changes, rotations and occlusions. We present an efficient approach to enhance recognition accuracy in this study, advocates transfer learning to fine-tune the parameters of the pre-trained model (VGG19 model) and non-pre-trained model convolutional neural networks (CNNs) for the task of image classification. The VGG19 network and convolutional network derive two channels of expression related characteristics from the facial grayscale images. The linear fusion algorithm calculates the class by taking an average of each classification decision on training samples of both channels. Final recognition is calculated using convolution neural network architecture followed by a softmax classifier. Seven basic facial emotions (BEs): happiness, surprise, anger, sadness, fear, disgust, and neutral facial expressions can be recognized by the proposed algorithm, The average accuracies for standard data set's "CK+," and "JAFPE," 98.3% and 92.4%, respectively. Using a deep network with one channel, the proposed algorithm can achieve well comparable performance.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Purnachand N

School of Electronics Engineering, VIT-AP University

Amaravati, Andhra Pradesh, India

Email: chandunece@gmail.com

1. INTRODUCTION

Emotional expressions are the most important as it encodes non-verbal ways of expressing interior emotions and intentions. The facial action coding system (FACS) is a suitable structure that uses action units (AU) to identify human facial behavior on the face [1]. Multifarious methods confide in the extraction of the facial region [2] or an automatic solution to identification [3]. Primary facial recognition extraction methods include local directional pattern (LDP), linear discriminant analysis (LDA), local binary patterns (LBP), principal component analysis (PCA), and convolutional neural networks [4]. AUs are core indicators, making a declaration about the correspondent emotion activation point [5], [6]. AUs may not only be utilized to reveal emotions and textures. When the face reveals emotion and multiple filters can be applied for the exposure of facial emotions Convolutional neural networks are specific kinds of artificial neural networks and have been working with reasonable performance as a feature extractor [7]. Facial expression identification typically requires three phases of preparation consisting of the acquisition of the face, the retrieval of the features, and the classifier [8]. Despite recent rapid advances, facial emotion recognition (FER) remains difficult due to a variety of challenges such as improvements in lighting and accessories,

partial occlusions, capturing head deflection of the facial areas, the identification rate remains low due to the influence of different ambient as well as differences in individual people's traits. Many characteristics may be retrieved and learned for a good face expression detection system using deep learning and specifically convolutional neural networks (CNNs). Noteworthy is that in the case of facial expressions, many indications originate from a few regions of the face, such as the lips and eyes, while other portions of the face play a little role in their production. As a result, the machine learning framework should ideally just focus on the most essential portions of the face, while being less sensitive to other facial regions. Handcrafted features for FER assignments are no longer adequate. A suboptimal solution to these problems can be offered by deep learning methodologies. In most fields of machine learning and computer vision, CNN has been highly successful, the researchers use CNN for object detection in a wide range of applications [9], [10]. We present a deep learning-based system for face expression identification in this paper that considers the preceding observation. FER tasks rely on detecting facial expressions and identifying faces based on RGB or grayscale pictures. Traditional FER tasks depend on hand-crafted features. Features may be divided into three primary categories: appearance, geometry, and motion characteristics, respectively. Pixel intensity [11], Gabor texture [12], LBP [13], and histogram of oriented gradients (HOG) [14], are some of the most often utilized appearance characteristics. These features from the full facial region are considered, but the eyes, nose, and mouth are not taken into consideration. Therefore, FER tasks employ geometric characteristics, which are represented by the geometric connections of facial landmark points identified from local areas that are significantly linked to expression variations [15]. Furthermore, combining different features is a trend that has great potential [16]. Two-stage multi-task framework to explore FER. With the use of linear-chain orticotropin-releasing factor (CRF), hidden CRF, and hidden layer variables [17] created an interactive, multi-dimensional model of the hidden layer. As a result of this mode, a similarity analysis is used to determine how an expression changes. Already existing methods for facial recognition using hand-crafted features have a limited recognition capability.

Numerous investigations have recently in consideration of deep learning, studied FER problems in pattern recognition, FER has had remarkable success [18]. Using deep belief networks (DBNs), trained a multi-layer perceptron (MLP) to detect distinct facial expressions based on the learning features. MLP surpasses both SVM and RF classifiers [19]. CNN for FER and reported its satisfactory performance in the "CK+" dataset. A data augmentation strategy was proposed to address the lack of labeled samples for CNN training. Several pre-processing technologies were also used to preserve expression-related features in facial images [20]. Combined several CNNs to study FER [21]. It was possible to combine these CNNs by learning the set weights of the network response also trained several deep CNNs for robust FER [22]. AU-inspired deep networks (AUDNs) [23]. As a result of AUDN's focus on a single face picture input modality, its recognition capabilities are limited. In a highly deep neural network better characteristics specific for expression representation. Four inception layers followed a max-pooling layer and two convolutional networks. However, it is impossible to train this network without the use of computational power (especially GPUs) [24]. Novel facial recognition approach using a guided image filter and a convolutional neural network [25]. Handcrafted features are the foundation of FER approaches. The use of facial depth pictures as an input to deep networks is, unfortunately, rare. All of the previous studies have made substantial progress in the field of emotion identification when compared to previous efforts, but they lack a clear technique for identifying key facial areas for emotion detection. By utilizing a linear fusion network-based architecture, we attempt to solve this issue by focusing on the emotions with an accuracy of standard data set's "CK+," and "JAFPE," 98.3% and 92.4%, respectively. This paper focuses on the problems of characteristics extraction and facial expression detection. First of all, binary facial image channels, including gray images and LBP images, will be used by FER using convolution neural networks. Secondly, a methodology for fine-tuning is used to optimize the use of a well-trained pre-Trained network (VGG19 model on ImageNet). Provided by linear fusion to both channel outputs. Final recognition is calculated using convolution neural network architecture followed by a softmax classifier. Processes the outcomes and does facial expression projections from benchmark facial expressions (happiness, sadness, anger, surprise, disgust, fear, and natural). Improvements on VGG16 leading to VGG19 overcomes AlexNet's limitations and improves recognition accuracy. VGG19 consists of some unnecessary ReLU units in the center of the network in contrast to VGG16 [10]. Our analysis is structured as follows. Section 2 is about the introduction. Section 2 is about the overview of the dataset's and also includes comprehensive CNNs and some pre-trained deep network models. The proposed model is defined in section 3. The test results and interpretation for section 4 are announced. Section 5 provides the conclusion statement.

