Received: August 13, 2024. Revised: September 24, 2024.

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Facial Emotion Recognition of Online Learners Using a Hybrid Deep Learning Model

Evangeline D[1,2]* **Parkavi A[2,3]**

[1]*Department of Information Science and Engineering, Ramaiah Institute of Technology, Bangalore, India*
[2]*Department of Computer Science and Engineering, Ramaiah Institute of Technology, Bangalore (Affiliated to Visvesveraya Technological University, Belagavi, Karnataka, India)*
[3]*Department of Computer Science and Engineering, Ramaiah Institute of Technology, Bangalore, India*
* Corresponding author's Email: evangeline271088@gmail.com

Abstract: Human emotions are broadly categorized into seven types namely, sorrow, happiness, fear, anger, disgust, surprise, and neutral. Emotion detection is essential in various applications in the domains of education, gaming, advertising, application development, mental health diagnosis and treatment, surveillance, human- computer interaction, criminals' statement verification, etc., In literature, there is significant work on facial expressions to detect emotions in educational settings. Detection of emotions are carried out through soft biometrics and Convolution Neural Networks. This is a widely researched topic in the recent past. In this work, a thorough analysis of emotion recognition on basis of facial expressions is carried out. The work involves analysing effective emotion detection from various parts of face of participants involved in online classes. A new hybrid neural network model is derived by improvising InceptionV3 architecture in this work. On Online Learning Spontaneous Facial Expression Database (OLSFED), the proposed architecture produces very promising results exceeding the performance of InceptionV3. While comparing the proposed hybrid Inception V3 model with the other neural network models in existing literature, it was found that the proposed model also outperforms ResNet50, MobileNetV1 and EfficientNetB0 (with Swish activation function). On comparing the proposed model with state-of-the-art methodologies, it was observed that the proposed hybrid InceptionV3 architecture outperforms MobileNetV2 and CNN10 architectures. Also, from the various experiments done on OLSFED dataset, it was inferred that emotion recognition on periocular and mouth region together provides a very close performance to emotion recognition on the whole face.

Keywords: Emotion recognition, Student engagement, CNN, Soft biometrics, Inception V3.

1. Introduction

Human emotions are complex in nature as it is considered as response of human brain to the actions in the environment. It has laid the foundation of "Affective Computing [1] which is a promising domain that could benefit the people in various domains. Affective Computing is considered the confluence of three domains – Computer Science, Cognitive Science and Psychology. This Affective Computing is based on three processes - emotion recognition, emotion elicitation and emotional behavior generation of which emotion recognition is a critical phase. Human emotions, in general can be categorized on basis of the following modalities [2]:

- Behavioral / Expressive Modality
- Somatic / Physiological Modality
- Cognitive / Interpretive Modality
- Experiential / Subjective Modality

Behavioral or expressive modality is the widely employed modality in which emphasis is laid on facial expressions, speech and gestures. Many emotion recognition techniques employ these physical signals but few employ physiological signals like Electro-Encephalography (EEG), Electro-Cardiography (ECG), Galvanic Skin Response, etc., Based on the application, one has to

make a choice between physical and physiological signals for emotion recognition.

Emotion recognition can be made through visual or audio or behavioral cues. Such cues could be soft biometric traits. Hard biometric traits possess properties like universality, uniqueness, permanence, collectability, performance, acceptability and circumvention that makes the subject to be identified. But soft biometric traits lack uniqueness and permanence. However, they might be employed to uniquely identify individual. Some of the soft biometric traits are age, eye color, height, weight, gender, androgenic hairs, wrist, skin color, blood vessels, etc., Facial expressions which are also soft biometric traits are significant in emotion detection [3]. It has been observed that facial expressions contribute 55% in emotion analysis while speech and non-verbal communication contributes to 38% and 7% respectively [4]. Some literature focusses only on facial expressions for emotion recognition while many other use multiple modalities including speech, keyboard or mouse dynamics for improving the results. Actually, there are many action units on human face. Action units comprise of combination of multiple facial movements at various facial landmarks. Boredom, confusion, delight, surprise and frustration are some emotions that could be detected from these facial units. Eyes, mouth, nose, chin and forehead are the prime facial landmarks that could be employed for analyzing the engagement or emotions of students during online learning. In general, facial expression recognition algorithms focus on extraction of facial features like eyes, mouth, etc., followed by feature selection and classification of emotions. The various stages of Facial Expression Recognition is mentioned in Fig. 2. [5]. Feature extraction is a significant phase in Emotion Recognition. Hand-crafted or non-handcrafted features can be employed for feature extraction. While handcrafted features are those that are manually made by researches, non-handcrafted features are automatically obtained from Deep Learning approaches.

Emotion Recognition is widely employed in many applications like Psychiatric Treatment, Online / Offline Learner Emotion Detection, Emotion Recognition of kids and elderly patients, Abnormal Activity Recognition, etc., This can be done using obtrusive or unobtrusive emotion recognition methodologies. Obtrusive emotion recognition involves emotion recognition from physiological traits like Electro-Encephalogram (EEG), Electro-Dermal Activity (EDA), Galvanic Skin Response (GSR), Photo Plethysmography (PPG), Electro Cardiogram (ECG), etc., Such emotion recognition methodologies require hardware devices to measure and track the emotions. There is a great possibility that it might cause discomfort for the subjects. The subjects may alter their emotions as they are well aware their emotions are tracked as it is being monitored by devices. On the other hand, unobtrusive emotion recognition involves tracking emotions through facial expressions, gestures, eye movements, speech and keystroke dynamics. So, unobtrusive emotion recognition systems can serve as a significant method to track emotions of online learners. Considering the above-mentioned perspectives, the objectives of the paper are thus stated as follows:

- To conduct a thorough analysis of emotion recognition from various parts of face.
- To employ relevant dataset of online learners for emotion recognition.
- To study the performance of existing neural architectures and state-of-the-art methods for emotion detection from various parts of face.
- To hybridize InceptionV3 which is an existing neural network architecture for emotion classification.

The significance of the proposed work is as follows:

- Focus on various parts of face for emotion detection is not much explored in the existing literature. Experimental analysis on various architectures signifies performance improvement in facial emotion recognition with focus on certain facial parts.
- Hybridization of InceptionV3 by incorporating Fused MB Convolution Blocks promises higher accuracy for emotion classification when compared with Inception V3 and other architectures.

The rest of the paper is organized as follows: Section 2 elaborates the methodologies of existing literature. Section 3 describes the proposed methodology. Section 4 presents experimental results and analysis. Section 5 concludes the work and highlights future research directions.

2. Related works

Facial emotion recognition is best understood with the pre-requisite knowledge of Emotion models, facial / behavioral / physiological traits, etc.,

Ekman and Friesen established the fact that all emotions are universally constant but may differ culturally under certain circumstances wherein people may hide their emotions [6]. There are 43 facial muscles that can allow one to masticate (chew) or express feeling through smiling, raising / lowering

brows, etc., Action units (AUs) are slight movements of facial muscles. Combination of such action units can represent facial expressions, leading to a multi-label classification problem [7]. Facial expressions along with corresponding action units are summarized in Table 1. There are three wide groups of models namely: Discrete models, Dimensional models and Cognitive models. While discrete emotional model (also known as Categorical emotional model) classifies human expressions into six emotions namely, happiness, sadness, surprise, anger, fear, and disgust, Plutchik's emotional wheel model (also known as componential model) focuses on eight emotions namely - joy, trust, fear, surprise, sadness, anticipation, anger, and disgust. Dimensional Model focuses on emotional states being represented across various dimensions. Pleasure – Arousal – Dominance model, also known as Circumplex Model of Emotion is a widely employed dimensional model. Pleasure (Valence) represents extent of happiness from distress to ecstasy. Arousal indicates physiological and psychological activity. Dominance gives the impact of environment on people's emotions [8]. Cognitive Emotional Models focus on development of agents that can mediate interaction between humans and robots. Appraisal models and Action Tendency models are types of Cognitive Emotion model [9]. The general steps in Facial Emotion Recognition Systems are represented in Fig 1.

Facial Emotion Recognition (FER 2013) [10-13], Extended Cohn Kanade (CK+) [10-17], Japanese Female Facial Expression (JAFFE) [13, 15] ,Karolinska Directed Emotional Faces (KDEF) [13, 14] Real-world Affective Faces (RAF) Dataset [11], Chinese Academy of Science – Pose, Expression, Accessories and Lighting (CAS-PEAL) [7], Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [18], Bosphorous Databases [17, 19], etc., are widely employed datasets. Apart from these datasets, many authors have created their own datasets suitable for their study [11, 19-21]. Few datasets focus on specific application like monitoring of patients [12], student monitoring [22, 23] etc., Some datasets may have single person in an image while some datasets may have multiple faces in one image [51]. Some of the other datasets are ICSAD [23], AffectNet [24], DAiSEE [25], OLSFED [26] and MCFER [12]. The details of the datasets are tabulated in Table 2.

