JOURNAL OF INFORMATION SCIENCE AND ENGINEERING 37, 15-39 (2021)
DOI: 10.6688/JISE.202101_37(1).0002

A Novel Dual CNN Architecture with LogicMax for Facial Expression Recognition

ALAGESAN BHUVANESWARI AHADIT AND RAVI KUMAR JATOTH
Department of Electronics and Communication
National Institute of Technology, Warangal
Telangana, 506004 India
E-mail: ahadit.ab.ml@gmail.com; ravikumar@nitw.ac.in

Facial expressions convey important features for recognizing human emotions. It is a challenging task to classify accurate facial expressions due to high intra-class correlation. Conventional methods depend on the classification of handcrafted features like scale-invariant feature transform and local binary patterns to predict the emotion. In recent years, deep learning techniques are used to boost the accuracy of FER models. Although it has improved the accuracy in standard datasets, FER models have to consider problems like face occlusion and intra-class variance. In this paper, we have used two convolutional neural networks which have vgg16 architecture as a base network using transfer learning. This paper explains the method to tackle issues on classifying high intra-class correlated facial expressions through an in-depth investigation of the Facial Action Coding System (FACS) action units. We have used a novel LogicMax layer at the end of the model to boost the accuracy of the FER model. Classification metrics like Accuracy, Precision, Recall, and F1 score are calculated for evaluating the model performance on CK+ and JAFFE datasets. The model is tested using 10-fold cross-validation and the obtained classification accuracy rate of 98.62% and 94.86% on CK+ and JAFFE datasets respectively. The experimental results also include a feature map visualization of 64 convolutional filters of the two convolutional neural networks.

Keywords: convolutional neural networks, transfer learning, facial action coding system, action units, Pearson correlation, data augmentation, dlib facial landmark predictor, vgg16, logicMax

1. INTRODUCTION

Facial emotions play a vital role in human nonverbal communication, and it helps to understand the inner feelings of humans. Human emotion analysis requires automated Facial Expression Recognition (FER) of unique facial features. According to research, nonverbal communication may represent nearly 67% of information during the interaction of humans [1]. Facial expressions can be voluntary or involuntary actions that can be usually observed with a naked eye. At times, few expressions may not be visible to the naked eye. Hence, there is a challenge to identify emotions automatically. There is much evidence that few facial expressions can be mapped to a particular emotion like a smiling expression can be related to an emotional state of happiness [2]. Humans have the instinctive, natural ability to comprehend the emotion of a person just by observing facial expressions. There is a rapid attraction in the research of automatic facial emotion recognition in recent years. The applications of this topic include, but are not limited to, human-machine interaction, advanced driver assistant systems (ADAS), clinical psychology, and entertainment busi-

Received December 31, 2019; revised March 18 & April 27, 2020; accepted May 21, 2020.
Communicated by Flavia C. Delicato.

ness. Ekman and Friesen proposed six basic emotions related to cross-cultural studies [3]. The essential facial emotions discussed are Anger, Disgust, Fear, Happiness, Sadness, and Surprise. There are other means of identifying human emotions through speech, text, and other biomedical data such as EEG [4]. Emotion recognition through facial features is simple as it does not require sophisticated sensors or transducers for extracting information.

Fig. 1. Sample images from CK+ database [5].

Automated FER models are first attempted through conventional methods using handcrafted features like non-negative matrix factorization (NMF) [6] and local binary patterns (LBP) [7] that widely use geometrical patterns of salient facial features. There are many classical methods for automated FER models that use computer vision algorithms and machine learning classifiers for feature extraction and segregation according to the patterns observed in the features. The algorithms like local binary patterns (LBP) [7] do extract features by logically analyzing the three-dimensional (height, width, colour) image matrix rather than learning directly from massive datasets. These handcrafted features depend on necessary corners, edges, and other critical spatial features. Classifiers like Decision Trees, Artificial Neural Networks (ANN) are trained with actual labels and are used to classify the emotions. These classifiers take considerable time to train and take less time to predict the test data. Modern FER models use deep learning state of the art techniques like convolutional neural networks that learn the features directly from the massive dataset of images through a series of convolutional layers. The shift in technology dramatically improves the accuracy rate of FER models [8-10]. The main disadvantage of moving to deep learning FER techniques is, it requires a considerable amount of training data to construct a good model. The process of quick training the model is possible only through advanced GPUs that shorten the training duration. FER modelled through deep neural networks have good accuracy in detecting the facial expressions because of its adaptable nature to extract suitable features even in the presence of noise. In this paper, we have designed a novel dual deep convolutional neural network using transfer learning and Logic Max. The entire work contribution of the paper is explained in the below points.

- The proposed model analyzes the facial expressions spatially by introducing two convolutional neural networks for the upper face and lower face emotion classification. The partition of the face into upper and lower parts is achieved using dlib's facial landmark predictor. The detailed process is explained in Sections 3 and 4.1.
- The proposed work investigates the correlation between different facial expressions through exploratory data analysis of respective FACS action units. The process of deriving the correlation matrix is explained in Section 4.4.

- A new layer which is known as "LogicMax" is introduced in this paper. It makes the final logical prediction of emotion by setting up a priority table. LogicMax reads the output of lower face CNN and upper face CNN models and decides the final emotion class of the face. The method of setting up a priority table is explained in Section 4.5.
- The accuracy of the proposed model outperforms other state-of-the-art FER techniques on CK+ [5] and JAFFE [11] dataset. The information about the performance metrics of the model is explained in Section 6. The methods of testing the model performance of different FER techniques are also discussed in this section.

2. RELATED WORK

Paul Ekman [3] has identified anger, disgust, fear, happiness, sadness, and surprise as six basic emotions, later neutral was added to the list. Ekman introduced FACS [12], which is one of the iconic works made in the field of facial emotion recognition, helped many researchers to extend the work in this field. There are numerous works in the field of facial expression recognition. The architectures used in this field can be broadly classified into the following categories

1. Pretraining and fine-tuning based Neural Networks.
2. Multiple feature input networks.
3. Spare blocks and layers based deep neural networks.
4. Ensemble-based deep neural networks.
5. Generative adversarial networks based FER models.

Pretraining and fine-tuning based Neural Networks use pre-trained networks like AlexNet [13], VGG [9], VGG-face [14], and GoogleNet [10]. The motivation behind using these pre-trained networks is to avoid overfitting. Kahou *et al.* [15] discussed the advantages of pre-trained models. The multi-stage fine-tuning method can further boost the performance of the FER. Multiple feature input networks are designed to tackle the problems of image rotation, scaling, and illumination effects. Instead of feeding normal RGB images, handcrafted features like Scale-invariant feature transform (SIFT) and mapped local binary pattern features are given as input to the deep neural networks.

Spare blocks and layers based deep neural networks are used to improve the performance of FER. A novel loss function known as the center loss is introduced to improve the discriminative power of CNN. Center loss [16], along with the softmax layer, is used at the end of the CNN layer to obtain a good threshold for classification. Many loss functions like island loss [17], and triplet loss [18] are deployed into CNN models to boost the discriminative power of FER. Ensemble-based deep neural networks can be again classified into multi-architecture ensembles, feature level ensembles, and decision-based ensembles. Multi-architecture ensemble models [19] use the different error functions like log-likelihood loss and hinge loss to feed weights to respective networks inside the model adaptively. Feature level ensembles concatenate important features derived from different networks in the model into a one-dimensional feature matrix. Decision-based ensembles adapt classification based on rules like majority voting [20], simple average [20], and weighted average [19]. Generative adversarial networks (GANs) are widely used in recent years in the field of facial expression recognition. The models trained with GANs can

perform image synthesis, which is realistic and accurate. They can overcome class imbalance issues in different datasets by adding more training images to the dataset. Zhang *et al.* [21] introduced a GAN-based FER model to synthesize images with various expressions under random poses for multi-view facial expression recognition.

3. ROLE OF FACS IN THE PROPOSED CNN

Deep Neural Networks like Convolutional Neural Networks (CNN) try to derive both low and high-level features automatically through training with good datasets. Low-level features are important lines, edges, and corner points, which can be extremely useful in predicting the overall class. Initial stages of a CNN extract the low-level features, and as we move deeper into the network, it tries to combine these low-level features into a meaningful class. Facial Emotions are tough to classify since the problem is a sub-classification task, which involves identifying the emotional classes that have a very slight variance. Paul Ekman and Wallace V. Friesen developed a system known as Facial Action Coding System (FACS) [12] that identifies different facial expressions on any human face. The important facial features are deconstructed and properly taxonomized according to their property using FACS. FACS helped to generate and classify independent actions of muscles/muscle contraction and relaxation known as "Action Units" (AUs). A combination of different action units on the face denotes a particular emotion, as shown in Table 1. Each emotion triggers different facial expressions, and if the FER model tries to analyze the facial expressions accurately, then the classification of emotions becomes easy. Table 1 explains the importance of different action units for the corresponding emotion.

