Hindawi
Scientific Programming
Volume 2023, Article ID 6859284, 21 pages
https://doi.org/10.1155/2023/6859284

Research Article

Ensemble Convolution Neural Network for Robust Video Emotion Recognition Using Deep Semantics

E. S. Smitha,[1] **S. Sendhilkumar**,[1] **and G. S. Mahalakshmi**[2]

[1]*Department of Information Science & Technology, College of Engineering, Anna University, Chennai, Tamilnadu, India*
[2]*Department of Computer Science & Engineering, College of Engineering, Anna University, Chennai, Tamilnadu, India*

Correspondence should be addressed to E. S. Smitha; smithaengoor@gmail.com

Received 13 February 2023; Revised 14 April 2023; Accepted 2 May 2023; Published 17 May 2023

Academic Editor: Dongpo Xu

Human emotion recognition from videos involves accurately interpreting facial features, including face alignment, occlusion, and shape illumination problems. Dynamic emotion recognition is more important. The situation becomes more challenging with multiple persons and the speedy movement of faces. In this work, the ensemble max rule method is proposed. For obtaining the results of the ensemble method, three primary forms, such as $CNN_{HOG\text{-}KLT}$, $CNN_{Haar\text{-}SVM}$, and CNN_{PATCH} are developed parallel to each other to detect the human emotions from the extracted vital frames from videos. The first method uses HoG and KLT algorithms for face detection and tracking. The second method uses Haar cascade and SVM to detect the face. Template matching is used for face detection in the third method. Convolution neural network (CNN) is used for emotion classification in $CNN_{HOG\text{-}KLT}$ and $CNN_{Haar\text{-}SVM}$. To handle occluded images, a patch-based CNN is introduced for emotion recognition in CNN_{PATCH}. Finally, all three methods are ensembles based on the Max rule. The $CNN_{ENSEMBLE}$ for emotion classification results in 92.07% recognition accuracy by considering both occluded and nonoccluded facial videos.

1. Introduction

Human emotions are inevitable in day-to-day interactions. They catalyse improvising communication. Generally, humans use face, hand, voice, and body gestures to express their feelings. Among all these, human faces have been the most prominent and expressive medium in carrying out emotions during interactions. Facial emotion recognition is the technology used to reveal information from one's emotional state or sentiments by analysing facial expressions from both static and videos. This is a part of affective computing.

Emotions are integral to human communication: during smiles, showing greetings and respect to others, frowning in confusion, raising voice during arguments, and so on. They are the best means of nonverbal communication, irrespective of culture, religion, or race. It assists in determining how someone feels by obtaining information about their emotional state. So, it can be used for verification and recognition purposes.

In conventional market research, companies use surveys and customer reviews (verbal methods) to understand the demands and needs of customers. The other method is behavioural, in which companies record video feeds of users interacting with a product. They manually analyse the video to observe a user's reactions and emotions. This method is useful, but it is time-consuming and tedious. Furthermore, it raises the overall cost. Market research firms can now easily automate video analysis and detect their users' facial expressions using artificial intelligence (AI)-enabled facial emotion recognition systems. This saves time and labour while also lowering costs. Market research firms can use facial recognition systems to scale their data collection efforts.

1.1. Benefits of Emotion Detection

1.1.1. Assess Personality Traits in Interviews. Personal interviews are an excellent way to interact with potential candidates and determine whether they are a good fit for the

position. However, analysing a candidate's personality in such a short period of time is not always possible. Furthermore, many categories of discussion and judgement add to the complexity. Through facial expressions, emotion detection can assess and measure a candidate's emotions. It assists interviewers in comprehending a candidate's mood and personality traits. Human resources can use this technology to develop recruiting strategies and policies to get the most out of their employees.

1.1.2. Product Testing and Client Feedback. When customers try a product, emotion detection technology can help the product industry understand their genuine emotions. Companies can set up a product testing session, record it, and then analyse it to detect and assess the facial emotions that emerge during the session. Because AI powers emotion detection, it can evaluate user reactions to new product launches.

1.1.3. Enhances Customer Service. Emotion detection improves the user experience in nearly every industry. This technology enables retailers to create more personalised customer offers by analysing their browsing and purchasing habits. Furthermore, healthcare providers can use facial recognition to create better care plans and deliver services much more quickly. In Phycology and Crime Prediction. Human emotion recognition has applications in psychology. Emotions, which are active most of the time, control our behaviours unconsciously. Emotions can significantly influence criminal behaviour. According to criminal psychologists [1], there are nine levels of emotional motivation for criminal conduct: bothered, annoyed, indignant, frustrated, infuriated, hostile, wrath, fury, and rage. Using emotion analysis, the phycologist can understand a person's emotions and emotional fluctuations. As a result, emotion recognition can be used to predict crime.