2.1 Facial pre-processing

Viola-Jones [27] is an effective face detection algorithm that has four steps namely selection of Haar like features, creation of integral image, running Adaboost training and creation of classifier cascades. There are three types of Haar features like edge, line and four sided features. Adaboost enables to identify the best features to provide a boosted classifier. Usually, facial regions in all images are cropped to same dimensions. Kazemi-Sullivan [28] is a significant facial landmark detection algorithm that identifies landmark facial points along the left and right jaw line, chin, left and right eyebrows, bridge and bottom of nose, left and right eye, and outer and inner lip regions. There are 68 facial landmark points. Face captured during image acquisition must be resistant to scale and translation variations. Hence, facial feature normalization is required. Let $p_{lEye}(t)$ and $p_{rEye}(t)$ indicate the left and right outermost eye landmark points at time t. The midpoint $p_{mid}(t)$ gives the midpoint between the two eyes [23].

$$p_{mid}(t) = \frac{p_{lEye}(t)+p_{rEye}(t)}{2} \quad (1)$$

$$\hat{p}_x(t) = p_x(t) - p_{mid}(t) \quad (2)$$

Here, $p_x(t)$ and $\hat{p}_x(t)$ are the original and translated coordinates of feature x at time t as given in Eq. (1) and Eq. (2). When $d_u(t)$ is the distance between $p_{lEye}(t)$ and $p_{rEye}(t)$ at time t, then the scale normalized feature point x at time t is given in Eq. (3).

$$\tilde{p}_x(t) = \frac{\hat{p}_x(t)}{d_u(t)} \quad (3)$$

The size of some datasets may not be sufficient. So, image may be subjected to horizontal or vertical mirroring. The resultant images are populated in the datasets so that the various approaches must be robust to scale and rotation [10]. ZCA whitening, rotation, zoom and shear variations, shift in height, width and channel are adopted in [29].

2.2 Video pre-processing

In [21], the first frame captured in video is considered as the key frame. The next key frame extracted from the video requires to be distinct. This decision is done based on cosine similarity which is given below in Eq. (4).

$$Cosine\ Similarity = \frac{a.b}{ab} = \frac{\sum_{i=1}^{n} a_i * b_i}{\sqrt{\sum_{i=1}^{n} a_i^2} * \sqrt{\sum_{i=1}^{n} b_i^2}} \quad (4)$$

Where, a_i and b_i indicate feature number i in feature vector a and b respectively. When the cosine

similarity is greater than 0.998, the frame is not considered as the key frame. For every frame in the video captured, sliding window and "winner take all" method had been employed to attain sequential motion detection. This technique can reduce the impact of noise interference, thereby making the system stable [23].

2.3 Occlusion

Background clutter is very common during image acquisition. Certain works were very keen in emotion recognition in classroom environment either on online or offline modes. In classrooms, students may use notebooks, smartphones, books that can contribute to background clutter [23]. Posture of the subject can be estimated using OpenPose, a human pose estimation library that can detect human body, foot, wrist, elbow and facial keypoints [23, 30]. Let $j_{lw}(t)$ and $j_{rw}(t)$ denote the center of left wrist and right wrist at time t respectively. The object is definitely positioned at the center of these coordinates as given in Eq. (5). To find out the top-left and bottom-right coordinates of the object, Eq. (6), Eq. (7) and Eq. (8) could be employed.

$$Obj_{mid}(t) = \frac{j_{lw}(t) + j_{rw}(t)}{2} \quad (5)$$

$$Obj_{topleft} = Obj_{mid}(t) - L_{shoulder}(t) \quad (6)$$

$$Obj_{bottomright} = Obj_{mid}(t) + L_{shoulder}(t) \quad (7)$$

$$L_{shoulder}(t) = |j_{lw}(t) - j_{rw}(t)| \quad (8)$$

2.4 Feature extraction

Facial cues are focused in certain works [31] where the students' concentration during learning can be estimated. Blinking, yawning, shaking and nodding are emphasized in [30, 31]. Eye Aspect Ratio (EAR), as given in Eq. (9) can be computed from the six facial landmarks on the edges of the eye.

$$EAR = \frac{\|p_{38} - p_{42}\| + \|p_{39} - p_{41}\|}{2\|p_{37} - p_{40}\|} \quad (9)$$

When the ratio is less than 0.2 for a preset period of 150 ms, then it is determined as closed eye. Otherwise, if EAR fluctuates within the time period, it is determined as blinking. The vertical distance d_{mouth} between the midpoints of the upper and lower lip (given by facial landmarks p_{52} and p_{58}) is given in Eq. (10):

$$d_{mouth} = |p_{52} - p_{58}| \quad (10)$$

When d_{mouth} ranges between 0.2 to 0.4 units, the subject is not yawning. When d_{mouth} exceeds 0.5 times the width of shoulder, it is considered yawning. The subject is considered to have nodded head once when the pitch angle threshold of head up and down movement is 0.3 and considered to have shaken the head once when roll angle threshold of head left right movement is 0.5. Fixation, saccade, eye blink and pupil diameter indicate the characteristics of eye movement to determine FCDE (Feature of coordinate difference of eye movement). While fixation is the time period during which the eyes remain relatively stationary, saccade is the time during which there is rapid eye movement from one target to another. FCDE is divided into two features $FCDE_s$ and $FCDE_f$, that denotes FCDE in saccade and fixation respectively [20]. When n number of frames are sampled during saccade, the eye movement coordinate of the first frame (x_{p1}, y_{p1}) gives the start of the eye movement trajectory and the same coordinate positions at different frames are given by (x_{vi}, y_{vi}).

$$FCDE_s = \frac{\sum_{i=1}^{n} \sqrt{\left((x_{pi} - x_{vi})^2 + (y_{pi} - y_{vi})^2\right)}}{n} \quad (11)$$

$$FCDE_f = \frac{\sum_{i=1}^{n} \sqrt{((x - x_{vi})^2 + (y - y_{vi})^2)}}{n} \quad (12)$$

During the computation of $FCDE_f$, the eye movement coordinate (x, y) does not vary. FCDE is the average of $FCDE_s$ and $FCDE_f$ which are computed as in Eq. (11) and Eq. (12). Mouse and keyboard dynamics is focused in [21] .Number of mouse clicks (pressed or released) in all the lines of log file n is given by the click number as mentioned in Eq. (13).

$$Click\ Number = \sum_{i=1}^{n} Action_i \in pressed\ or\ released \quad (13)$$

Mouse speed can be given as rate of change of distance i.e., pixels as in Eq. (14).

$$Mouse\ Speed = \frac{Distance\ (Pixels)}{Time} \quad (14)$$

The total number of keystrokes is given in Eq. (15):

$$Keystroke = \sum_{i=1}^{n} Action_i \in Key \quad (15)$$

Considering the mistakes in typing $misNum$, typing speed was given in Eq. (16).

$$Typing\ Speed = \frac{\left(\frac{Keystroke}{5}\right) - misNum}{Time} \quad (16)$$

In Intelligent Tutoring Systems, the work [21] focusses on composite engagement of learners in online learning environment. The composite engagement value was computed in Eq. (17).

$$e_1 = \left(\frac{vEL_1}{1} * 10000\right) + \left(\frac{vEL_2}{2} * 1000\right) + \left(\frac{vEL_3}{3} * 1\right) \quad (17)$$

Where, vEL_1, vEL_2 and vEL_3 indicate strong, high and medium engagement levels respectively. The work [11] focusses on determining Engagement Index EI to find out the academic emotion as given in Eq. (18).

$$EI = Emotion\ Probability(EP) * Weight\ of\ Emotion(WE) \quad (18)$$

The emotions considered here are happy, surprised, neutral, angry, fear and sad that take the weights 0.6, 0.6, 0.9, 0.25. 0.3 and 0.3 respectively. While emotion probability is given by deep CNN, emotion weight gives the learner's emotional state at that instant of time.