Table 1. FACS action units for different emotions [12].

Emotion	Facial Muscle	Corresponding Action Units
Anger	Brow lowerer+Upper lid raiser+Lid tightener+Lip tightener	4+5+7+23
Disgust	Nose wrinkle+Lip corner depressor+ Lower lip depressor	9+15+16
Fear	Inner brow raiser+Outer brow raiser+Brow lowerer+ Upper lid raiser+Lid tightener+Lip stretcher+ Jaw drop	1+2+4+5+7+20+26
Happiness	Cheek raiser+Lip corner puller	6+12
Sadness	Inner brow raiser+Brow lowerer+Lip Corner depressor	1+4+15
Surprise	Inner brow raiser+Outer brow raiser+ Upper lid raiser (Slight)+ Jaw drop	1+2+5B+26

Table 2. Intensity level classification of action units in FACS [12].

Alphabet	A	B	C	D	E
Intensity Level	Trace	Slight	Pronounced	Extreme	Maximum

FACS has scaled the intensity of the action units by introducing levels from A to E, where A is the weakest and E as the strongest intensity, as shown in Table 2. From the Table 1, it is evident that various emotional states have the same facial muscle moments, for example, Disgust and Sadness emotions trigger the Lip Corner Depressor (Action unit 16). There is a reasonable probability of misclassifying the emotions due to these

similarities in the different emotion classes. FACS help in modelling an excellent deep neural network by exposing the correlation between the emotions. There is much difference in the emotion class happiness and surprise because there is no intersection of action units in both the classes. FACS can convey important information regarding the probability of differentiating two emotions through the study of their respective action units. We have designed a new architecture that uses FACS information along with dual CNN in predicting the emotion class. The inclusion of FACS information in the CNN model improved the accuracy of the model and helped in a better understanding of the role of action units in emotion classification.

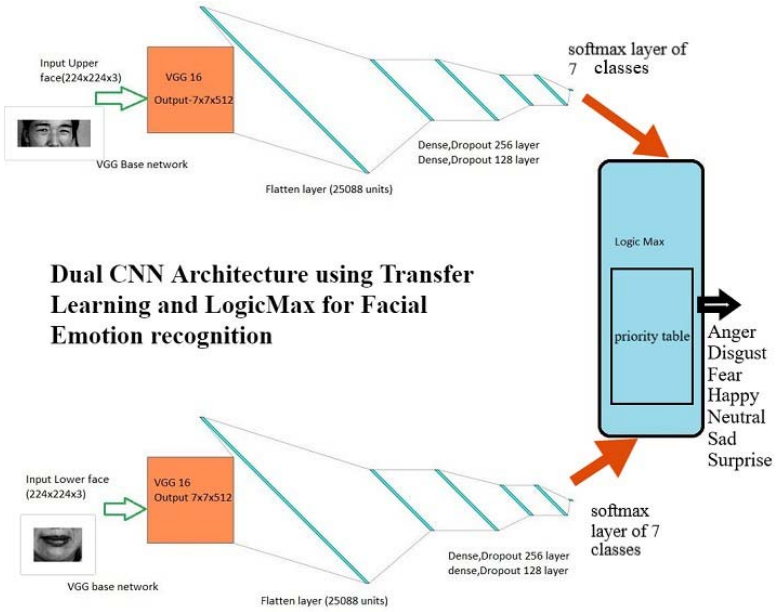


Fig. 2. Overview of the proposed architecture.

On analyzing the action units in emotions, the majority of them lie on the crucial facial landmarks like eyebrows AU (1,2,4), eyes AU (5,7), and lips AU (23,16,26,12,23). A face can be symmetrically divided into two parts, either vertically or horizontally. When a face is split vertically, two images share identical action units since both are mirror images. When a face is horizontally split, we obtain two asymmetric images that have different landmarks. The upper half has eyebrows, eyes, and nose as essential landmarks, and the lower part of the face has a mouth and chin as crucial landmarks. To improve accuracy, we designed two separate deep convolutional networks to identify the emotion on both the upper and lower parts of the face. Each convolutional neural network tries to extract different features in their respective sections (upper or lower region of the face). In doing so, CNN models can spatially concentrate on feature extraction of their respective landmarks. This complex architecture improves the efficiency of the model and also trace each landmark's behavior in different emotions. Some action units are more pronounced when compared to others, for example, emotions like surprise. In some situations, the emotion predicted on the upper face contradicts the emotion predicted on the other half of the face.

During the mismatch, there has to be a logical conclusion on the emotion of the subject. A new layer called "LogicMax" helps to solve the problems of mismatch by building a priority table. LogicMax is a layer that is fitted at the end of the two CNN models to take a logical conclusion of the emotion class of the overall face. LogicMax is a novel layer that predicts the final emotion class of the overall face by analyzing both the outputs of two CNN. Thus, FACS information along with a logical approach can be imparted inside the logicMax layer to improve the efficiency of classification.

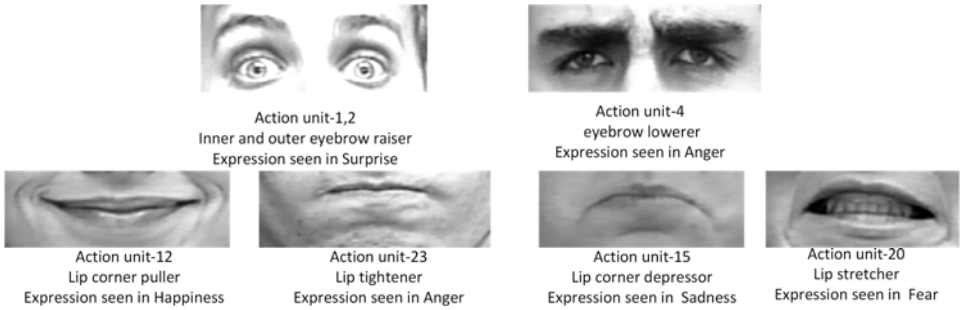


Fig. 3. Action units observed in few facial expressions.

4. DESIGN OF THE PROPOSED FER MODEL USING TRANSFER LEARNING AND LOGIC-MAX

ImageNet [22] is one of the knowledge transfer projects which provides huge datasets that are useful for training models. Powerful models like Inception, VGG-16, and Resnet are trained on ImageNet data, which consists of thousands of image categories. As the models are pre-trained with a huge database, they have a good ability to extract the features like edges, corners, and different shapes. It is wise to implement these models on our problem statement as it saves a lot of computation and time.

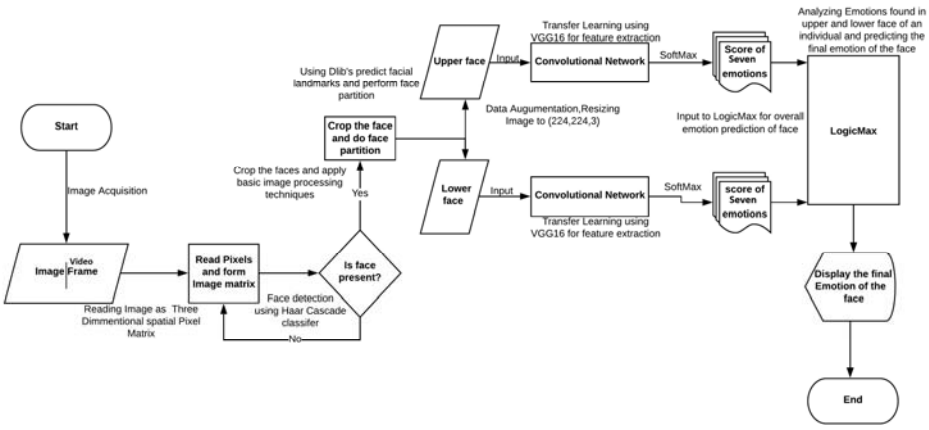


Fig. 4. Flow chart of the model process.

The proposed model, as shown in Fig. 4 has two CNN architectures that use VGG 16

architecture [9] as its base network through transfer learning. The pre-trained network gives better feature extraction and also saves much time when compared to training a whole new model. The pre-trained weights of the VGG16 network are loaded into their corresponding convolutional filters. There are many advanced pre-trained models like Resnet 101, Inception V2, and Inception V3, but VGG 16 is selected in this paper as it has a good trade-off between loading time vs. feature extraction [23]. The proposed work considers dual CNN architecture so, the VGG 16 has been chosen for its simplicity and fastness. The VGG16 base-network weights are disabled from training since it has good pre-trained weights for feature extraction.