2. DATASET'S DESCRIPTION

Emotion plays a major role in interpersonal communication and also in improving lifestyle. Face emotion detection can also extract a set of face-related attributes, such as head pose, age, emotion, facial hair,

and glasses. These attributes are common predictions, not authentic classifications. When users add themselves to a face service few attributes are useful, to make sure that the application is getting high-quality face data (for example, our application could advise users to take off their sunglasses if the user is wearing sunglasses). The trial was consummate on two publicly accessible facial expression databases: JAFFE (Japanese female facial expression) and Cohn-Kanade (CK+). We test the proposed methodology extensively in our dataset in facial emotional expression classification. The experimental results of the proposed method outperform the most state-of-the-art facial emotional expression recognition systems. Data sets are used for training and 10-fold cross-validation testing purposes by splitting the data into 80%, training, and 20%, for testing. The data sets' statistics are shown in Figure 1.

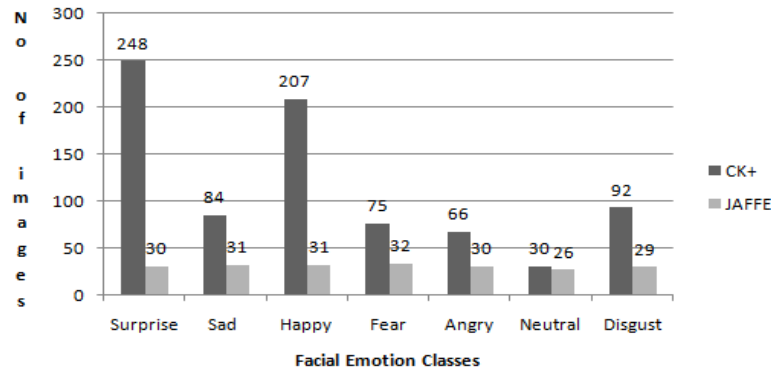


Figure 1. Data set's statistics

2.1. Convolutional neural networks

Compared to hand-crafted functionalities, CNN instantly memorizes functions for learning deep visual variations by using a wide variety of training data and can easily equate its test process on acceleration GPU cores. The following layers are included in CNN architecture.

2.1.1. Convolutional layer

Performance over the input is transformed into ground breaking layers. Enable V_k to be the kernel size n, m filter added to the input X and the number of CNN neuron input ties is n and m . The resulting layer output is determined accordingly.

$$C(X_{u,y}) = \sum_{i=-n/2}^{n/2} \sum_{j=-m/2}^{m/2} V_k(i,j)X_{u-i,y-j} \tag{1}$$

2.1.2. Max polling

Max Pooling is a convolution method in which the Kernel takes the highest value from the region it convolves. Max Pooling basically tells the Convolutional Neural Network that only that information will be carried forward if it is the greatest information available in terms of amplitude. By using the maximum function, where m is the filter size the output is calculated a follow. Max Pooling reduces X_i input.

$$m(X_i) = \max \{x_{i+1,j+1} \mid K \leq \frac{m}{2}, l \leq \frac{m}{2} k, l \in N\} \tag{2}$$

2.1.3. Rectified linear unit (ReLU)

The rectified linear activation unit, or ReLU, is one of the few landmarks in the deep learning revolution. It's simple, yet it's far superior to reaming activation functions like sigmoid or tanh. ReLU is known to be used by the neural cell to calculate its X , output using the following activation function.

$$R(X) = \max (0, x) \tag{3}$$

2.1.4. Fully connected layer (FC)

The output from the convolutional layers represents high-level features in the data. While that output could be flattened and connected to the output layer, adding a fully-connected layer is a best way of learning non-linear combinations of these features. FC layer also termed as multi-layer perceptron (MLP) binds all of the preceding layers' neurons to each neuron of its layer. The number of neurons in the completely connected layer was defined as X with a size of K and l .

$$f(x) = \sigma(w * x) \quad (4)$$

2.1.5. Output layer

There are three sorts of layers in a standard neural network: one or more input layers, one or more hidden layers, and one or more output layers. More advanced, novel neural networks may include several layers of any sort, and each layer may be designed differently. The output layer is one of the hot vectors that represent the input image class. Therefore, the number of groups is dimensional. The output vector class X is derived.

$$c(x) = \{i | \exists A_j \#i : x_j \leq x_i\} \quad (5)$$

2.1.6. Softmax layer

A neural network's activation function is an essential component. A neural network is a basic linear regression model without an activation function. The activation function offers the neural network non-linearity. The fault is distributed back over the Soft-max. Enable N to be the input vector dimensions and soft-max will then calculate mapping to,

$$S(X_i) = \sum_{i=1}^n e^{X_i} \quad (6)$$

2.2. Pre-trained models for classification

We define pre-trained model types for classification. Training when we were allotted to a deeper network. Further, training a deeper network is complicated by overfitting and convergence effects. These types of issues are optimally resolved in the transfer learning architectures. Transfer learning models are disunited into four states: instance, parameter-based, feature-representation, and relational-knowledge. Transfer learning deploys a pre-existing architecture, trained on a large dataset, and further processing in the next stages. Our work guarantees consistency by reducing the costs of training with new models of deep learning. In this article, the top five pre-trained models for image classification that are state-of-the-art and are widely used in the industry as well as research. The individual models are explained in much more detail below. That is inception-v3, VGG-16, VGG-19, ResNet50 and EfficientNet.

2.2.1. Inception

The image dataset was used in the 2012 large scale visual recognition challenge (ILSVRC). Inception-v3 supplies elevated accuracy outcomes in comparison with Inception-v1 and Inception-v2. Table 1 shows the architecture of inceptionv3. CNN is the convolution layer, P is the maximum-pooling layer, MX is the mixed layer and F is the fully connected layer. Inception v3 parameters are less than 25M, it enables path fusion, length of feature 2048, and the training set contains Image net 126M images of 1k subjects.