2.5 Multi-modal fusion

Modalities like eye movement, audio and video are given specific weights and weighted decision level fusion is carried out in [20]. The weights w_1, w_2 and w_3 are given as 0.225, 0.675 and 0.1 respectively for the modalities - eye movement, audio and video. Weighted joint fine tuning is applied in [18] wherein, image sequence from video I, speech S, and facial landmark in the image L are integrated using Eq. (19).

$$W_1 O_I + W_2 O_L + W_3 O_S \quad (19)$$

Where, W_1, W_2 and W_3 take the values 0.2, 0.2 and 0.6 respectively.

2.6 Multi-modal fusion

Though Classification of emotions is done using classifiers like SVM [11], Naïve Bayes [21], Ada SVM [17], k-NN [14], Random Forest [15], etc., Convolution Neural Networks [32] are widely used in many image processing techniques and is extensively employed for emotion recognition. CNNs have three types of layers: convolutional layers that yield activation map, pooling layers that provides summary statistic of neighborhood and fully-connected layers that maps input to output. LSTM [18], EC-CNN [20], MobileNet V1 [12], HM-RNN [27], Mini Xception [21], Efficient SwishNet [13], VGG19 [10], ResNet50 [11] and ResNet18 [7], InceptionV3 [11], etc., are some of the neural network architectures used for facial emotion recognition in various works. Literature survey is briefly tabulated in Table 3.

Table 1. Action Units, Facial Expressions and Emotions

Emotion	Action Units	Facial Expressions
Anger (hostility)	AU4+AU5+AU7+AU23	Lowered eyebrows, glared eyes, tightened narrowed lips, tightened lower eyelids,
Disgust (displeasure)	AU9+AU15+AU16	Narrowed eyebrows, wrinkled nose and curled upper lip.
Happiness (pleasure)	AU6+AU12	Corners of the mouth are pulled upward, and the large orbital muscles around the eyes are contracted
Sadness (grief)	AU1+AU4+AU15	Drooping eyelids, lowered mouth corners, downcasted eyes, and pouted lips.
Surprise	AU1+AU2+AU5+AU26	Raised eyebrows, wrinkled forehead, widely opened eyes, dropped jaw, and a opened mouth
Fear	AU1+AU2+AU4+AU5+AU7+AU20+AU26	Raised, pulled together eyebrows, tensed lower eyelids, and a lightly opened mouth

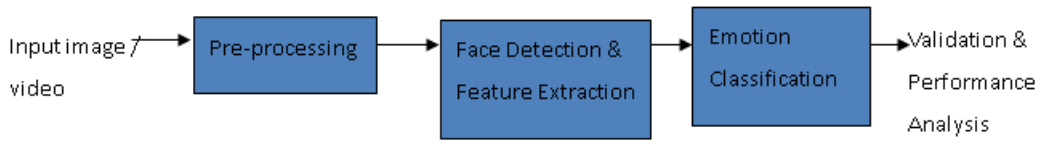

Figure. 1 Facial Emotion Recognition Systems

Table 2. Summary of various datasets

Dataset	Ethinicity	Nature	Size	Emotions
FER - 2013	Multiple ethinicities	Acted and Spontaneous	Training set - 28,709 Test set - 3,589	Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral
CK+	Africa America Asia	Acted and Spontaneous	1200	Angry, Disgust, Fear, Happy, Sad, Surprise, Contempt Neutral
JAFFE	Japan	Acted	213	Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral
KDEF	Sweden	Acted	4900	neutral, happy, angry, fear, disgust, sad, surprise
RAF DB	Multiple ethinicities	Collected from Internet	29672	12 classes of compound emotions
CAS-PEAL	China	Elicited	99594	Smile, frown, surprise, close eyes, open mouth
RAV-DESS	North America	Acted	7356 video recording	Speech - calm, happy, sad, angry, fearful, surprise, and disgust
DAiSEE	Indian	Spontaneous	9068 video snippets	Engaged, Bored, Confused, Frustrated

Table 3. Summary of Literature Review on Emotion Recognition

Work	Dataset	Features	Methodology	Merits	Demerits	Accuracy
[10]	FER-2013 CK+	Face	Variants of VGG Net	Multi-attention mechanism is employed Delivered Duty Unpaid (DDU) loss function is used	Many parameters of VGG Net is not explored	MCSA VGG Net with accuracy of 0.64, 0.91, 0.63 for Angry, Happy and Sad on FER-2013 and 0.97, 1, 0.96 for Angry, Happy and Sad on CK+
[21]	FER-2013, RAF and New dataset with 110 subjects	Face + Keystrokes + Mouse Dynamics	Mini Xception with ReLU Activation function	-	-	95.23% 90.47%, 76.19% for single, dual and multiple modalities
[11]	FER-2013 CK+, RAF-DB and Newly constructed dataset with 20 learners	Face	Inception-V3 VGG-19 ResNet-50	Pre-trained faster R-CNN model	Low volume test set	89.11 (Inception-V3) 90.14 (VGG-19) 92.32 (ResNet-50)
[12]	FER-2013 and MCFER dataset with 15 subjects	Face	MobileNet V1 with joint loss, center loss and softmax loss	The mobile app could recognise facial expressions when there are multiple faces in a single frame	Speed decreases with more number of faces	95.89%
[13]	CK+, FER-2013 JAFFE, KDEF FERG	Face	Efficient SwishNet model with SwishNet activation functions	Generalizability of the model was focussed by means of cross- corpora evaluation	Occlusion and face covered with glasses, mask, hair, etc., are not covered	100% (CK+), 64.2% (FER-2013), 95.02% (JAFFE) 85.5% (KDEF), 100% (FERG)

[14]	CK+ KDEF	Face	Web-shaped model and k-NN classifer	Simple method with good results Works even with head pose variations	Neutral expression is not recognised properly. Results of KDEF different from CK+	91% for CK+ (8 classes) and 68% for KDEF
[15]	JAFFE CK+	Face	Parabola model SVM, MLP and Random Forest	Classification accuracy is 100% for angry and happy. Robust to illumination, scale and rotation variations	Classification accuracy of negative emotions is inferior	Average classification accuracy is 96.32% on JAFFE and 96.36% on CK+
[16]	FER-2013	Face	Haar Cascade and CNN	Four categories of learner are focussed Happy / Surprise: Active Angry / Disgusted: Passive Fear / Sad: Non – Listener Neutral : Evaluative Listener	Negative emotions like disgust and angry have less samples in training set.	94.44%
[17]	CK+ MMI Bosphorous BU-3DFE BP4D Spontaneous	Face	3D head tracking algorithm based on ellipsoidal model Gabor features and AdaSVM for classification	Robust to pose variations	For Bosphorous dataset, the results are varying when compared to other algorithms	98.2 % (CK+) 97.2 % (MMI) 98.9% (Bosphorous) 93.5% (BU-3DFE) 97.2% (BP4D Spontaneous)
[7]	CK dataset	Face	Keras CNN PCA ResNet18	Female and male expression were studied separately.	Hate and neutral facial expression were not recognised easily	47.49%, 71.5% and 71.2%
[19]	CK+ & Bosphorous dataset and Own dataset	Face	RBF Neural Network	-	68 facial landmarks must be optimized	MSE is 0.01826
[18]	RAVDESS	Facial landmarks, Image sequences and audio signal	CNN and LSTM models	Syncing speech and image is accomplished Use of multiple modalities and weighted joint fine tuning has increased accuracy by 10%	-	87.11% for weighted joint fine tuning
[20]	Own dataset with 68 subjects	Audio-visual features and eye movement	FE-CNN and EC-CNN	Deep and shallow features are fused at feature-level, decision level and model-level	Boredom, confusion, happiness and interest are the only four emotions considered	Model-level fusion, decision level fusion and feature-level fusion yields 80.16%, 81.90% and 76.02% respectively