To the base network, we have added a flatten layer, a dense layer of 256 neurons, and a dropout layer. It is followed by another dense layer with 128 neurons, a dropout layer, and a softmax layer of seven output classes, refer to the model design in Fig. 5. The training of the network involves only in updating the weights of the added layers to the base network. This same model is used twice to determine the emotion class on the upper face and lower face. We have implemented this model using the Keras framework. In the Fig. 2, representation of the model with the VGG16 as a base network and other added layers is shown.

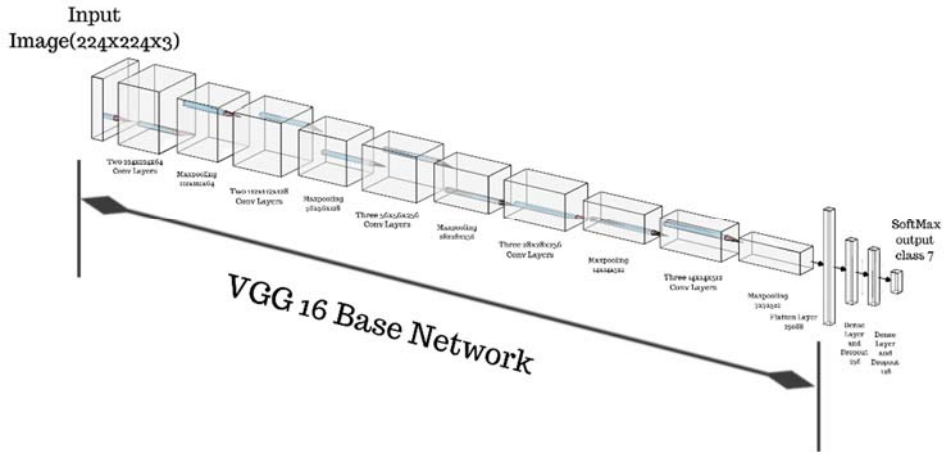


Fig. 5. Structure of single convolutional neural network used in the model.

4.1 Partition of Face

The paper considers lower and upper face parts for emotion prediction. The research in the study of facial emotion analysis in humans has revealed that the eye and mouth movements alone play an important role in the display of micro expressions [24]. The upper half of the face considers eyes as an essential indicator of emotion analysis and the lower half considers mouth as an important landmark to understand the facial expression. The face can also be divided into 3 or 4 parts but the increase in the division increases the number of CNN models which can complicate the practical design of the FER model. Localization of faces on image or video frame is done using Haar Cascade Classifier. The extracted face is to be partitioned into the lower and upper parts. The partition of the face

into more than two parts can make the algorithm complex and highly computational. We have used the Dlib library [25] and Open CV tools [26] to partition the face. Dlib's facial landmark detector is a helpful tool to identify important facial landmarks like eyes, nose, eyebrows, and jawline. Facial landmark extraction using Haar cascades is also possible but, the training to detect landmarks requires huge training of the classifier with positive and negative images to produce an accurate cascade classifier for landmarks. Kazemi and Sullivan [27] have implemented a facial landmark detector using an ensemble of regression trees. The algorithm can produce 68 coordinates of important facial landmarks. The coordinates of point 33 (Fig. 6 b) lie on the exact center of a face. We have successfully implemented a program using Opencv tools to partition a face using the dlib landmark information. The pixels that lie above point 33 belong to the upper part of the face, which includes eyes, nose, and eyebrows pixels, and the pixels that lie below point 33 belong to the lower part of the face, which includes mouth as an important landmark. The two face parts are given as inputs to their corresponding CNN models for determining the emotion class, refer to Fig. 2.

(a) A sample image taken from CK+ dataset. (b) Partition using dlib's landmark predictor.

Fig. 6. Partition of the face.

4.2 Data Augmentation

The extracted face parts are data augmented with unique parameters. Data Augmentation is a technique used to create artificial images from the dataset by transforming the image geometrics and adding random noise. The important image transformations done in data augmentation are Rescaling, Rotation, Shear, Zooming, Width Shifting, Height Shifting, Horizontal flipping, and Vertical Shifting. The combination of different parameters is to be carefully chosen to generate a good synthetic dataset. Data Augmentation is useful to eliminate the overfitting problem [28]. Overfitting in machine learning occurs when the model tries to memorize the patterns instead of learning to detect complex patterns in the training data. Detection of facial expressions should be robust even in case the image is tilted, mirrored, or zoomed. Data augmentation should be carefully performed since it can also lead to serious underfitting problems. Generally, the training and validation error helps in analyzing overfitting and underfitting problems in deep neural networks. If the model has good training accuracy but has very less validation accuracy, then the model is overfitting to the data. If the training accuracy is very less than that of validation accuracy, then the model is undergoing underfitting. It is seen that the width shifting of the training set during data augmentation does decrease the accuracy of the model since the important

landmarks get affected by high width shifting so, width shifting is not performed during the data augmentation. Only the training set is data augmented; the validation dataset is not data augmented but only rescaled. Table 3. shows the magnitude of variations of each operation during the process of data augmentation. Various combinations of values are applied, and the data shown in the Table 3. gave us the best results, and overfitting problems are avoided through the proper data augmentation process.

Table 3. Data augmentation on training set.

Operation	Scaling Factor
Rescale	1/255
Rotation Range	0-30
Shear Range	0-0.15
Zoom Range	0-0.15
Height Shift Range	0-0.2
Width Shift Range	No
Horizontal Flip	Yes
Fill Mode	Nearest

4.3 Training and Validating the Model

We have used Google's Collabalortoy GPU to train our model. Google's Colab provides GPU Nvidia 1×Tesla K80, having 2496 CUDA cores and CPU Xeon Processor of the frequency of 2.3 GHz. Input images are resized to (224×224) as the VGG16 model is trained for (224×224) sized images. The RMSprop optimizer is used in training the model. The loss function categorical cross-entropy is used as an error function for training the weights of the neural layers. The cross-entropy loss function is a widely used loss function in classification problems for deep neural networks [29]. For each batch input of images, the softmax layer produces the predicted outputs which contain CNN scores of all emotion classes. The softmax layer is a function that transforms arbitrary random values into a proper ordered probability distribution. SoftMax layer function gives output ranging between (0, 1). The total number of classes present in the CNN model is 7. Let us consider t_i and y_i be the target and the softmax score of ith class of a sample.

$$\text{Softmax score for each class } i = 1 \text{ to } 7 : f(y)_i = \frac{e^{y_i}}{\sum_j^{N=7} e^{y_j}} \tag{1}$$

$$\text{Categorical Cross entropy error} : -\sum_{i=1}^{N=7} t_i \log(y_i) \tag{2}$$

The sum of all outputs from the softmax layer equals one. In Multi-Class classification problems, the targets are one-hot encoded, which makes only the positive emotion class appear in the categorical loss function.

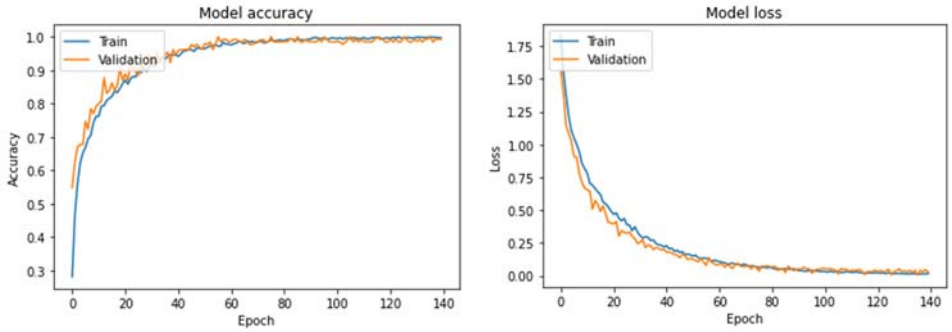


Fig. 7. Train and validation graphs of upper face CNN model on CK+ dataset during 10-fold cross-validation process.

4.4 Exploratory Data Analysis of FACS Action Units and Emotions

The LogicMax is an important layer added to the end of two CNN models. The CNN models predict the emotion class of their respective inputs (lower and upper face). The action units are shown in the Table 4 are sufficient for analyzing the emotions since the FACS considers these action units important for predicting the emotions on the face, refer to the Table 1. If both the lower and upper face CNN models predict the same emotion class, there is no perplexity involved in the decision making of the overall emotion of the face. But, if the lower and upper face CNN models predict different emotion class, there has to be a logical conclusion on the overall emotion of the face. This logical conclusion can be sought by doing exploratory data analysis on the different action units of emotions. The correlation between different emotions on the lower and upper face provides an important basis for designing the logicMax layer. The spatial distribution of important action units of emotions is shown in Fig. 8. The combination of the action units shown in the Table 4. form important basics in identifying the different emotions according to FACS (refer to the Table 1). The categorical action units are one hot encoded for correlation analysis which is shown in the Table 4.