Table 1. Inception V3 architecture

input image size	Architecture
229x229x3	CNN0, CNN1: C2:64x3x3, P0:3x3 CNN3:80x1x1, CNN4:192x3x3 P1:3x3, M0:35x35x256 MX1, MX2:35x35x288 MX3, MX4:17x17x768 MX5, MX6, MX7:17x17x768, MX7:17x17x768, MX8:8x8x1280 MX9, MX10:8x8x1280, P3:8x8, F:1000

2.2.2. Visual geometry group (VGG)

VGG is an object recognition model pre-trained, such as Inception-v3. In this article, we discuss two variants of visual geometry group models namely- VGG16 and VGG19. Both uses features extracted from the pre-trained VGG16 and VGG 19 networks trained on the ImageNet dataset for detecting facial expressions. The input greyscale images are pre-processed during the training step by performing strength normalization and resizing on the pixel values. However, VGG should be designed with 224x224x3 inputs, we specify `pooling = 'avg'` before classification layer (that way VGG can handle inputs of any size). In another way, VGG16 supports down to 48x48 images as an input. We instantiate our model with `Keras.applications.vgg16.VGG16(include_top=True, weights='imagenet', input_shape=(72,72,3))` and then add own model head fully convolutional or dense. These images are given as input to the VGG network. The VGG contains five pooling layers. Fine-tuning enables one to update the model architecture by eliminating the layer heads that were previously entirely connected, offering new, newly initialized layers, and training the new FC layers to predict our input group. Training minimizes the average prediction and log-loss after the softmax layer. CNN is the convolution layer, P is the maximum pooling layer, MX is the mixed layer and F is the fully connected layer. VGG16 parameters 138M, it disables path fusion, the length of feature 4096, and

training set containing ImageNet 126 M images of 1k subjects. VGG19 has 144M parameters, it enables path fusion, the length of feature 4096, and the training set contains ImageNet of 126M images of 1k subjects. The architecture of VGG16 and VGG19 architecture shown in Table 2 and Table 3.

Table 2. Vgg16 architecture

Input image size	Architecture
224 X 224 X 3	CNN0,1: 64x3x3 P0:3x3, CNN2,3:128x128x3, P1:3x3 CNN4,5:256x256x3 CNN6:256x256x3, CNN7:512x512x3, P2:3x3 CNN8:256x256x3, CNN9:512x512x3, P3:3x3 CNN10,11:512x512x3 CNN12:512x512x3, F0, F1:4096, F2:1000

Table 3. Vgg19 architecture

Input image size	Architecture
224 X 224 X 3	CNN0,1: 64x3x3 P0:3x3, CNN2,3:128x128x3, P1:3x3 CNN4,5,6,7:256x256x3P2:3x3, CNN8:512x512x3, CNN9,10:512x512x3, CNN11,12:512x512x3, P3:3x3CNN13,14:512x512x3 CNN15:512x512x3, F0,1:4096, F2:1000

2.2.3. ResNet-50

ResNet-50 is an image recognition model which has been pre-trained deep learning of the CNN, a subset of deep neural networks. ResNet-50 is 50 layers deep and is trained on 1 million images in 1,000 groups in the ImageNet database. In addition, the algorithm has more than 23 million workable parameters, suggesting a deep architecture that allows image recognition easier. A highly efficient solution is a prequalified model. Specifically, the five stages of residual blocks consist of the ResNet-50 model. Every residual block has three layers of 1*1 and 3*3 convolution blocks. Every layer is fed into the next layer in traditional neural networks. Each layer in a residual block network flows into the next layer, called the identity connections; roughly 2–3 hop apart. The notation $k \times k, n$ in the CNN layer block shows the filter size k and n channels. The completely linked layer of 1000 neurons is denoted by fully connected (FC) 1000. Django rest framework (DRF) is a library that creates a versatile and efficient application programming interface using Django basic models. ResNet-50 architecture is shown in Figure 2. From the experimental outcomes, we observed that the VGG19 model achieved high accuracy on both data sets compared to the ResNet-50 since it is a very deep network with 144 million parameters (trainable and non-trainable). Due to this behavior VGG network efficiently captures the features (edges) by minimizing the overfitting problem. EfficientNet-B0 to B7, parameters are 5.3 million to 66 million, we take top three accuracy values of each EfficientNet-B0 to B7, the accuracy also increase several parameters increase model, the parameters like trainable and non-trainable can changes according to model weights, bias, for pre-trained model most of the cases it will use listed parameters in Tables 4 and 5. The bias and weights initially take random values, then we update using the backpropagation algorithm. These learned parameters are updated weights correctly, the model will predict the accurate prediction. The learned parameters are not updated properly, the recognition rate is down. The intuition behind selecting ImageNet and not selecting other architectures like ResNet-50, MobileNet, and EfficientNET. ResNet, MobileNET's are faster and smaller than other major networks like VGG-19 for there is a small trade-off and that trade-off is accuracy compared to ImageNet, MobileNET's are typically aren't as accurate but relatively small accurate reduction, top State-of-the-Art pre-trained models for image classification.

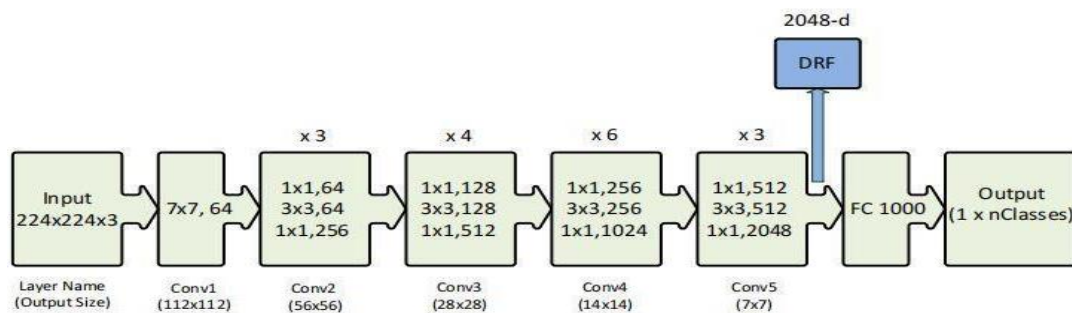


Figure 2. ResNet-50 architecture [26]