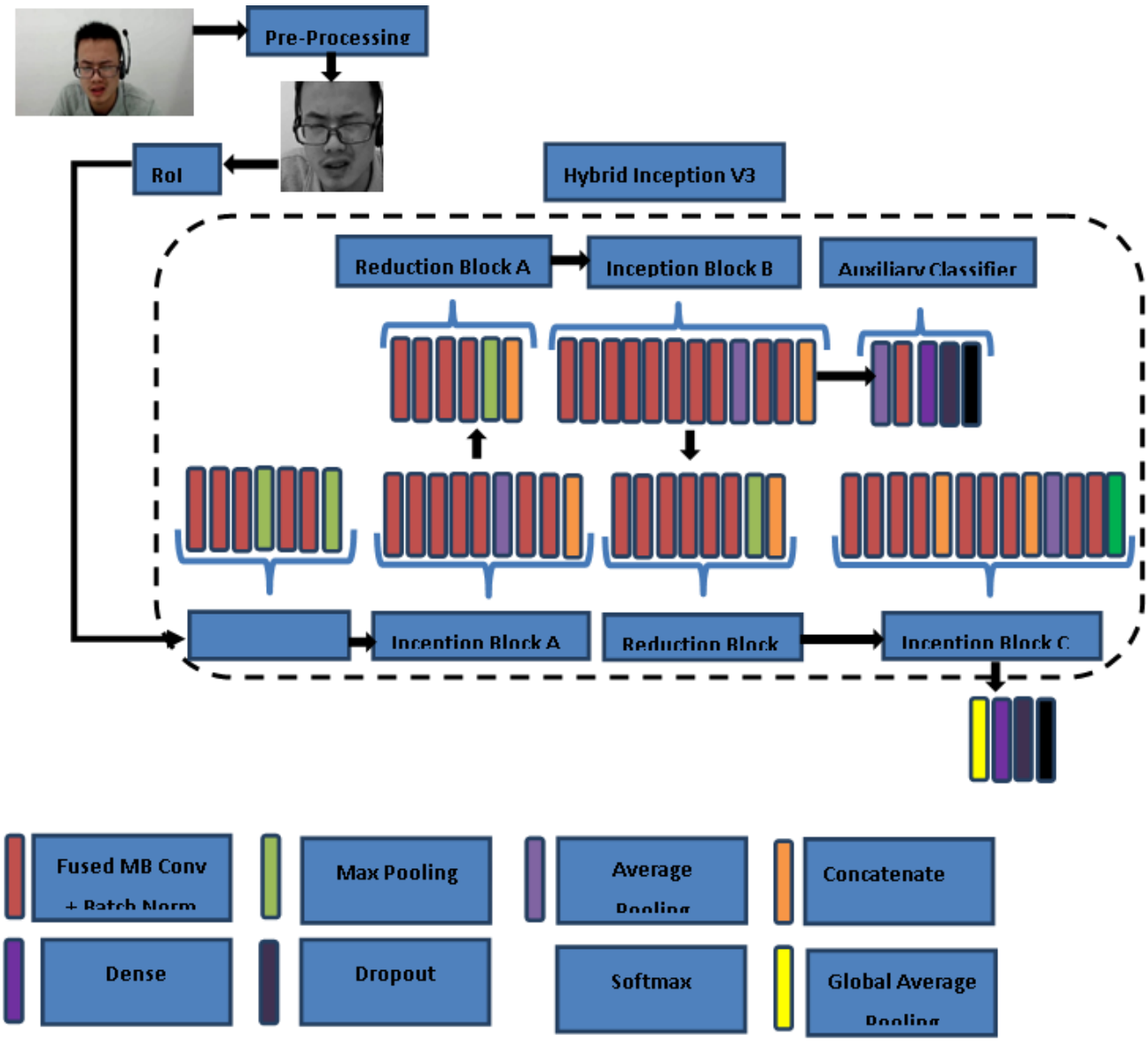

Figure. 2 Facial Emotion Recognition using Hybrid InceptionV3

From the extensive literature survey conducted, the following observations have been made. Many datasets contain images and videos of emotions of adults. Expression recognition in kids and senior citizens requires appropriate datasets because expression recognition could be different across varying age categories. Many of the datasets include images of people from a certain geographical region. Caucasian, Japanese, Chinese, Asian and American subjects' images are available in many datasets. Indian datasets are very much limited. In most of the datasets, certain negative emotions might have limited image samples and accuracy of determining those emotions is lesser compared to positive emotions. While exploring multi-modal biometrics, available datasets may be limited. Female facial expressions are richer than male facial expressions. Datasets may or may not contain occlusion which can lead to difficulties in pre-processing while classifying emotions in real-time. 3D datasets are very much limited. Very limited datasets are available on compound emotions. Datasets of images captured from real-time online learning environment are very much limited. Principal Component Analysis (PCA) is employed in few studies for dimensionality reduction. Action units are essential in emotion classification. Certain action units may be correlated. Only the uncorrelated features could be considered for the study. This could reduce time complexity without affecting accuracy as dimensionality reduction occurs. Facial images when revealed can cause security concerns as biometric traits could be extracted. Also, privacy becomes a questionable factor when they are processed at some other

locations. Images and videos acquired for emotion recognition could be processed at the user's end for emotion recognition. Use of Federated Learning is much recommended. There are several factors that may influence the identification of emotions. People belonging to Eastern ethnicities especially Japan may tend to infer emotions from cues of eye compared to their western counterparts. Social sensitivity and social anxiety may also pose challenges. The manner in which the emotions are misinterpreted needs to be studied. And emotion recognition systems may be designed by assigning varied weights for different facial features for improved emotion identification.

3. Proposed methodology

3.1 Datasets

OLSFED is a publicly available dataset that contains 30184 facial images of 82 participants in online classes [38]. The images were of size 1280*720 with JPEG compression standards. The expressions were captured with various illumination and occlusions. The emotions considered in this dataset are confusion, distraction, enjoyment, neutral and fatigue that are academic specific.

3.2 Pre-processing

Images of the dataset are RGB in nature and grayscale conversion is done. Also, facial regions are extracted using Haar Cascade Classifier.

3.3 Extraction of various facial parts

In the proposed work, OLSFED dataset is employed. The facial regions were split into several parts and performance of neural architectures like InceptionV3 and VGG-16 were employed. Facial regions were primarily experimented with left, right, periocular, mouth, non-periocular and non-mouth regions apart from the entire face. According to human anthropometry, the eyes reside in the top region of face. Periocular region is the feature-rich region surrounding eyes including eyelids, eyelashes, eyebrows, tear duct, etc. The periocular region can be extracted using Eq. (20) and Eq. (21) as in [33].

$$width_{peri} = 0.67 * \frac{height_{face}}{2} \tag{20}$$

$$height_{peri} = 2 * distance_{(eyebrow,eyecenter)}$$
$$height_{peri} = 2 * \left(0.21 + \frac{0.07}{2}\right) * \frac{width_{face}}{2} \tag{21}$$

The width of the periocular region of each eye is the width of the corresponding eyebrow. The height of periocular region is twice the distance between eyebrow and center of eye [33]. Assuming a as the half of the height of face and b as the half of width of face, Region of Interest (RoI) Extraction for periocular region is demonstrated in Fi 3.

Table 4. List of Notations

Notation	Description
$p_{lEye}(t)$ and $p_{rEye}(t)$	The left and right outermost eye landmark points at time t
$p_{mid}(t)$	The midpoint between the two eyes at time t
$\tilde{p}_x(t)$	scale normalized feature point x at time t
$Cosine\ Similarity$	cosine similarity
a_i and b_i	feature number i in feature vector a and b respectively
$j_{lw}(t)$ and $j_{rw}(t)$	Center of left wrist and right wrist at time t respectively.
$Obj_{mid}(t)$	Midpoint of object at time t
$Obj_{topleft}$	Top left of object at time t
$Obj_{bottomright}$	Bottom right of object at time t
$L_{shoulder}\ (t)$	Length of shoulder at time t
EAR	Eye Aspect Ratio
d_{mouth}	Vertical distance d_{mouth} between midpoints of upper and lower lip
FCDE	Feature of coordinate difference of eye movement
$FCDE_s$ and $FCDE_f$	Feature of coordinate difference of eye movement for fixation and saccade
$Click\ Number$	Number of mouse clicks
$Mouse\ Speed$	Speed of mouse
$Keystroke$	Number of keystrokes
$Typing\ Speed$	Typing Speed
vEL_1, vEL_2 and vEL_3	strong, high and medium engagement levels
EI	Engagement Index
EP	$Emotion\ Probability$
WE	$Weight\ of\ Emotion$
$width_{peri}$	Width of periocular region
$height_{face}$	Height of face
$height_{peri}$	Height of periocular region
$width_{face}$	Width of face
$left_{face}, top_{face},$	Left and top coordinate of face
$left_{mouth}$	Left of mouth region
$width_{mouth}$	Width of mouth region
top_{mouth}	Top coordinate of mouth region
$height_{mouth}$	Height of the bounding box encompassing mouth region