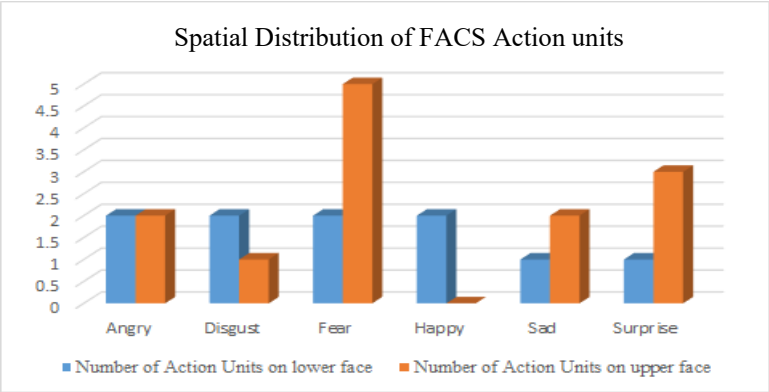
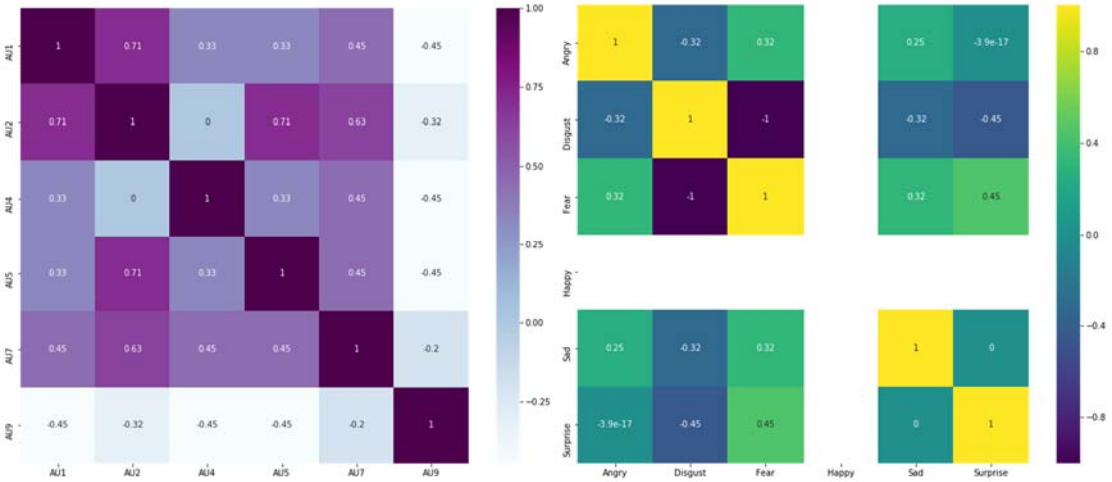


Fig. 8. Spatial distribution of action units on face.

Table 4. One hot encoding of action units for six emotions.

Action Units	Angry	Disgust	Fear	Happy	Sad	Surprise
AU1	0	0	1	0	1	1
AU2	0	0	1	0	0	1
AU4	1	0	1	0	1	0
AU5	1	0	1	0	0	1
AU6	0	0	0	1	0	0
AU7	0	0	1	0	0	0
AU9	0	1	0	0	0	0
AU12	0	0	0	1	0	0
AU15	0	1	0	0	1	0
AU16	0	1	0	0	0	0
AU20	0	0	1	0	0	0
AU23	1	0	0	0	0	0
AU26	0	0	1	0	0	1
AU27	1	0	0	0	0	0

There are no action units present on the upper face for emotion happiness. The emotion happiness has all the crucial facial action units present on the lower face (cheeks and mouth). Hence, according to FACS, if the lower CNN model predicts happiness, then it is not required to investigate the emotion class of the upper face. The correlation of emotions in lower and upper face are analyzed using the Pearson correlation matrix. It is clear that on the lower face, the emotions like Disgust and Sad have a good correlation since they share the same action unit 15 (Lip corner depressor) refer to Table 4. Therefore, there is a high probability that emotion disgust can be predicted as sad and vice versa by the lower face CNN model. In this case, it is essential to observe the upper face CNN model's prediction. It is clear that on the upper face, the disgust emotion has a unique action unit 9 (Nose wrinkle), refer to Table 1. We can also observe the disgust emotion has a poor correlation with all other emotions on the upper face, as shown in the correlation matrix refer to Figs. 9 (a) and (b). So, in the case of a mismatch, if the upper face predicts a disgust emotion class, there is no need to investigate the emotion of the lower face. We have de-



(a) Pearson correlation matrix of upper face action units.

(b) Pearson correlation matrix of upper face emotions.

Fig. 9. Correlation of different action units and emotions on upper face.

signed priorities in predicting the final emotion in case there occurs a mismatch between the prediction of two CNN models, refer to Table 5. The value of the correlation coefficient lies between −1.0 and 1.0. The value of the correlation coefficient determines the power of association. If the value of the correlation coefficient lies between 0.5 and 1.0, it suggests a strong positive association. The correlation coefficient between 0 and 0.5 suggests a weak positive association. The correlation coefficient below 0 to −1.0 indicates a negative correlation.

Table 5. Priority table inside the LogicMax layer.

Lower face CNN output class	Upper face CNN output class	LogicMax output	Justification
Neural/Disgust/ Happy	Fear	Fear	Fear has strong features on upper face
Anger	Fear/Sad/ Surprise/Happy	Anger	Anger has a type-2 feature lip tightener on mouth
Fear/Sad/ Disgust/Anger	Neutral	Neutral	Lack of emotion is seen on upper face for Neutral and Happy emotions
Sad	Fear/Sad/ Surprise/Happy	Sad	Sad has many type-3 features on the upper face
Happy	Sad/Neutral/ Surprise/Fear/Anger	Happy	Happy has a type-1 feature Lip corner puller on lower face
Surprise	Sad/Neutral/ Anger/Fear/Happy	Surprise	Surprise has a type-2 feature Jaw drop on lower face
Sad/Neutral/Surprise Fear/Happy/Anger	Disgust	Disgust	Disgust has a type-1 feature Nose wrinkle on upper face

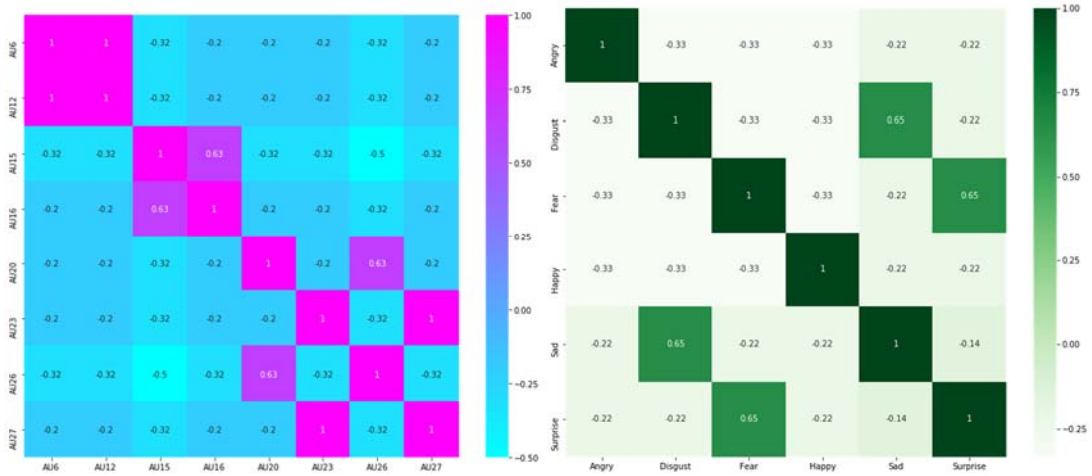
4.5 Logicmax and Priority Table

The LogicMax layer is a novel decision-making layer introduced in this paper. The logicMax, unlike softmax, can be tuned and modified according to the nature of the output class. The logicMax aims to impart human intelligence and logical thinking inside a CNN model. In the proposed model, the function of logicMax is to predict the overall emotion seen on a face by analyzing the emotions found in the lower and upper face regions. The real discriminative power of logicMax is utilized in the situation when the emotions predicted by the two CNN models mismatch. The mismatch is often seen in fa cial expression classification tasks since the correlation is high among different emotions. In these situations, the layer should choose any one of the two CNN model output. This prioritizing should be thoroughly performed by analyzing different features that appear in facial expressions. A set of rules is framed inside the priority table that decides the output class by analyzing the features. For creating a priority table, three types of features are considered. The three types of features are explained in the below points.