Table 4. Efficient Net's family efficiency on ImageNet

Model	Top accuracy	Parameters
EfficientNet-B0	77.3%	5.3M
EfficientNets-B1	79.2%	7.8M
EfficientNets-B2	80.3%	9.2M
EfficientNets-B3	81.7%	12M
EfficientNets-B4	83.0%	19M
EfficientNets-B5	83.7%	30M
EfficientNets-B6	84.2%	43M
EfficientNets-B7	84.2%	66M

Table 5. Pertained models for top accuracy

Model	Top accuracy	Parameters
ResNet-50	77.15%	25M
Inception_v3	78.8%	24M
VGG-16	78.2%	138M
VGG-19	84.2%	144M

2.2.4. EfficientNets

The EfficientNet category consists of eight models from B0 to B7, each with a corresponding number of models referring to variants with higher parameters and greater accuracy. EfficientNet is a community of neural network models. Since 2012, the ImageNet dataset has expanded as it has grown more complex. However, most of them are not efficient in terms of load handling. More efficient methods for smaller models have been introduced in recent years. So much so that when scaling down the model, scaling is performed on depth, distance, and resolution focusing on all three in combination has yielded more efficient results.

3. PROPOSED MODEL

This work proposes a new approach for facial emotion recognition using a linear fusion technique. The proposed method consists of the following steps: feature extraction from CNN and feature extraction from vgg19, and finally fusion of different outputs. Recognizing facial emotions with the help of the proposed model. When it comes to face detection, the present investigation uses the widely utilized Viola-Jones framework [27]. The entire methodology is depicted in Figure 3. The CK+ [28] and the JAFFE [29] facial pictures are used in benchmarking datasets, and real environments vary in rotation, even for the same subject. This variation is independent of facial emotions and may impact the accuracy of FER's recognition. The face area is aligned using rotation rectification to solve this problem.

$$[R_x, R_y, 1] = [R_x', R_y', 1] = \begin{bmatrix} \cos \theta & \sin \theta & 0 \\ -\sin \theta & \cos \theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad (7)$$

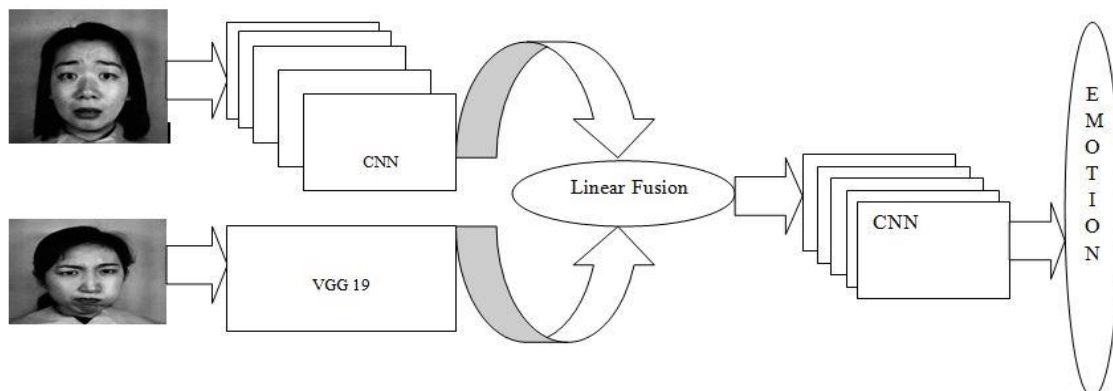


Figure 3. Proposed model

In which (R_x, R_y) represents the initial point in the face picture, and (R_x', R_y') represents the coordinate (x, y) after the rotation transformation is performed. θ denotes the angle of rotation from one eye center to the other, as measured from the center of the eye. On the horizontal axis, the axis starting value is zero, all identified face areas are reduced to 72×72 pixels after rotation rectification. While a smaller face area might increase the speed of FER, it can also lead to losing facial features, especially for the information acquired from facial LBP images, related facial regions, such as mouths, eyes, and eyebrows, are more remarkable in LBP images than in grayscale images.

Our design is to amend the exactness of emotion expression classification by modern convolutional network architecture. Including gray images and LBP images, will be used by FER using convolution neural networks. Activations of rectified linear unit (ReLU) after each convolution layer are added and soft-max classifiers are used for an activation function in the flattening layer. The first step of feature extraction from convolution neural network architecture is shown in Table 6. Then, VGG 19 is chosen for Facial features extraction, applying transfer learning our partial VGG19 network is trained on the Flatten model ImageNet dataset with a dropout ratio of 0.5, 64 dense layers of ReLU architecture, and seven activation points. Softmax classifier is deployed in the same. The input greyscale images are pre-processed during the training step by performing strength normalization and resizing on the pixel values. These images are given as input to the VGG network. The VGG contains five pooling layers. Rather than using VGG16, it is optimal to use VGG19 (VGG19 has more memory), and the best performance.

The input data has a size of $1 \times 72 \times 72$. After that, we'll work on the first four blocks. The fifth block Conv5_1ft, Summarizes the parameters for the layers in this block, which are mentioned in Table 7. The learning rates of the fifth block's layers are (0.001 used for layers of the fifth block) than their original values (0.01 used for layers of previous blocks) to ensure that they can learn more effective information. Because we decreasing 10 times the original value. Weight decay, which reduces your coefficients to zero, guarantees that you reach a local optimum with small-magnitude parameters. This is critical to prevent an over-fitting situation. In addition, by increasing the convexity of the objective function, the model becomes easier to optimize. When it comes to optimal weight decay, it depends on the total number of batch runs and weight updates.