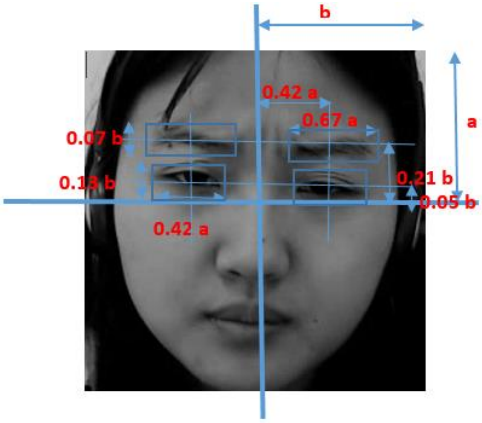

Figure. 3 Periocular Region of Interest Extraction

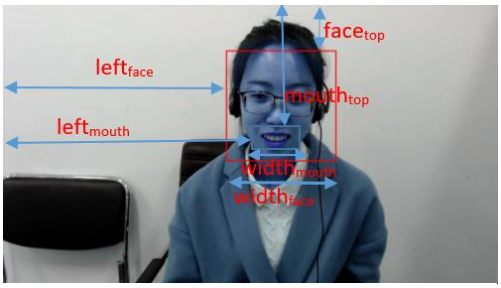

Figure. 4 Mouth Region of Interest Extraction

The mouth region can be extracted using Eq. (22), Eq. (23), Eq. (24) and Eq. (25) as given in [34]. When face is detected, left, width, top and height of facial region denoted as $left_{face}$, $width_{face}$, top_{face} and $height_{face}$ are determined. As mouth is located in lower half of the face, mouth region is extracted from facial region as demonstrated in Fig 4:

$$left_{mouth} = left_{face} + \frac{(width_{face} - left_{face})}{4} \quad (22)$$

$$width_{mouth} = width_{face} - \frac{(width_{face} - left_{face})}{4} \quad (23)$$

$$top_{mouth} = top_{face} + \frac{(height_{face} - top_{face})}{1.5} \quad (24)$$

$$height_{mouth} = height_{face} - \frac{(height_{face} - top_{face})}{1.5} \quad (25)$$

3.4 Hybrid architecture

MobileNetV2 architecture is a benchmark neural network architecture that proves to be very much efficient and accurate when compared with its predecessor architectures. One of the essential component present in MobileNetV2 is the presence of lightweight depthwise convolutions. It is well-noted that EfficientNet family of models employ MBConv layers in it. Though slow, this layer makes EfficientNet more efficient. Hence, simple convolution layers are replaced with a variant of MBConv layer called as Fused MBConv layer into InceptionV3 architecture. Fused MBConv layers are employed in EfficientNetV2 architecture. MBConv layers are composed of Convolution1*1 layer followed by depthwise convolution 3*3, Squeeze and

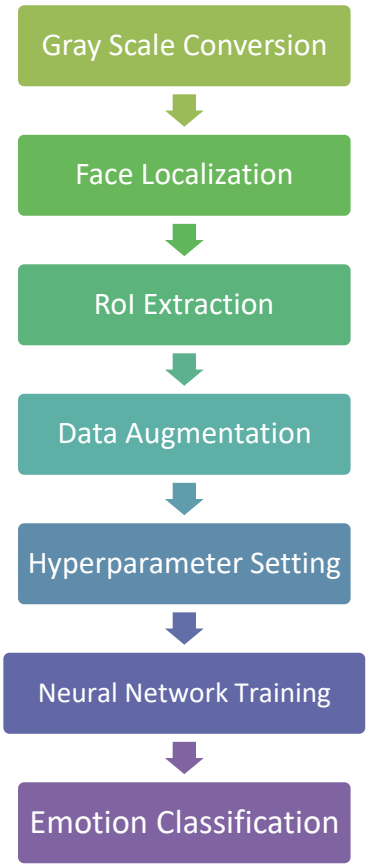

Figure. 5 Proposed Methodology

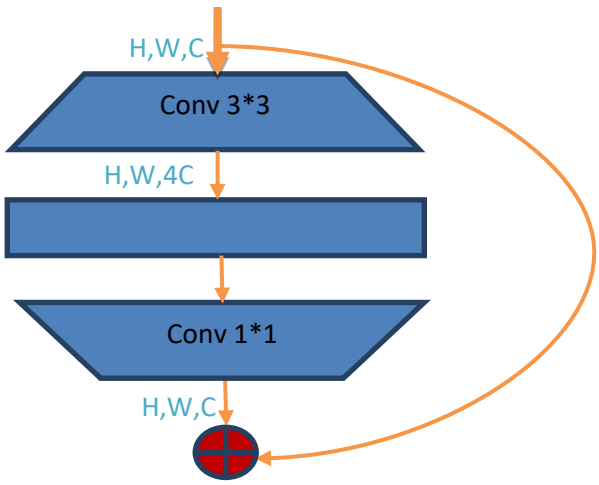

Figure. 6 Fused MB Conv Block

Excitation (SE) and finally, Convolution 1*1 layer. The fused MBConv layer employs Convolution 3*3layer instead of Convolution 1*1 layer and Depthwise Convolution 3*3 layer of MBConv layer.The methodology is demonstrated in Fig. 2, Fig. 5 and Fig. 6.

The list of notations used in Section 2 and Section 3 are tabulated in Table 4.

4. Experimental results

4.1 Performance of inceptionV3 and VGG-16 on various facial parts

Emotion recognition was experimented with OLSFED dataset using InceptionV3 and VGG16 architectures. In case of InceptionV3 and VGG16, the hyper-parameters used are 5 epochs, sparse categorical cross entropy as loss function and Adam optimizer. ReLu activation function was used in Dense layers and softmax activation function was used in the final layer. And the dataset is divided into training and test set with ratio of 80:20. To avoid overfitting, early stopping is adopted. Images are resized to a size of 224*224. Experiments were done on various parts of the face. Imagenet weights were employed.

For InceptionV3 architecture, it is observed that periocular region gives more accurate representation of human emotions with an accuracy of 83.6%. While comparing the losses, validation loss for periocular emotion recognition was 0.4586. Training accuracy was 75.6% and training loss was 0.6226. Non-periocular region gives less validation accuracy of 50% for human emotion recognition with validation loss of 1.33. Training accuracy was only 40% for non-periocular region. Mouth emotion recognition yielded validation accuracy of 93.5%, validation loss of 0.1989, training accuracy of 86.2% and training loss of 0.3837. When both periocular and mouth regions were employed, the training accuracy was 88.48% with training loss of 0.3307. The validation accuracy was 95.66% and the loss was 0.1410. It is to be noted that the validation accuracy was only 89.12% with loss of 0.3497 for the entire face. The same dataset was experimented with VGG-16 architecture. Periocular emotion recognition was successful with 89.32% training accuracy, 0.3119 training loss, 89.97% validation accuracy and 0.3078 validation loss. Mouth emotion recognition performed well with 95.09% training accuracy, 0.1480 training loss, 95.82% validation accuracy and 0.1448 validation loss. With periocular and mouth regions, validation accuracy was 99.31% and 0.0320 validation loss. It is to be noted that the validation accuracy was only 72.18% with loss of 0.7836 for the entire face. From these experiments held, it is observed that periocular and mouth regions together are more effective in emotion classification with lesser number of pixels than employing the entire face for emotion recognition. Entire face was 224*224 with 50176 pixels. But the periocular region is of size 75*54 and the mouth region is of size 168*75. Hence, 16650 pixels are effective for emotion recognition compared to the entire face. The experiments were also held on left and right parts of the face. It must be noted that accuracy was slightly higher for right side of face during experimentation with both InceptionV3 and VGG-16 architectures. Experiments were conducted with 12th Gen Intel(R) Core(TM) i5-1235U 1.30 GHz processor with 16.0 GB RAM on Windows 10 Pro Version 22H2. Python Tensorflow library was employed for implementing InceptionV3 and VGG-16 architectures.

Table 5. Peak Performance of various architectures on OLSFED Datasets

Archite-cture	Region	TA	TL	VA	VL
Inception V3	Periocular	75.69	0.6226	83.63	0.4586
	Non-Periocular	40.08	14.2492	50.88	1.3304
	Mouth	88.58	0.3310	93.63	0.1929
	Non-Mouth	86.57	0.3586	94.84	0.1803
	Periocular + Mouth	88.48	0.3307	95.66	0.1410
	Left Face	67.46	0.7805	87.34	0.4052
	Right Face	83.17	0.4534	91.32	0.2446
	Whole Face	78.77	0.5875	89.12	0.3497
VGG16	Periocular	89.32	0.3119	89.97	0.3078
	Non-Periocular	95.75	0.1609	95.88	0.1477
	Mouth	95.09	0.1480	95.82	0.1448
	Non-Mouth	98.94	0.0483	98.79	0.0528
	Periocular + Mouth	98.82	0.0399	99.31	0.0320
	Left Face	96.44	0.1434	94.04	0.1908
	Right Face	97.61	0.0909	97.53	0.0910
	Whole Face	71.56	0.8412	72.18	0.7836

4.2 Performance of hybrid inceptionV3 on various facial parts

When experimenting Hybrid InceptionV3 architecture on OLSFED dataset, enhanced results were obtained.