- **Type 1 features:** Analyze and locate the unique features present in different emotions. In the one-hot encoding Table 4, emotions like happiness and disgust have unique action units (12, 9), respectively. The action unit 12 in happiness is found on the lower face,

and action unit 9 in disgust emotion is found on the upper face. These action units are unique since they are not found in other emotions. These features are given the highest priority in the LogicMax. When a mismatch of emotion class between the two CNN models at the softmax layer occurs, these unique features are examined primarily in the input. These features can be termed as "Type 1 features". The correlation heat map charts shown in Fig. 10 (b), 9 (b), and one hot encoding Table 4 provides important information about the Type 1 features.

- **Type 2 features:** The features which have a poor correlation with all other features are the next vital patterns that need to be examined in the LogicMax layer. If the unique features (Type 1) are not available in the input, these types of features are explored in the input. These features can be easily discriminated against as they have a weak correlation with other standard features. The anger class has an action unit 23 present on the lower face, which has a weak correlation with other action units, refer to Fig. 10 (b). The detection of anger class in the lower face suggests there is a higher probability that the subject displays anger emotion. These features can be termed as "Type 2 features".



(a) Pearson correlation matrix of lower face action units.

(b) Pearson correlation matrix of lower face emotions.

Fig. 10. Correlation of different action units and emotions on lower face.

- **Type 3 features:** The next type of features are not so important as the Types 1 and 2 features since they are trivially found among different emotions. These features are given less importance in the priority table as they are not unique and appear in two or more emotions. For example, action units 4 and 5 can be observed in emotions like anger, fear, surprise, and sadness. These features create perplexity to the classifier as they are found in different classes and have a chance to increase the correlation among different emotion classes. These features can be termed as "Type 3 features".

The basics of Types 1, 2, 3 help in the construction of the priority table. The priority table contains a set of conditional statements for picking the most appropriate emotion of the entire face by analyzing the emotions found on the lower and upper face regions. The

priority table has significantly boosted the accuracy of the model during the 10-fold cross-validation process. The detailed explanation of the priority table is written in the form of a pseudo code. If the lower CNN and upper CNN output mismatch and do not possess the combinations shown in the Table 5, then the upper face CNN model output is considered as the final output class for the entire face.

5. DEALING WITH OCCLUSION

Face occlusion occurs when extraneous objects block the person's face causing problems in facial expression recognition. Typical examples of face occlusion occur in cases like a face covered with a scarf, a subject wearing glasses, subjects' beard, a subject wearing a cap and hat, *etc.*. The proposed model can deal with occlusion problems under few constraints. The following points are to be noted in face occlusion cases.

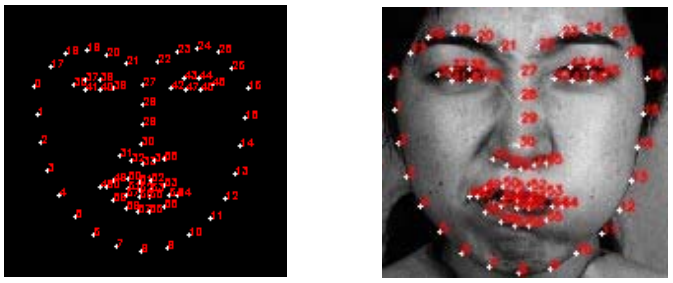
- Facial expression recognition in occluded faces is prone to misclassification. The occlusion of important facial landmarks can hinder the performance of FER models since they provide crucial information about the action units. The expressions like happiness only have essential information only in the lower face since there are no action units at the upper face, refer to the Fig. 9 (b), 10 (b), and Table 4.
- The proposed work considers dlib's landmark detector, which predicts 68 important landmarks of a face. The 68 points of the landmark of the face are shown in Fig. 11. The spatial information of the landmarks discussed are explained in the Table 6.

Algorithm 1: LogicMax

```
procedure PRIORITY TABLE
    l ← predicted emotion of lower CNN model
    u ← predicted emotion of upper CNN model
    f ← LogicMax's emotion prediction on overall face
    an ← Anger
    di ← Disgust
    fe ← Fear
    ha ← Happy
    ne ← Neutral
    sa ← Sad
    su ← Surprise
match:
    if l == u then
        u ← f
        Display: There is no mismatch
mismatch:
    if l = u then
        con1 ← (u == fe)&(l ! = sa)&(l ! = an)&(l ! = su).

        if (u == fe)&(l ! = sa)&(l ! = an)&(l ! = su) then
            fe ← f.
```

22: $con2 \leftarrow (l == an)\&(u\ != di)\&(u\ != ne)$.
23:
24: **if** $(l == an)\&(u\ != di)\&(u\ != ne)$ **then**
25: $an \leftarrow f$.
26: $con3 \leftarrow (u == ne)\&(l\ != ha)\&(l\ != su)$.
27:
28: **if** $(u == ne)\&(l\ != ha)\&(l\ != su)$ **then**
29: $ne \leftarrow f$.
30: $con4 \leftarrow (l == sa)\&(u\ != di)\&(u\ != \underline{ne})$.
31:
32: **if** $(l == sa)\&(u\ != di)\&(u\ != ne)$ **then**
33: $sa \leftarrow f$.
34: $con5 \leftarrow (f == ha)\&(u\ != fe)\&(u\ != di)$.
35:
36: **if** $(l == ha)\&(u\ != fe)\&(u\ != di)$ **then**
37: $ha \leftarrow f$.
38: $con6 \leftarrow (l == su)\&(u\ != di)$.
39:
40: **if** $(l == su)\&(u\ != di)$ **then**
41: $su \leftarrow f$.
42: $con7 \leftarrow (u == di)$.
43:
44: **if** $(u == di)$ **then**
45: $di \leftarrow f$.
46:
47: **if** $(con1|\ con2|con3|con4|con5|con6|con7) ==$ (False) **then**
48: $u \leftarrow f$.
49: **Display:** Mismatch found and LogicMax has executed successfully
50: **OUTPUT:** Print the final output: f



(a) 68 landmark points (b) 68 landmark points on an image

Fig. 11. Information on landmarks.

- Since it is difficult to retrieve the entire landmark information, the available landmarks (refer to Table 6) are tried to retrieve from the occluded image. If the retrieved information contains the upper or lower face facial region, then the region of interest points are selected and given as input to the respective CNN model for emotion prediction.

Table 6. Spatial information about the landmark points using dlib's 68 point landmark predictor.

Facial regions	Landmark points	Spatial Information
mouth	points 48 to 68	Lower
right eyebrow	points 17 to 22	Upper
left eyebrow	points 22 to 27	Upper
right eye	points 36 to 42	Upper
left eye	points 42 to 48	Upper
nose	points 27 to 35	Upper

- The accuracy rates of the lower face and upper face CNN models analyzed during the cross-validation process are greater than 90%. The upper face CNN model proposed in the paper requires eyebrows, eyes, and nose to predict the emotion. The lower face CNN model requires the mouth region to predict the emotion. In this way, the proposed work can predict the emotion of an occluded face to a possible extent.

6. EXPERIMENTAL RESULTS AND DISCUSSION

In this section, we discuss the facial expression datasets, method of testing the model, results, performance comparison of our model with other significant FER models, and different metrics for evaluation. We then provide filter visualization of the CNN models in Figs. 20 (a) and (b).

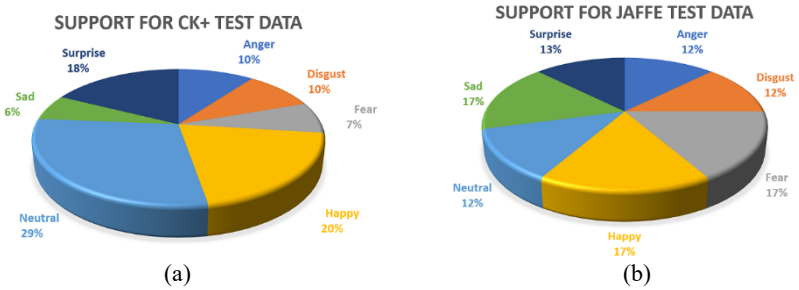
6.1 Databases

We have used the two most popular facial expression databases extended Cohn-Kanade database (CK+) and the Japanese Female Facial Expression (JAFFE) database, for testing the model performance.