Table 6. CNN architecture for image classification

Layer	Filters	Kernel Size	Stride	Activation Function
Convolution(C1)	64	7x7	4x4	ReLU
Pooling		3x3	2x2	
Convolution(C2)	32	5x5	1x1	ReLU
Convolution(C3)	64	5x5	1x1	ReLU
Convolution(C4)	32	5x5	1x1	ReLU
Pooling		3x3	2x2	
Flatten				softmax

Table 7. Parameters for the proposed model

Parameters	Value
Learning rate	0.001
Weight Decay	0.01
Momentum	0.001
Optimizer	Adam

This is critical to prevent an over-fitting situation. In addition, by increasing the convexity of the objective function, the model becomes easier to optimize. When it comes to optimal weight decay, it depends on the total number of batches runs and weight updates. According to our empirical research of Adam, the lower the optimum weight decay is, the longer the runtime/number of batch runs are. Momentum is utilized to reduce weight change variations across successive iterations, and it started at a zero initial value. The partial VGG19 network is used to extract an expression related feature vector from face grayscale images. The CNN is used to extract a feature vector from LBP face pictures. There are two cascaded full-connect layers on each feature vector to reduce the size. Both training models obtain corresponding model weights. Corresponding weights to linear fusion approach to new model CNN architecture. CNN is the most common network model among the many possible deep learning models. The final CNN architecture is shown in Table 8. The VGG 19 system uses a tuning system to extract emotions from gray-level images. The vector is derived by CNN from LBP face images. The two cascaded fully connected layers are communicated by each vector. VGG19 and CNN both construct a fused vector $fu = \{p1, p2 p7\}$. The first element is followed by (8),

$$F_u = \alpha \cdot S_i (1 - \alpha) I_i \quad (8)$$

where α weights of the gray-scale images. Evaluated by cross-validation. The categorical probability distribution of soft-max, the input is a set of multi-class.

$$Y_i = \sum_{j=1}^k \frac{e^{X_j}}{e^{X_j}} \quad (9)$$

The cost function, which is defined by $f(y = k/x)$.

$$Loss(y, z) = -\sum_{i=1}^k Z_i \cdot \log(Y_i) \quad (10)$$

True label Z_i indicates and Y_i is the output of a soft-max function. Backpropagation is based on the optimization algorithm of gradient descent.

Table 8. Proposed CNN architecture

Layer	Filters	Kernel Size	Drop out	Activation Function
Convolution(C1)	6	5x5		ReLU
Pooling		2x2		
Convolution(C2)	16	5x5		ReLU
Pooling		2x2		
Convolution(C3)	120	5x5	0.25	ReLU
Flatten			0.5	ReLU
Dense (84)				
Dense layer				Softmax

The novelty of the proposed methodology is linear weighted fusion is a procedure where the resulting fused picture is more informative and complete than any of the input of the picture by combining relevant information from a set of images into the individual picture. Image fusion techniques can improve the quality of application data. The merger technique was used to integrate VGG-19 and CNN decisions. The fusion algorithm calculates the class by taking an average of each classification decision on training samples. More reliable (higher precision) classification is weighed and significantly contributes to decision making. We have used the fusion rule of the weighted sum to evaluate the best opportunity of a fusion result of a particular emotion, comparison multimodal with unimodal models. Due to the high capacity of deep CNN, the average feature fusion model improves the performance substantively. Taking an average of multiple models will reduce the variance. This shows that multimodal offers a very effective way of improving accuracy rates, as originally proposed.

4. EXPERIMENTAL RESULTS AND ANALYSIS

Built on the Tensor flow system, we test the achievement of the proposed model on the windows internet Google Colab platform. Two publically available datasets are used for facial emotion images. The results are illustrated in Figures 4(a) and (b). Displaying precision curve by red and loss by green and stabilizes loss following 25 to 50 epochs, and their test results are also shown. The average accuracy of recognition for 'CK+ and JAFFE' data sets is 98.3% and 92.4%, respectively. We also measure our process by assessing accuracies based on single-channel facial images. We test the feasibility of our methodology. Compared to other CNN related approaches, our approach produces improved efficiency with features like histogram of oriented gradients (HOG), support vector machine (SVM), and K-NN Our method's benefit is accomplished by making good use of the complementary of various facial image sources, while the other strategy only uses FER approaches. Comparisons between our approach and the other state-of-the-art FER approaches are shown in Table 9.

4.1. Some of the classification faults are discussed

The SVM algorithm is not suitable for large sets of data. SVM does not do that well as there is more noise in the collection, i.e. overlap of target groups. In the scenarios where the number of features for each data point exceeds the number of training data samples, the SVM would be underperforming. And the SVM Model is a representation of examples as dots in space, mapped in such a way that the examples of the Separate categories or classes are classified into two categories. Division of the plane that maximizes the

margin between the two there is various groups. This is because the separation plane has the widest distance to the Lower, the closest training data points in any class. Error in a generalization of the total classifier.

HOG has a decent score for human identification. However, it has a drawback that is very susceptible to the rotation of the image. KNN-Precision relies on data accuracy. With large data, the predictive stage can belong. Responsive to the data size and insignificant functions. The KNN algorithm's drawback is that it uses both Equal characteristics with correlations in computation. This can lead to mistakes in classification, particularly if there is just a small subset of traits that are helpful for classification.