The validation accuracy of hybrid model was 86.14% which was 3% higher than the validation accuracy of InceptionV3 for the periocular region. For the mouth region, the validation accuracy of the hybrid model was 95.88% which was 2% higher than the test accuracy of InceptionV3. When both the regions were only considered ignoring the rest of the face, the validation accuracy was 95.69% which was similar to the validation accuracy of InceptionV3.

When the entire face was considered, the validation accuracy was 96.26% which was 7% higher than the validation accuracy of InceptionV3 architecture. When the performance of the hybrid InceptionV3 and VGG16 are compared, the accuracy of VGG16 was higher. The results of all the experiments conducted as mentioned above are given in Table 5, Fig. 7 and Fig. 8.

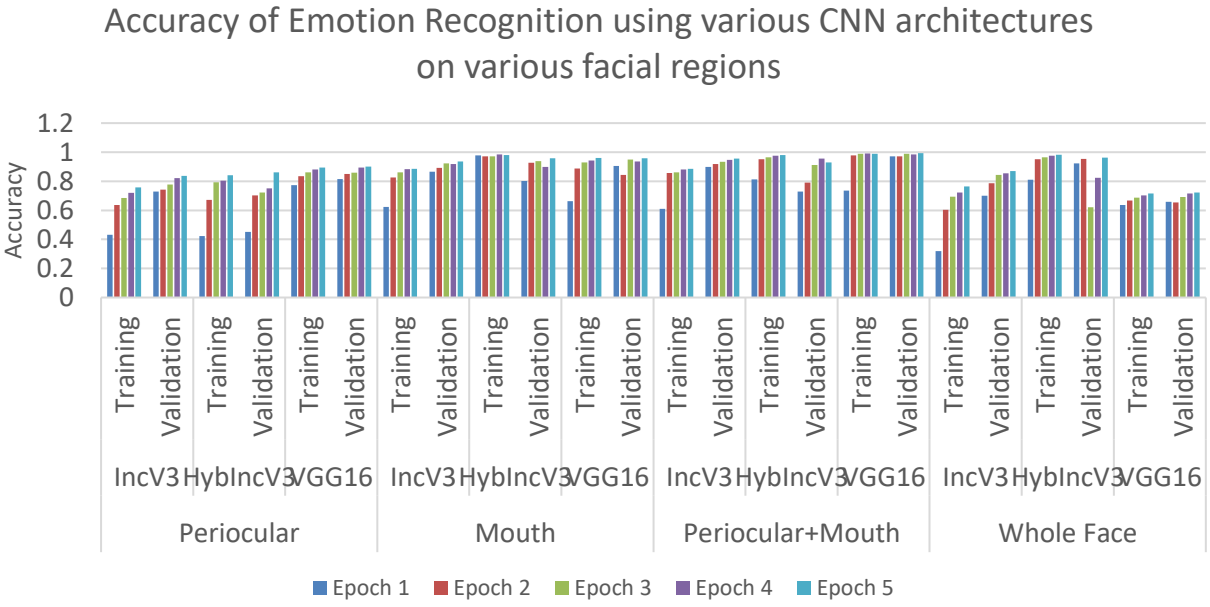

Figure. 7 Facial emotion recognition in multiple Comparison of Training and Validation Accuracy of Facial Emotion Recognition using InceptionV3, Hybrid Inception V3 and VGG-16 on various facial regions

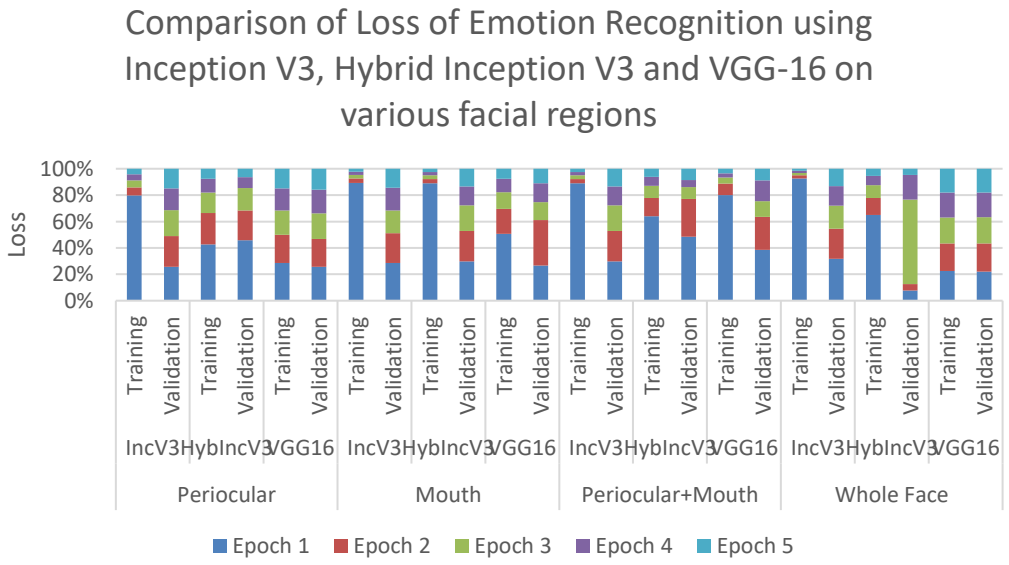

Figure. 8 Comparison of Training and Validation Loss of Facial Emotion Recognition using InceptionV3, Hybrid Inception V3 and VGG-16 on various facial regions

Received: August 13, 2024. Revised: September 24, 2024.

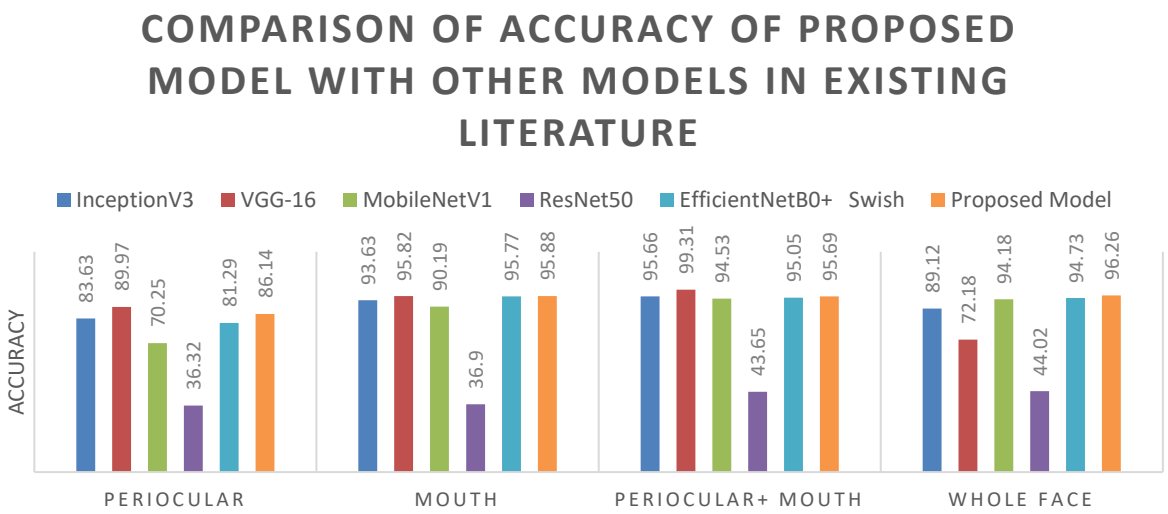

Figure. 9 Comparison of Accuracy of Proposed Hybrid InceptionV3 Model with InceptionV3, VGG-16, MobileNetV1, ResNet50 and EfficeintNetB0+Swish Models