1. CK+ database [5]: The extended Cohn-Kanade, widely known as CK+, is a facial expression dataset for classification of action units and facial emotion recognition. The dataset has posed as well as non-posed expressions. The Extended CohnKanade (CK+) dataset consists of 593 sequences across 123 different subjects. Considering the most appropriate method, most of the papers have taken the last three or five frames of the sequence and used them for image-based facial expression recognition. Each sequence in the database contains frames varying from 10 to 60, and in every sequence, the frames are captured such a way that there is a shift in expression from a neutral to the peak intensity of specific emotion. Among the given sequences, only 327 sequences with 118 subjects have the expression labels of anger, contempt, disgust, fear, happiness, sadness, and surprise based on the Facial Action Coding System (FACS). In this paper, we have considered the last three to four frames from each labeled sequence for classification. The seven labels taken in this experiment for expression classification are anger, disgust, fear, happiness, neutral, sadness, and surprise. A total of 1479 images are derived from the labeled sequences. The process of extracting two parts of a face is achieved using the dlib library. The images are split into two halves to get upper and lower face using the dlib library. For the lower face

and upper CNN models, 1331 images are used for training the model and 148 images for testing the model using the 10fold cross-validation process. The emotions predicted by the two CNN models are given as input to the logicMax layer. If there exists a mismatch in the expression classification between the two CNN models then, the LogicMax predicts the final emotion by applying the rules set in the priority table as discussed in the LogicMax and priority table in Section 4.5.

2. JAFFE database [11]: Japanese Female Facial Expression has a total of 213 samples, which are posed expressions taken from ten Japanese female subjects. Each subject in the dataset has nearly three to four images of six basic expressions (anger, disgust, fear, happiness, sadness, and surprise) and 1 image of neutral expression. This dataset, unlike CK+, has few images for each expression. Data Augmentation plays an important role in this dataset as it could help to extend the number of training samples. We have taken the entire 213 images in the dataset for training and testing the CNN models. The upper face CNN and the lower face CNN models are trained with 192 images and tested with 21 images in each fold during the 10-cross validation process.



(a) (b)

Fig. 12. Information about the samples used in test data for different classes.

Table 7. Information about the *k*-fold validation process on CK+ and JAFFE datasets.

Dataset	Total Samples	Training Samples	Testing Samples	Method of Testing	Data Selection
CK+	1479	1331	148	10-fold cross validation	last three to four samples of a sequence (includes peak expression frame)
JAFFE	213	192	21	10-fold cross validation	All the samples from the dataset are taken for training and testing

6.2 Classification Metrics

The important evaluation metrics of the FER model discussed in this paper are Accuracy, Precision, Recall, and F1-score. Let TP represents True Positives, FP represents False Positives, FN represents False Negatives, and FP represents False Positives.

1. **Accuracy:** Accuracy (*Acc*) is useful in evaluating model performance. However, when there exists a class imbalance problem, it is necessary to consider other im-

portant metrics like precision and recall.

$$Acc = \frac{TP + TN}{TP + FN + TN + FP} \tag{3}$$

2. **Precision:** The precision (P) highlights the ability of the model to pick the desired class. P depends on TP and FP. False Positives are the number of predictions the model misclassifies as positive when the true label is negative.

$$P = \frac{TP}{TP + FP} \tag{4}$$

3. **Recall:** Recall (R) is the other classification metric that conveys the ability of the model to predict all the classes of interest in a dataset. R depends on TP and FN. FN is the number of predictions the model misclassifies as negative when the true label is positive.

$$R = \frac{TP}{TP + FN} \tag{5}$$

4. ***F*1 Score:** It is necessary to maintain good precision and recall for any model. The goal of a good classifier is to pick the correct class without any mistake (precision) and, at the same time, pick as many as correct classes (recall). A good trade-off is to be maintained between precision and recall. $F1$ score provides a decent blend of two metrics recall, and precision. $F1$ score is the harmonic mean of recall and precision.

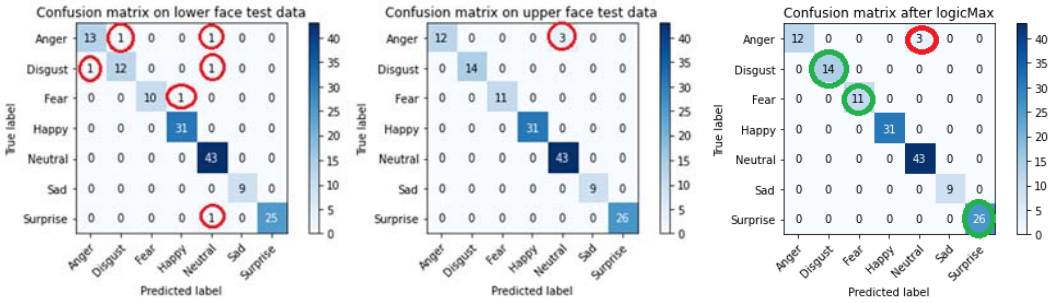


Fig. 13. Confusion matrices on third fold CK+ test data during 10-fold cross-validation.

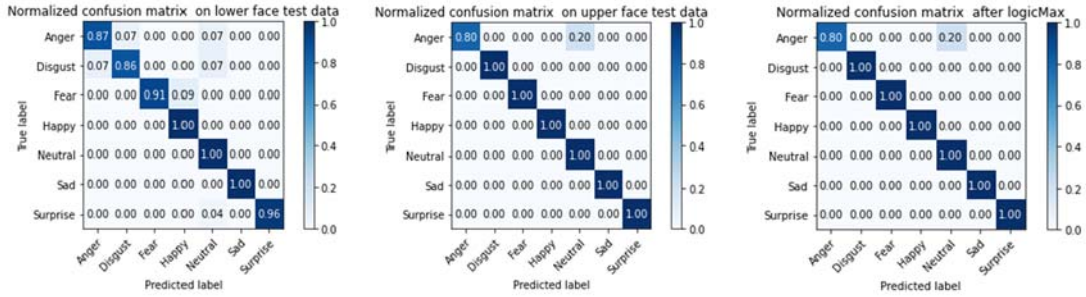


Fig. 14. Normalized confusion matrices on third fold CK+ test data during 10-fold cross-validation.

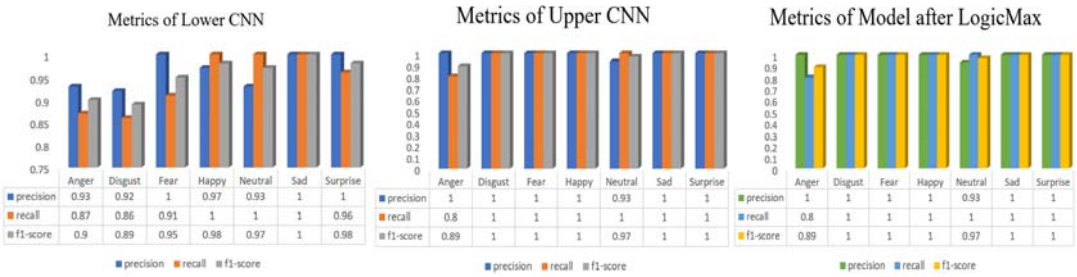


Fig. 15. Classification metrics of the model on CK+ test data (third fold cross-validation).

$$F1\ Score = 2 * \frac{P * R}{P + R} \tag{6}$$

6.3 *k*-fold Validation Process: Testing the Model Performance

Testing in machine learning and deep learning is a fundamental process in evaluating the performance of the model. The popular validation techniques mentioned in the literature are hold out method, *k*-fold cross-validation and leave one out cross-validation. In this paper, we have used the *k*-fold cross-validation procedure since it is a widely used evaluation technique in various state of the art methods.

In the *k*-fold cross-validation process, the total data is randomly partitioned into *k* equal-sized parts as shown in Fig. 16. In these *k* parts, one part is retained as the validation data for testing, and the other $(k - 1)$ parts are used for training the model. The cross-validation process is then repeated *k* times, with each of the *k* folds used exactly once as the validation data. This process results in *k* individual results, and the scores are then averaged to produce a single estimation. Each sample is used for validation exactly once, which reduces the bias on the data. The value of *k* is arbitrary. The 10-fold cross-validation is commonly used in many FER models for the evaluation process. The upper and lower face parts are partitioned from the selected data samples and given as input to the respective CNN models. In this paper, we have used the scikit learn's [30] *k*-Folds cross-validator to split the dataset into *k* ($k = 10$) consecutive folds with a shuffle. The 10 folds are created such that each fold has nearly 10% of the total data samples for testing. Table 7 provides information about the number of training and testing data samples used in upper and lower face CNN models. The training samples are shown in the Table 7. are used to train the CNN models, and the testing is done using the two CNN models and logicMax layer. There is no need for logicMax during the training process since the use of logicMax is required only during the testing phase.