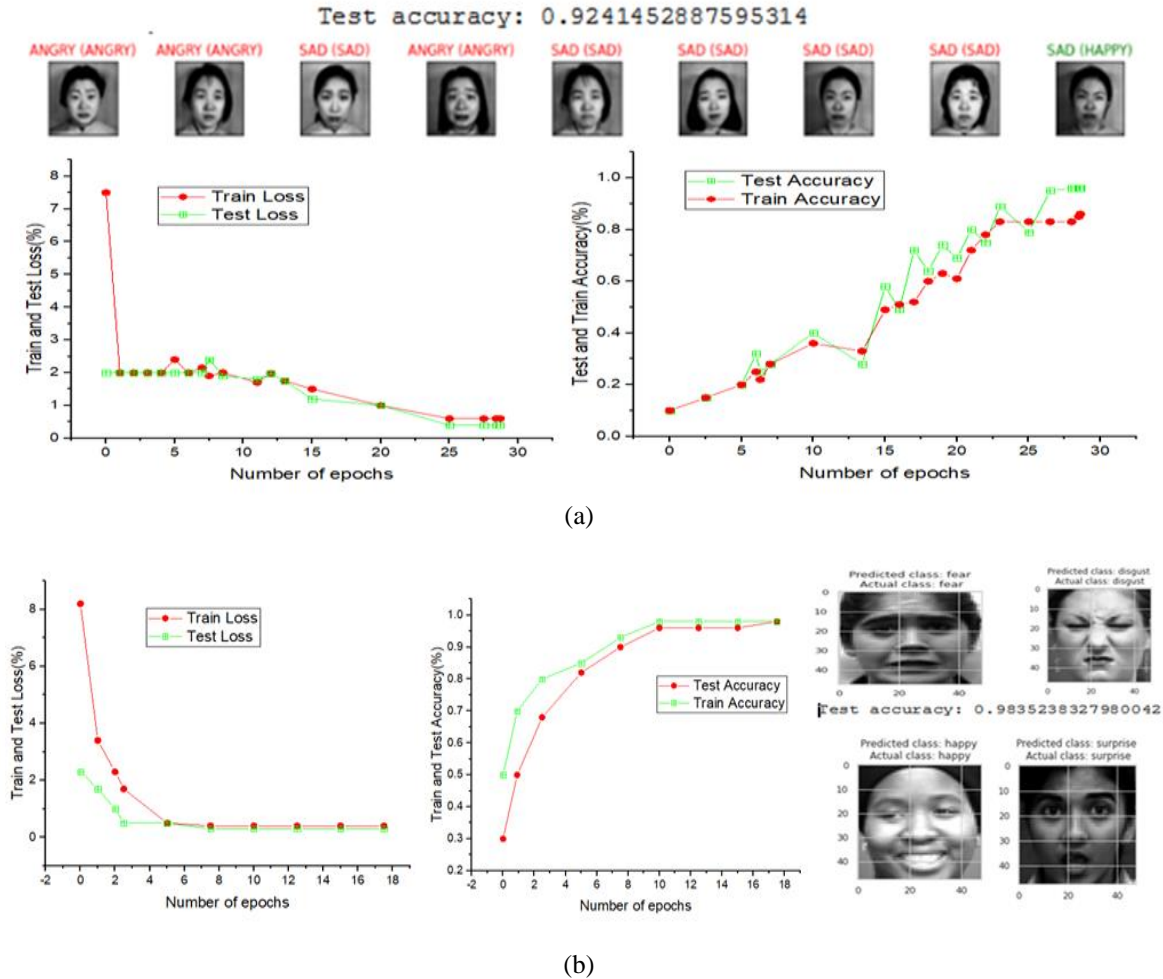


Figure 4. JAFFE training accuracy and validation loss, (a) and their test analysis test results, CK+ training accuracy and validation loss and (b) their test analysis test results. Number of epochs in x axis, train and test loss, their accuracy (%) in y axis

Table 9. Comparison of recognition rate with some of the existing techniques

Method	JAFFE Database	CK+ Database
LBP+ORB features [30]	88.55%	
Fisher face [31]	89.2%	
Deep Features + HOG [32]	90.58%	
Sun <i>et al.</i> [33]	92.00%	
CNN [34]	76.5%	
Proposed Model	92.4%	
LBP+ORB features [30]		93.2%
Deep features + HOG [32]		94.17%
Inception [35]		93.2%
Dynamic cascade classifier [36]		97.8%
Attentional convolutional [37]		98.0%
Proposed Model		98.3%

4.2. Qualitative analyses of the proposed model

Webcam facial image samples are obtained for testing. The robustness of our procedures is measured. The facial area observed in each sub-figure is a green rectangular. Figures 5(a) and (b). Demonstrates some good identification cases: Extremely precise facial gestures. The accuracy of multiple facial expressions is over 0.95 like anger and happy, including when a notebook occludes the topics partly Figure 5(c). Illustrates some instances of failed facial recognition, as "Unknown" or mislabel.