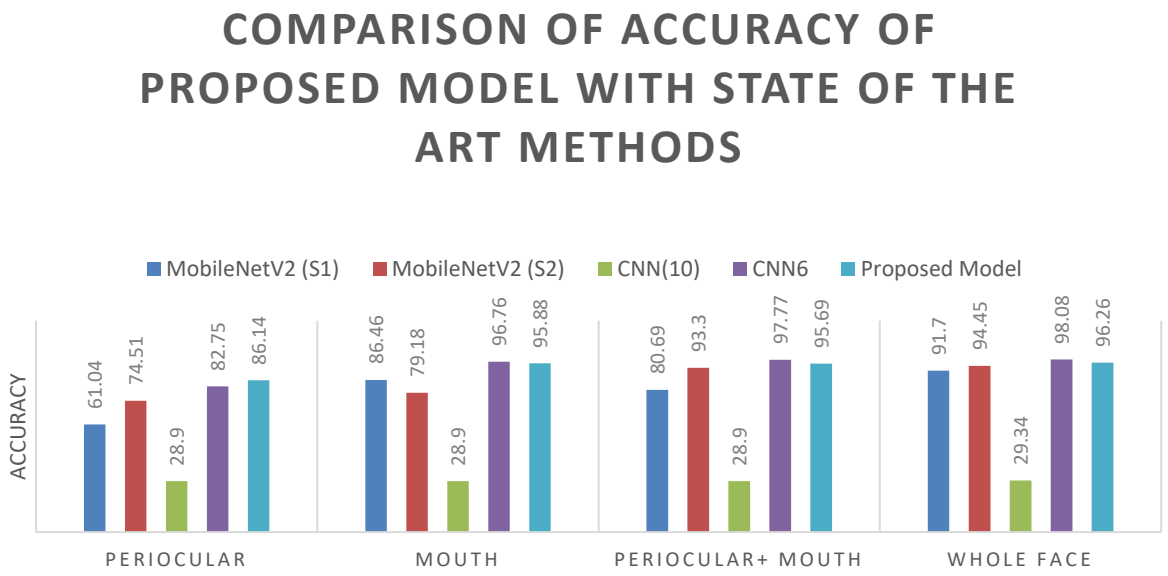

Figure. 10 Comparison of Accuracy of Proposed Hybrid InceptionV3 Model with MobileNet V2 (Scenario1), MobileNetV2 (Scenario2), CNN10 and CNN6 Models

4.3 Performance of proposed model with existing and state-of-the-art models

While analysing the methodologies of the existing literature, it was observed that VGG-16 and its variants, ResNet50, EfficientNet and its variants, MobileNet and Inception architectures were predominantly employed. Hence, the peak performance of proposed method is compared with the peak performance of VGG-16 [10], ResNet50[11], MobileNetV1 [12], InceptionV3 [11] and EfficientNetB0 (with Swish activation function) [13] on the whole face and its various parts on OLSFED dataset. Comparison of accuracy of all these models is illustrated in Fig. 9. From Fig. 9., it was observed that the proposed model gives the second highest accuracy for periocular + mouth and periocular regions next to VGG-16. For different RoI and facial regions, it has surpassed the performance of InceptionV3, MobileNetV1, ResNet50 and EfficientNetB0. On comparison with the models mentioned in Fig. 9., the periocular + mouth RoI of face proves to be effective in emotion recognition and gives higher or very close accuracy to the entire face region. While conducting experiments, same hyper-parameters were used. Sparse Categorical Cross Entropy Loss function, Adam Optimizer, 5 epochs and batch size of 32 were employed. The state-of-the-art neural networks are summarized in Table 6.

Table 6. Summary of State-of-the-art Neural Network Architectures

Architecture	Description
MobileNetV2 (Scenario 1) [35]	Depthwise separable convolutions and Inverted residual structure with linear bottlenecks are employed. While in Scenario 1 (S1) all the convolutional layers are freezed, 50% of the convolutional layers are unfreezed in Scenario 2 (S2).
MobileNetV2 (Scenario 2) [35]	
CNN6 [36]	There are five convolutional – maxpool layers followed by a fully connected layer and softmax layer.
CNN10 [37]	There are two convolutional layers, Leaky ReLU layer, Batch Normalization, Max-pooling, Dropout layer, Flatten layer, Dense, Dropout and Dense Layer

The peak performance of MobileNetV2 (Scenario 1 abbreviated as S1) [35], MobileNetV2 (Scenario 2 abbreviated as S2) [35], CNN6 [36] and CNN10 [37] are also compared against the proposed model as represented in Fig. 10. From Fig. 10, it was observed that CNN10 has the poorest performance for different RoI and face. CNN6 architecture gives the best performance on mouth, periocular + mouth and whole face region. The second best performance was given by the proposed Hybrid Inception V3 model. To infer the significance of periocular + mouth RoI over the whole face region, performance of emotion recognition on these regions together and face was compared for all the models in Fig. 10. The performance of emotion recognition on periocular + mouth RoI was 0.31% to 1.15% less than emotion recognition on the whole face for all the models in Fig. 10.

5. Conclusion

Emotion recognition is a key research area in the recent times. There are several applications for facial emotion recognition techniques: emotion monitoring of elderly patients and autism disordered children, emotion analysis of online learners, gaming and entertainment sectors. Emotion recognition is very essential for engagement detection in online learners.

A hybrid Inception V3 model is proposed in this work. This model replaces convolution layer with Fused MB Convolution Layer and gives greater accuracy than Inception V3 model while classifying emotions. Experiments were conducted on OLSFED dataset to compare the performance of the proposed model against benchmark neural network architectures like VGG-16, ResNet-50, EfficientNetB0 (with Swish activation function), InceptionV3 and MobileNetV1 while employing same set of hyper-parameters. The performance of Hybrid Inception V3 exceeds all the benchmark architectures for only the mouth region and entire face region. Hybrid InceptionV3 records the second highest performance for only the periocular region and the periocular + mouth region, with VGG-16 emerging as the best performing architecture. On comparing Hybrid InceptionV3 with state-of-the-art architectures like CNN10, CNN6, MobileNet V2 (S1) and MobileNet V2 (S2), the proposed model gives second best performance, with CNN6 emerging as the best performing architecture.

In this paper, while performing emotion recognition on various facial regions, it was observed that periocular and mouth regions separately give promising results for emotion recognition. The performance of facial emotion recognition and emotion recognition on periocular + mouth regions on face were thoroughly compared. Experimental results prove that emotion recognition on periocular + mouth regions give accuracy very close to emotion recognition on entire face even with lesser number of pixels.

In future, the following issues could be addressed by the researchers:

1. Facial emotion recognition in medically altered faces needs great focus.
2. Facial emotion recognition in multiple ethnicities may have deviations.
3. Facial emotion recognition is quite challenging when the subjects are wearing facial masks as during the Covid-19 pandemic.
4. Again, facial emotion recognition from surveillance cameras is very challenging.
5. Data calibration is a necessity when working on faces of subjects from different geographical origins.
6. Changes in emotion and the frequency of change of emotion needs to be studied in controlled and uncontrolled environments.
7. Optimal bag of soft biometric traits for emotion recognition needs to be determined while data collection could be time-consuming.
8. Hand gestures are vital in deducing emotions with accuracy.
9. Integration of emotion detection systems into relevant applications like Intelligent Tutoring Systems, Gaming and Entertainment Applications, etc.,
10. Re-aligning people with negative emotions to experience positive emotions needs special focus.
11. Defining optimal RoI in human faces for emotion detection.

12. Literature on Privacy and security issues of Facial Emotion Recognition is very limited.
13. The degree of correlation between positive emotions and task evaluations needs to be studied.
14. Optimal fusion of multi-modal soft biometrics requires further study.
15. Alignment, translation and representation are some of the challenges that need to be addressed.
16. Emotions of groups and sub-groups of population needs to be determined and analysed to accelerate personalized and group academic development.
17. Studies on thermal images for emotion recognition is very much limited.

Conflicts of Interest

The authors declare no conflict of interest.

Author Contributions

Conceptualization, methodology and formal analysis: Evangeline D and Parkavi A; investigation: Evangeline D; writing-original paper draft: Evangeline D; writing review and editing: Evangeline D and Parkavi A.

Acknowledgments

The work did not receive any funding.

References

[1] S. B. Daily, M. T. James, D. Cherry, J. J. Porter, S. S. Darnell, J. Isaac and T. Roy, "Emotions and affect in human factors and human-computer interaction", *Affective computing: Historical foundations, current applications, and future trends*, pp. 213-231, 2017.

[2] E. Hudlicka, "Computational Modeling of Cognition-Emotion Interactions: Theoretical and Practical Relevance for Behavioral Healthcare", *Emotions and Affect in Human Factors and Human-Computer Interaction*, pp. 383-436, 2017.

[3] I. Michael Revina and W. R. Sam Emmanuel, "A Survey on Human Face Expression Recognition Techniques", *Journal of King Saud University - Computer and Information Sciences.*, pp. 619-628, 2018.