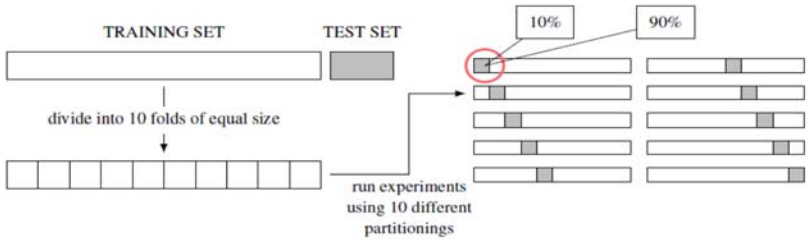


Fig. 16. *k* fold validation process, ($k = 10$).

6.4 Evaluation of Model Performance and Comparison with Other Models Under Similar Test Conditions

In this section, we have shown the results of different metrics to evaluate the performance of the two CNN models. We have performed a 10-fold cross-validation process for evaluating the accuracy scores on both datasets. The confusion matrices of CNN models and the effect of logicMax layer are discussed for both datasets. The important performance metrics accuracy, precision, recall, and F1-Score are shown for CNN models during cross-validation, refer to Figs. 15 and 19. The importance of the LogicMax can be understood from the confusion matrices shown in the Figs. 13 and 17. The differences in output during emotion classification between the lower face CNN and the upper face CNN model are corrected by the LogicMax layer, as seen in confusion matrices are shown in the Figs. 13 and 17. A typical example of the advantage of Logicmax is seen in the confusion matrices, refer to Fig. 19. Few samples under disgust are misclassified in the lower CNN model, but due to the unique action unit of disgust on the upper face, the sample gets correctly classified in the upper CNN model and using the logicMax layer the correct class is selected by the model. The normalized confusion matrices are shown in the Figs. 14 and 18. for understanding the model. The logicMax on analyzing the outputs from the CNN models predict the correct class. Emotions like disgust and Sadness which are usually hard to recognize have achieved accuracy using the proposed algorithm. Emotions of happiness, surprise, and disgust have scored good accuracy in both CK+ and JAFFE datasets. It is important to perform comparison with other models under similar test conditions. The

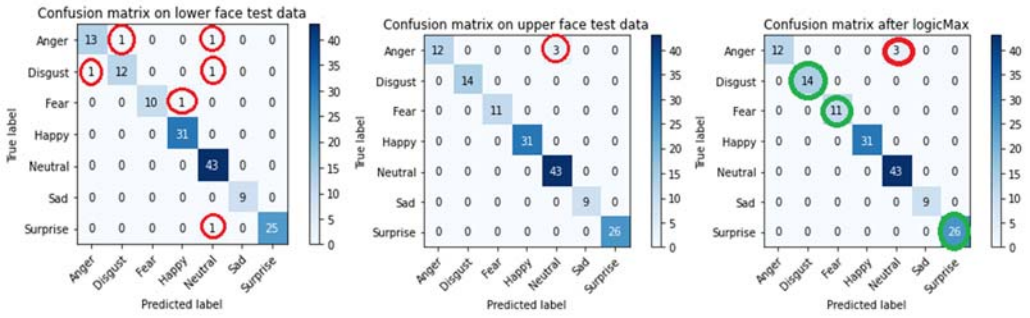


Fig. 17. Confusion matrices on eighth fold JAFFE test data during 10-fold cross-validation.

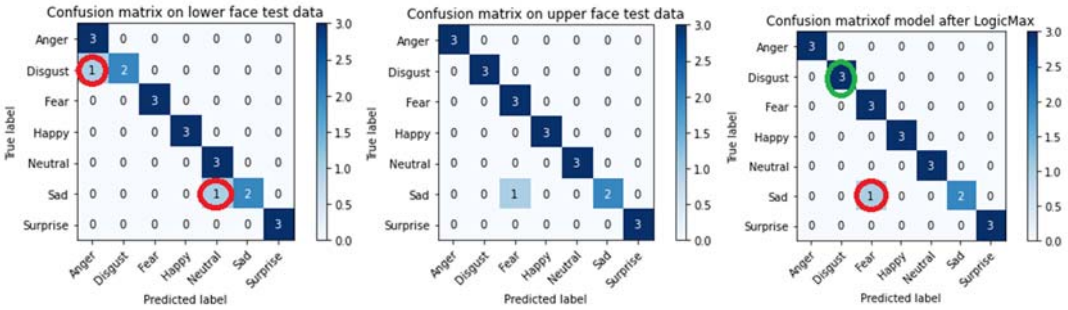


Fig. 18. Normalized confusion matrices on eighth fold JAFFE test data during 10-fold cross-validation.

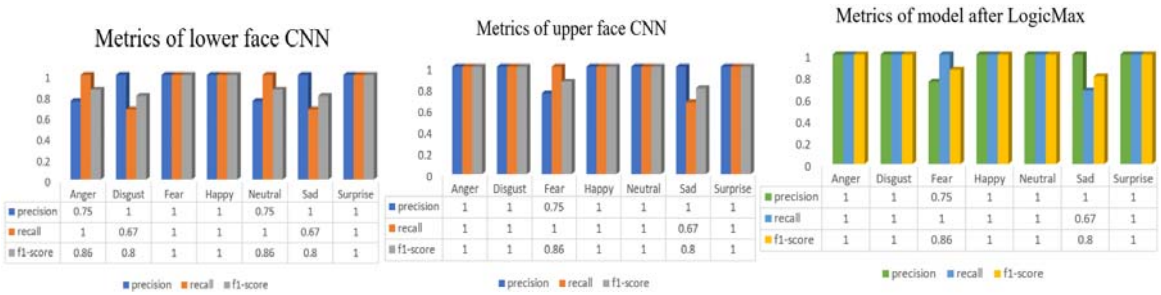


Fig. 19. Classification metrics of the model on JAFFE test data (on eighth-fold test data during 10-fold cross-validation).

important test conditions that have to be maintained are number of samples taken from the dataset for training and testing, number of classes (emotions) that the model is able to classify, the method of validating the test data and the iterations used for validating the data. The overall accuracy of the 10-fold cross-validation process on the CK+ and JAFFE datasets are 98.62% and 94.86% respectively. The proposed model is compared with another state of the art models which have used similar test conditions, refer to the Tables 8 and 9.

Table 8. Comparison of FER accuracy and other parameters on different models for CK+ dataset.

	Authors	Method	Test Procedure	Data Selection	Number of Classes	Performance (%)
CK+	Zhao *et al*. [31]	The Peak-Piloted Deep Network (PPDN), 2016	10-fold cross validation	Last three frames of each sequence (near to peak expression)	6	97.3%
	Siyue Xie and Haifeng Hu [32]	Facial expression recognition with FRR-CNN, 2017	10-fold cross validation	last three frames of each sequence (near to peak expression)	6	92.06%
	Jung *et al.* [33]	Joint Fine-Tunning in Deep Neural Networks for Facial Expression Recognition 2015	10-fold cross validation	Not mentioned	7	97.2%
	Sherly Alphonse and Dejey Dharma [34]	Novel directional patterns and a Generalized Supervised Dimension Reduction System (GSDRS), 2019	10-fold cross validation	last three to four frames of each sequence (near to peak expression)	7	97.71%
	Yang *et al.* [35]	Facial expression recognition by de-expression residue learning, 2018	10-fold cross validation	last three frames of each sequence (near to peak expression)	7	97.3%
	The proposed work	A Novel Dual CNN Architecture with LogicMax for Facial Expression Recognition	10-fold cross validation	last three to four frames of each sequence (near to peak expression)	7	98.62%

Table 9. Comparison of FER accuracy and other parameters on different models for JAFFE dataset.

'Dataset	Authors	Method	Testing Procedure	Data Selection	Number of Classes	Performance (%)
JAFFE	M.K.Mohd Fitri Alif et al. [36]	Fused convolutional neural network for facial expression recognition, 2018	10-fold cross validation	All the images in the dataset	7	83.72
	Caifeng Shan et al. [37]	Facial expression recognition based on Local Binary Patterns: A comprehensive study	10-fold cross validation	All the images in the dataset	7	81%
	Zhao et al. [38]	Facial Expression Recognition via Deep Learning, 2015	10-fold cross validation	All the images in the dataset	7	90.95%
	The Proposed work	A Novel Dual CNN Architecture with LogicMax for Facial Expression Recognition	10-fold cross validation	All the images in the dataset	7	94.86%

The problem faced during the training of our model on the CK+ database is the class imbalance issue. In the CK+ dataset, emotions like happiness and surprise have more number of labels when compared to the labels of disgust and sadness (refer to Fig. 12). The class imbalance issue can be partially solved by creating more samples through data augmentation. However, the additional images created during the data augmentation is useful only during the training process but not used in testing. In recent years, the class imbalance issue is solved by generating synthetic data through GANs.