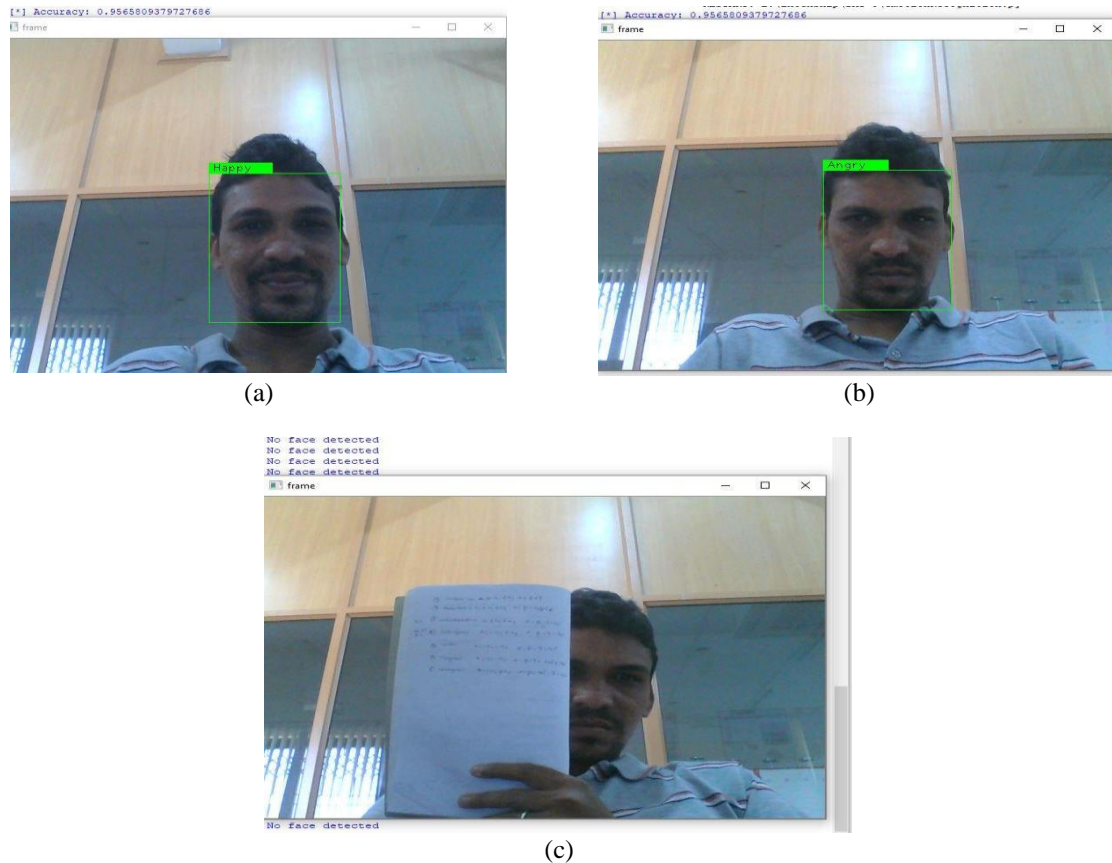


Figure 5. Using facial web cam of successful recognition of facial expressions with better accuracy for the: (a) happy expression in angry partial occlusions, (b) head deflection and a notebook occludes partially, and (c) failed to recognition facial expression

5. CONCLUSION

In this study, an improved FER technique can process both grayscale and LBP pictures of the face at the same time. We contend that the two picture channels now being utilized complement one another, collect both local and global information from facial images, and enhance recognition. To make full use of the characteristics that have been retrieved from the various picture channels, a weighted fusion method is presented. To automatically extract the characteristics of face emotions from facial grayscale pictures, a partial VGG19 network is built. To train the network using ImageNet's basic parameters, fine-tuning is performed. CNN is built to automatically extract face expression characteristics from LBP pictures. After that, a weighted fusion approach is presented to combine the two characteristics to make maximum use of the complementary face information. The results of the recognition are based on the linear fusion method to a novel model of CNN architecture using corresponding weights from a softmax classifier. The proposed methodology was compared to some of the prior methods. Based on the comparison results, it appears that the proposed methodology outperforms some of the existing methods. Based on the outcomes of the experiments, we may infer that our suggested approach can be used for facial emotions recognition. Therefore a modern model must be developed that can lower training time and delivery in real-time applications. Our work for the future will be to simplify the network further and accelerate the algorithm. We plan to concentrate on other facial images channels to expand the fusion network further.




REFERENCES

- [1] T. F. Cooles, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 681–685, Jun. 2001, doi: 10.1109/34.927467.
- [2] I. Kotsia and I. Pitas, "Facial expression recognition in image sequences using geometric deformation features and support vector machines," *IEEE Transactions on Image Processing*, vol. 16, no. 1, pp. 172–187, Jan. 2007, doi: 10.1109/TIP.2006.884954.
- [3] K. Anderson and P. W. McOwan, "A real-time automated system for the recognition of human facial expressions," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 36, no. 1, pp. 96–105, Feb. 2006, doi: 10.1109/TSMCB.2005.854502.
- [4] K. Nurzynska and B. Smolka, "Smiling and Neutral Facial Display Recognition with the Local Binary Patterns Operator," *Journal of Medical Imaging and Health Informatics*, vol. 5, no. 6, pp. 1374–1382, Nov. 2015, doi: 10.1166/jmih.2015.1541.
- [5] V. Bettadapura, "Face Expression Recognition and Analysis: The State of the Art," Mar. 2012, [Online]. Available: <http://arxiv.org/abs/1203.6722>.
- [6] S. L. Happy and A. Routray, "Automatic facial expression recognition using features of salient facial patches," *IEEE Transactions on Affective Computing*, vol. 6, no. 1, pp. 1–12, Jan. 2015, doi: 10.1109/TAFFC.2014.2386334.
- [7] J. Donahue *et al.*, "DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition," Oct. 2013, [Online]. Available: <http://arxiv.org/abs/1310.1531>.
- [8] R. Samad and H. Sawada, "Extraction of the minimum number of Gabor wavelet parameters for the recognition of natural facial expressions," *Artificial Life and Robotics*, vol. 16, no. 1, pp. 21–31, Jun. 2011, doi: 10.1007/s10015-011-0871-6.
- [9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017, doi: 10.1145/3065386.
- [10] M. Shaha and M. Pawar, "Transfer Learning for Image Classification," in *Proceedings of the 2nd International Conference on Electronics, Communication and Aerospace Technology, ICECA 2018*, Mar. 2018, pp. 656–660, doi: 10.1109/ICECA.2018.8474802.
- [11] M. R. Mohammadi, E. Fatemizadeh, and M. H. Mahoor, "PCA-based dictionary building for accurate facial expression recognition via sparse representation," *Journal of Visual Communication and Image Representation*, vol. 25, no. 5, pp. 1082–1092, Jul. 2014, doi: 10.1016/j.jvcir.2014.03.006.
- [12] C. Liu and H. Wechsler, "Gabor feature based classification using the enhanced Fisher linear discriminant model for face recognition," *IEEE Transactions on Image Processing*, vol. 11, no. 4, pp. 467–476, Apr. 2002, doi: 10.1109/TIP.2002.999679.
- [13] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on Local Binary Patterns: A comprehensive study," *Image and Vision Computing*, vol. 27, no. 6, pp. 803–816, May 2009, doi: 10.1016/j.imavis.2008.08.005.
- [14] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn, "DISFA: A spontaneous facial action intensity database," *IEEE Transactions on Affective Computing*, vol. 4, no. 2, pp. 151–160, Apr. 2013, doi: 10.1109/T-AFFC.2013.4.
- [15] H. Kobayashi and F. Hara, "Facial interaction between animated 3D face robot and human beings," in *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, 1997, vol. 4, pp. 3732–3737, doi: 10.1109/icsmc.1997.633250.
- [16] L. Zhong, Q. Liu, P. Yang, J. Huang, and D. N. Metaxas, "Learning Multiscale Active Facial Patches for Expression Analysis," *IEEE Transactions on Cybernetics*, vol. 45, no. 8, pp. 1499–1510, Aug. 2015, doi: 10.1109/TCYB.2014.2354351.
- [17] S. Jain, C. Hu, and J. K. Aggarwal, "Facial expression recognition with temporal modeling of shapes," in *Proceedings of the IEEE International Conference on Computer Vision*, Nov. 2011, pp. 1642–1649, doi: 10.1109/ICCVW.2011.6130446.
- [18] A. Dhall, R. Goecke, J. Joshi, K. Sikka, and T. Gedeon, "Emotion recognition in the wild challenge 2014: Baseline, data and protocol," in *ICMI 2014 - Proceedings of the 2014 International Conference on Multimodal Interaction*, Nov. 2014, pp. 461–466, doi: 10.1145/2663204.2666275.
- [19] X. Zhao, X. Shi, and S. Zhang, "Facial expression recognition via deep learning," *IETE Technical Review (Institution of Electronics and Telecommunication Engineers, India)*, vol. 32, no. 5, pp. 347–355, Mar. 2015, doi: 10.1080/02564602.2015.1017542.
- [20] A. T. Lopes, E. De Aguiar, A. F. De Souza, and T. Oliveira-Santos, "Facial expression recognition with Convolutional Neural Networks: Coping with few data and the training sample order," *Pattern Recognition*, vol. 61, pp. 610–628, Jan. 2017, doi: 10.1016/j.patcog.2016.07.026.
- [21] Z. Yu and C. Zhang, "Image based static facial expression recognition with multiple deep network learning," in *ICMI 2015 - Proceedings of the 2015 ACM International Conference on Multimodal Interaction*, Nov. 2015, pp. 435–442, doi: 10.1145/2818346.2830595.
- [22] B. K. Kim, J. Roh, S. Y. Dong, and S. Y. Lee, "Hierarchical committee of deep convolutional neural networks for robust facial expression recognition," *Journal on Multimodal User Interfaces*, vol. 10, no. 2, pp. 173–189, Jan. 2016, doi: 10.1007/s12193-015-0209-0.
- [23] M. Liu, S. Li, S. Shan, and X. Chen, "AU-inspired Deep Networks for Facial Expression Feature Learning," *Neurocomputing*, vol. 159, no. 1, pp. 126–136, Jul. 2015, doi: 10.1016/j.neucom.2015.02.011.
- [24] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," Mar. 2016, doi: 10.1109/WACV.2016.7477450.
- [25] S. Yallamandaiah and N. Purnachand, "An effective face recognition method using guided image filter and convolutional neural network," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 23, no. 3, pp. 1699–1707, Sep. 2021, doi: 10.11591/ijeecs.v23.i3.pp1699-1707.
- [26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Jun. 2016, vol. 2016-December, pp. 770–778, doi: 10.1109/CVPR.2016.90.
- [27] Y.-Q. Wang, "An Analysis of the Viola-Jones Face Detection Algorithm," *Image Processing On Line*, vol. 4, pp. 128–148, Jun. 2014, doi: 10.5201/ipol.2014.104.
- [28] 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 *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, CVPRW 2010*, Jun. 2010, pp. 94–101, doi: 10.1109/CVPRW.2010.5543262.
- [29] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with Gabor wavelets," in *Proceedings - 3rd IEEE International Conference on Automatic Face and Gesture Recognition, FG 1998*, 1998, pp. 200–205, doi: 10.1109/AFGR.1998.670949.
- [30] B. Niu, Z. Gao, and B. Guo, "Facial Expression Recognition with LBP and ORB Features," *Computational Intelligence and Neuroscience*, vol. 2021, pp. 1–10, Jan. 2021, doi: 10.1155/2021/8828245.
- [31] Z. Abidin and A. Harjoko, "A Neural Network based Facial Expression Recognition using Fisherface," *International Journal of Computer Applications*, vol. 59, no. 3, pp. 30–34, Dec. 2012, doi: 10.5120/9531-3956.