[4] Y. Miao, H. Dong, J. M . Al Jaam and A.E. Saddik, "A Deep Learning System for Recognizing Facial Expression in Real-Time", *ACM Transactions on Multimedia Computing, Communications, and Applications*, Vol. 15, No. 2, pp. 1-20, 2019.

[5] H. C. Lee, C. Y. Wu, T. M. Lin, "Facial Expression Recognition Using Image Processing Techniques and Neural Networks", In: *Pan JS., Yang CN., Lin CC. (eds) Advances in Intelligent Systems and Applications - Volume 2. Smart Innovation, Systems and Technologies*, Berlin, Heidelberg, 2013.

[6] M. B. Akcay and K. Oğuz., "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers", *Speech Communication*, Vol. 116, pp. 56-76, 2020.

[7] P. M. Ashok Kumar, J. B. Maddala and K. M. Sagayam, "Enhanced Facial Emotion Recognition by Optimal Descriptor Selection with Neural Network", *IETE Journal of Research.*, Vol. 65, No. 5, pp. 2595-2614, 2021.

[8] R. A. Khalil, E. Jones, M. I. Babar and T. Jan, "Speech Emotion Recognition Using Deep Learning Techniques: A Review", *IEEE Access*, Vol. 7, pp. 117327-117345, 2019.

[9] P. Ekman, "Universals and cultural differences in facial expressions of emotion", In: *Proc. of Nebraska Symposium on Motivation*, Vol. 19, pp. 207-283, 1971.

[10] L.Yao, Y. Wan, N. Hongjie and B. Xu, "Action unit classification for facial expression recognition using active learning and SVM", *Multimedia Tools and Applications*, Vol. 80, pp. 24287-24301, 2021.

[11] K. R. Scherer, "What are emotions? And how can they be measured?", *Social science information*, Vol. 44, No. 4, pp. 695-729, 2009.

[12] Y. Wang, W. Song, W. Tao, A. Liotta, D. Yang, X. Li, S. Gao, Y. Sun, W.Ge, W. Zhang and W.Zhang., "A systematic review on affective computing: emotion models, databases, and recent advances", *Information Fusion*, Vol. 83-84, pp. 19-52, 2022.

[13] M. Horvat, A. Stojanovic and Z. Kovacevic, "An overview of common emotion models in computer systems", In: *Proc. of 45th Jubilee International Convention on Information, Communication and Electronic Technology (MIPRO)*, Opatija, Croatia, pp. 1008-1013, 2022.

[14] M. M. T. Zadeh, M. Imani and B. Majidi, "Fast Facial emotion recognition Using Convolutional Neural Networks and Gabor Filters", In: *Proc. of 5th Conf. on Knowledge Based Engineering and Innovation (KBEI)*, Tehran, Iran, pp. 577-581, 2019.

[15] T.S. Ashwin and R. M. R. Guddeti, "Affective database for e-learning and classroom environments using Indian students' faces, hand

gestures and body postures", *Future Generation Computer Systems*, Vol. 108, pp. 334-348, 2020.

[16] T. S. Ashwin and R. M. R. Guddeti, "Unobtrusive Behavioral Analysis of Students in Classroom Environment Using Non-Verbal Cues", *IEEE Access*, Vol. 7, pp. 150693-150709, 2019.

[17] L. Qianqian et al., "Research on Behavior Analysis of Real-Time Online Teaching for College Students Based on Head Gesture Recognition", *IEEE Access,*, Vol. 10, pp. 81476-81491, 2022.

[18] J. Bao, X. Tao and Y.Zhou, "An Emotion Recognition Method Based on Eye Movement and Audiovisual Features in MOOC Learning Environment", *IEEE Transactions on Computational Social Systems*, Vol. 11, No.1, pp. 171-183, 2022.

[19] K. Altuwairqi, S. K. Jarraya, A. Allinjjawi and M. Hammami, "Student behavior analysis to measure engagement levels in online learning environments", *Signal, Image and Video Processing*, Vol. 15, No. 7, pp. 1387-1395, 2021.

[20] E. M. Albornoz, D. H. Milone and H. L . Rufiner, "Spoken emotion recognition using hierarchical classifiers", *Computer Speech and Language*, Vol. 25, No. 3, pp. 556-570, 2011.

[21] H. Cheng and X. Tang, "Speech Emotion Recognition based on Interactive Convolutional Neural Network", In: *Proc. of 3rd International Conf. on Information Communication and Signal Processing (ICICSP)*, Shanghai, China, pp. 163-167, 2020.

[22] R. Y. Cherif, A. Moussaoui, N. Frahta and M.Berrimi, "Effective speech emotion recognition using deep learning approaches for Algerian dialect", In: *Proc. of 2021 International Conf. of Women in Data Science at Taif University (WiDSTaif)*, Taif, Saudi Arabia, pp. 1-6, 2021.

[23] S. W. Byun and S. P. Lee, "Human emotion recognition based on the weighted integration method using image sequences and acoustic features", *Multimedia Tools and Applications*, Vol. 80, pp. 35871-35885, 2021.

[24] S. L. Happy, P. Patnaik, A.Routray and R. Guha, "The Indian Spontaneous Expression Database for Emotion Recognition", *IEEE Transactions on Affective Computing*, Vol. 8, No. 1, pp. 131-142, 2017.

[25] F. Ye, "Emotion Recognition of Online Education Learners by Convolutional Neural Networks", *Computational Intelligence and Neuroscience*, Vol. 4316812, 2022.

[26] S. Gupta, P. Kumar, R. K. Tekchandani, "Facial emotion recognition based real-time learner engagement detection system in online learning context using deep learning models", *Multimedia Tools and Applications*, Vol. 82, No. 8, pp. 11365-11394, 2023.

[27] H. Wang, D. P. Tobón V., M. S. Hossain and A. E. Saddik, "Deep Learning (DL)-Enabled System for Emotional Big Data", *IEEE Access*, Vol. 9, pp. 116073-116082, 2021.

[28] T. Dar, A. Javed, S. Bourouis, H. S. Hussein and H. Alshazly, "Efficient-SwishNet Based System for Facial Emotion Recognition", *IEEE Access*, Vol. 10, pp. 71311-71328, 2022.

[29] P. Barra, L. D. Maio and S. Barra, "Emotion recognition by web-shaped model", *Multimedia Tools and Applications*, Vol. 82, pp. 11321-11336, 2022.

[30] V. S. Avani, S. G. Shaila and A. Vadivel, "Geometrical features of lips using the properties of parabola for recognizing facial expression", *Cognitive Neurodynamics*, Vol. 15, No. 3, pp. 481-499, 2021.

[31] M. B. Myneni, H. Akkineni and C. Srinivasulu, "An Approach for Learner Categorization Based on Emotions in Intelligent Adaptive E-Learning Environment", *Journal of Mobile Multimedia*, Vol. 18, No. 6, p. 1709-1732, 2022.

[32] E. Owusu, J. K. Appati and P. Okae, "Robust facial expression recognition system in higher poses", *Visual Computing for Industry, Biomedicine, and Art volume*, Vol. 5, No. 14, pp. 1-15, 2022.

[33] P. Kumari and K. R. Seeja, "Periocular biometrics: A survey", *Journal of King Saud University - Computer and Information Sciences*, Vol. 34, No. 4, pp. 1086-1097, 2022.

[34] N. Puviarasan and S. Palanivel, "Lip reading of hearing impaired persons using HMM", *Expert Systems with Applications*, Vol. 38, No. 4, pp. 4477-4481, 2011.

[35] E. S. Agung, A. P. Rifai and T. Wijayanto, "Image-based facial emotion recognition using convolutional neural network on emognition dataset", *Scientific Reports*, Vol. 14, No. 14429, pp. 1-22, 2024.

[36] A. Qazi, M. Farooq, F. Rustam, M. Villar, C. Rodríguez and I. Ashraf, "Emotion Detection Using Facial Expression Involving Occlusions and Tilt", *Applied Sciences*, Vol. 12, No. 11797, pp. 1-24, 2022.

[37] E. G. Dada, D. O. Oyewola, S. B. Joseph, O. Emebo and O. O. Oluwagbemi, "Facial Emotion Recognition and Classification Using the Convolutional Neural Network-10 (CNN-10)",

Applied Computational Intelligence and Soft Computing, Vol. 2023, No. 2457898, pp. 1-19, 2023.

[38] C. Bian, Y. Zhang, F. Yang, W. Bi and W. Lu, "Spontaneous facial expression database for academic emotion inference in online learning", *IET Computer Vision*, Vol. 13, No. 3, pp. 249-353, 2019.