6.5 Filter Visualization

It is important to visualize the filters in our CNN model. Each CNN tries to capture low-level features like edges and corner points in initial layers. In the VGG network Fig. 5, it is clear when we move deeper into the network, there is a decrease in the kernel size from the 2nd convolution layer (224×224) to the last convolution layer (14×14). The increase in the feature maps is also seen in the architecture from 64 feature maps in initial convolution layers to 512 feature maps in the last convolution layer. Since it is difficult to visualize all the filters from each layer, we have shown the first 64 convolutional filters in the first layer of our model in Fig. 20.

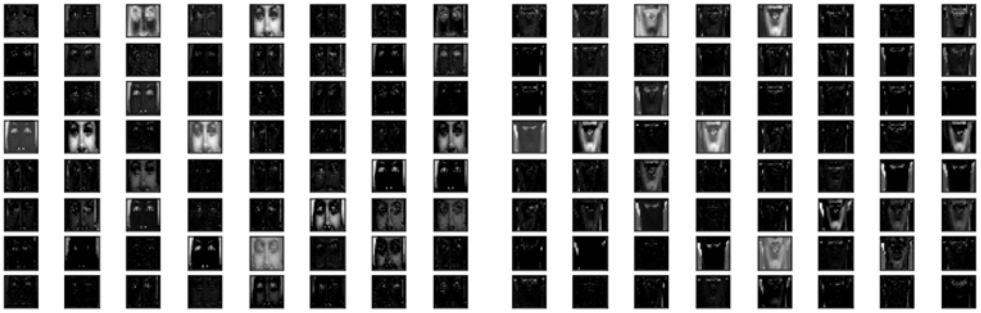


Fig. 20. Filter visualization from both CNN models.

7. CONCLUDING REMARKS

The classification of facial expressions using FACS and LogicMax has improved the accuracy rates on CK+ and JAFFE datasets. The performance of the model and other

parameters are compared with other state-of-the-art techniques, and the proposed model achieved a good accuracy score. This work improves the precision of classifying emotions like Happiness, Disgust, and surprise by implementing a dual CNN architecture. The LogicMax analyzes the predicted emotions found on the upper and lower face and decides the final class by selecting the most appropriate emotion. The proposed work can be extended by using other correlation methods on action units. In the future, the proposed model can be implemented on embedded hardware platforms.

REFERENCES

1. A. Mehrabian, "Communication without words," *Psychology Today*, Vol. 2, 1968.
2. C. Darwin and P. Prodger, *The Expression of the Emotions in Man and Animals*, Oxford University Press, USA, 1998.
3. P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion," *Journal of Personality and Social Psychology*, Vol. 17, 1971, p. 124.
4. P. C. Petrantonakis and L. J. Hadjileontiadis, "Emotion recognition from EEG using higher order crossings," *IEEE Transactions on Information Technology in Biomedicine*, Vol. 14, 2009, pp. 186-197.
5. 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 *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, 2010, pp. 94-101.
6. R. Zhi, M. Flierl, Q. Ruan, and W. B. Kleijn, "Graph-preserving sparse nonnegative matrix factorization with application to facial expression recognition," *IEEE Transactions on Systems*, *Man*, *and Cybernetics*, Part B, Vol. 41, 2010, pp. 38-52.
7. 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, 2009, pp. 803-816.
8. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1097-1105.
9. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.
10. C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1-9.
11. M. J. Lyons, S. Akamatsu, M. Kamachi, J. Gyoba, and J. Budynek, "The Japanese female facial expression (Jaffe) database," in *Proceedings of the 3rd International Conference on Automatic Face and Gesture Recognition*, 1998, pp. 14-16.
12. R. Ekman, *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression using the Facial Action Coding System*, Oxford University Press, USA, 1997.
13. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1097-1105.
14. O. M. Parkhi, A. Vedaldi, A. Zisserman, *et al.*, "Deep face recognition," in *The British Machine Vision Association*, Vol. 1, 2015, p. 6.

15. S. E. Kahou, C. Pal, X. Bouthillier, P. Froumenty, C. Gülcehre, R. Memisevic, P. Vincent, A. Courville, Y. Bengio, R. C. Ferrari *et al.*, "Combining modality specific deep neural networks for emotion recognition in video," in *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, 2013, pp. 543-550.
16. Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Proceedings of European Conference on Computer Vision*, 2016, pp. 499-515.
17. J. Cai, Z. Meng, A. S. Khan, Z. Li, J. O'Reilly, and Y. Tong, "Island loss for learning discriminative features in facial expression recognition," in *Proceedings of the 13th IEEE International Conference on Automatic Face and Gesture Recognition*, 2018, pp. 302-309.
18. F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 815-823.
19. S. E. Kahou, C. Pal, X. Bouthillier, P. Froumenty, C. Gülcehre, R. Memisevic, P. Vincent, A. Courville, Y. Bengio, R. C. Ferrari *et al.*, "Combining modality specific deep neural networks for emotion recognition in video," in *Proceedings of the 15th ACM International Conference on Multimodal Interaction*, 2013, pp. 543-550.
20. B.-K. Kim, S.-Y. Dong, J. Roh, G. Kim, and S.-Y. Lee, "Fusing aligned and nonaligned face information for automatic affect recognition in the wild: a deep learning approach," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016, pp. 48-57.
21. F. Zhang, T. Zhang, Q. Mao, and C. Xu, "Joint pose and expression modeling for facial expression recognition," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3359-3368.
22. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A largescale hierarchical image database," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248-255.
23. J. Huang, V. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama *et al.*, "Speed/accuracy trade-offs for modern convolutional object detectors," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7310-7311.
24. M. Iwasaki and Y. Noguchi, "Hiding true emotions: micro-expressions in eyes retrospectively concealed by mouth movements," *Scientific Reports*, Vol. 6, 2016, p. 22049.
25. D. E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, Vol. 10, 2009, pp. 1755-1758.
26. G. Bradski and A. Kaehler, *Learning OpenCV: Computer Vision with the OpenCV Library*, O'Reilly Media, Inc., CA, 2008.
27. V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1867-1874.
28. L. Perez and J. Wang, "The effectiveness of data augmentation in image classification using deep learning," arXiv preprint arXiv:1712.04621, 2017.
29. Z. Zhang and M. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," in *Advances in Neural Information Processing Systems*, 2018, pp. 8778-8788.

30. G. Corrado, "Scikit learn: Machine learning in python," *Journal of Machine Learning Research*, Vol. 12, 2011, pp. 2825-2830.
31. X. Zhao, X. Liang, L. Liu, T. Li, Y. Han, N. Vasconcelos, and S. Yan, "Peak-piloted deep network for facial expression recognition," in *Proceedings of European Conference on Computer Vision*, 2016, pp. 425-442.
32. S. Xie and H. Hu, "Facial expression recognition with FRR-CNN," *Electronics Letters*, Vol. 53, 2017, pp. 235-237.
33. H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, "Joint fine-tuning in deep neural networks for facial expression recognition," in *Proceedings of IEEE International Conference on Computer Vision*, 2015, pp. 2983-2991.
34. A. S. Alphonse and D. Dharma, "Novel directional patterns and a generalized supervised dimension reduction system (GSDRS) for facial emotion recognition," *Multimedia Tools and Applications*, Vol. 77, 2018, pp. 9455-9488.
35. H. Yang, U. Ciftci, and L. Yin, "Facial expression recognition by de-expression residue learning," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2168-2177.
36. M. Alif, A. Syafeeza, P. Marzuki, and A. N. Alisa, "Fused convolutional neural network for facial expression recognition," in *Proceedings of Symposium on Electrical, Mechatronics and Applied Science*, 2018, pp. 73-74.
37. 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, 2009, pp. 803-816.
38. X. Zhao, X. Shi, and S. Zhang, "Facial expression recognition via deep learning," *IETE Technical Review*, Vol. 32, 2015, pp. 347-355.

Alagesan Bhuvaneswari Ahadit received his B.Tech degree in Electronics and Communications Engineering from Kalasalingam University Tamil Nadu in 2013, M. Tech degree from National Institute of Electronics and IT Calicut in 2015. He is currently working as Research Scholar at National Institute of Technology Warangal. His research area includes the image and video processing, computer vision, machine learning applications on images, and deep learning for visual computing.

Ravi Kumar Jatoth received his B.E degree in Electronics and Communications Engineering from Osmania University, Hyderabad in 2003, M. Tech degree in Instrumentation and Control Systems from Jawaharlal Nehru Technological University Hyderabad in 2005. He received his Ph.D. from National Institute of Technology Warangal in 2014. He is currently working as Associate Professor at the National Institute of Technology Warangal. His research areas include process control system, digital signal processing, tracking, and soft computing.