- [32] H. Wang, S. Wei, and B. Fang, "Facial expression recognition using iterative fusion of MO-HOG and deep features," *Journal of Supercomputing*, vol. 76, no. 5, pp. 3211–3221, Aug. 2020, doi: 10.1007/s11227-018-2554-8.
- [33] X. Sun, S. Zheng, and H. Fu, "ROI-Attention vectorized CNN model for static facial expression recognition," *IEEE Access*, vol. 8, pp. 7183–7194, 2020, doi: 10.1109/ACCESS.2020.2964298.
- [34] N. Jain, S. Kumar, A. Kumar, P. Shamsolmoali, and M. Zareapoor, "Hybrid deep neural networks for face emotion recognition," *Pattern Recognition Letters*, vol. 115, pp. 101–106, Nov. 2018, doi: 10.1016/j.patrec.2018.04.010.
- [35] M. Li, H. Xu, X. Huang, Z. Song, X. Liu, and X. Li, "Facial Expression Recognition with Identity and Emotion Joint Learning," *IEEE Transactions on Affective Computing*, vol. 12, no. 2, pp. 544–550, Apr. 2021, doi: 10.1109/TAFFC.2018.2880201.
- [36] A. M. Ashir, A. Eleyan, and B. Akdemir, "Facial expression recognition with dynamic cascaded classifier," *Neural Computing and Applications*, vol. 32, no. 10, pp. 6295–6309, Mar. 2020, doi: 10.1007/s00521-019-04138-4.
- [37] S. Minaee, M. Minaei, and A. Abdolrashidi, "Deep-emotion: Facial expression recognition using attentional convolutional network," *Sensors*, vol. 21, no. 9, Feb. 2021, doi: 10.3390/s21093046.

BIOGRAPHIES OF AUTHORS



Usen Dudekula    received his B. Tech degree in Electronics and Communication Engineering (ECE) from Jawaharlal Nehru Technological University, Anantapuramu, and M. Tech degree in Embedded Systems from Jawaharlal Nehru Technological University, Anantapuramu. Currently, he is a Research Scholar at the School of Electronics Engineering, VIT-AP University, Andhra Pradesh-522237, India. His main research areas include Image Processing, Pattern Recognition, Machine Learning, and DeepLearning. He can be contacted at email: basha.834@gmail.com.



Purnachand N    received his M. Tech degree from VIT University, Vellore, India, and Ph.D. degrees from the University of Aveiro, Aveiro Portugal. He is currently working as Associate Professor at VIT-AP University, India. His areas of research include Image Processing, Video Processing, and Pattern Recognition. He can be contacted at email: chandunece@gmail.com.