1.2. Classes of Emotion. Emotions are generally classified as positive or negative. The six basic emotions are anger, happiness, fear, disgust, sadness, and surprise. Other emotions are an embarrassment, interest, pain, shame, shyness, anticipation, smile, laugh, sorrow, hunger, and curiosity. Emotions can be discrete or dimensional [2]. According to Natya Sastra [3], nine basic emotions are identified. They are love, laughter, sorrow, anger, courage, fear, disgust, surprise, and peace.

If a person is angry, the eyebrows are drawn together and lowered inside and the extremes are pulled outward, either in both eyebrows or in one. Vertical lines, generally two, appear between the eyebrows, lower lid raised, eyes focusing at the centre, stare or bulging and lips tightly pressed together with corners down or square shape, moustache or the upper part of the mouth curved, nostrils may be dilated, the lower jaw projecting out. If a person is happy, the corners of the lips are drawn wide and upwards, the lips parted and teeth exposed, or lips widened with wrinkles running from the outer nose to outer lip, a portion of cheeks below the eyes are raised, lower lid may show wrinkles or be tense, exhibiting crow's feet near the corner of both the eyes. Therefore, identifying facial expressions is the key to emotion recognition.

For facial expression recognition [4], extraction of facial features for capturing the changes in appearance is achieved via harvesting deep feature semantics. To achieve this, traditional CNNs are optimised using a soft-max loss, which penalises the misclassified samples, forcing the features of different classes to stay apart.

Deep convolutional neural networks (CNN) otherwise require massive training to produce better accuracy. Due to the limited public database available for facial expressions, data augmentation mechanisms must be employed. Cropping sample images at varied angles results in images at various positions at varied scales, which further reduces the sensitivity of the overall system. Therefore, if augmented for experimental purposes, the utmost care should be taken in data to preserve the model's robustness.

Deep CNNs can hierarchically learn the features from samples to represent all possible complex variations of input images [5]. The max pooling layers only consider input features' first-order statistics, which limits learning deep semantic features. This becomes further complicated when pose variations and/or partial facial occlusions are included. Occlusions and variant poses are two major factors causing the significant change in facial appearance. Removing occluded regions is not practical when real-time video emotion recognition is considered. Real-world occlusion is yet another difficult task for emotion detection research. Using CNN to ignore occlusion and pose variations might lead to inaccurate facial features.

In contrast to the above claim, human intelligence can exploit local facial regions and holistic faces for better perception of emotions in partial or complete facial occlusions. Due to the dynamism in the variation of local parts such as the eyes, nose, and mouth, the vital issue rests in the robust detection of such energy from every keyframe. Directly feeding the keyframes leads to underutilising prior knowledge hidden in consecutive frames. Hand-crafted facial descriptors are unsuitable for interpreting the powerful temporal features in facial images.

Several mathematical models capable of processing under adverse conditions have been proposed to address these challenges. Constrained frontal faces shall be analysed by facial expression classifiers [6]. Subspace analysis techniques [7] require extensive training and are not suitable. Recognition systems based on local feature representations [8] respond better under face illumination variations. However, occlusion and posing reduce the accuracy. Stacked supervised autoencoder [9] is better at solving the above problems. However, they need accurate training data which is occlusion-free. The proposed work concentrates on emotion recognition from videos performed via ensemble-based approaches, as there is less robustness in processing video data streams as in [6]. Though the work tends to achieve dynamic emotion recognition via crucial frame extraction, face recognition, and further processing of image-based emotion detection, the idea of SSPP is not addressed as it might be computationally expensive and lead to delayed recognition.

Therefore, the proposed work discusses using KLT tracking to track the recognised faces in video accurately. The extraction of visual facial features for facial recognition is crucial since the colour and shape of faces in video are similar. In this work, HOG features are used to precisely capture facial features such as directions and edges of the face and facial intensities, which are later fed to the SVM classifier for robust facial recognition. Avoiding a large number of layers and the need for colossal training, the proposed work encompasses nine layers of CNNs with extensive data augmentation. The primary goal is to analyse the emotion of all the persons in a given video.

1.3. Challenges. A facial expression is representative of a specific emotion, so it is not easy even for humans to recognise emotions accurately. Studies show that different people recognise different emotions in the same facial expression. And it is even more challenging for AI to differentiate between these emotions.

1.3.1. Technical Challenges. Emotion recognition shares many challenges. Identifying an object, continuous detection, and incomplete or unpredictable actions are the most widespread technical challenges of implementing emotion recognition.

Face occlusion is one of the main challenges in captured videos and pictures. Another commonly seen challenge is lighting issues. Identifying facial features and recognising unfinished emotions are crucial challenges in emotion recognition.

1.3.2. Psychological Challenges. Psychologists have studied the connection between facial expressions and emotions since the middle of the 19th century. Cultural differences in emotional expression are one of the main challenges. Infants and children indicate feelings differently than adults, so identifying children's emotions is another challenge.

An ensemble method of image classification is proposed to improvise emotion recognition. Ensemble learning aims to assemble diverse models or multiple predictions and boost prediction performance. The ensemble combines numerous learning algorithms to obtain their collective performance and improve the performance of existing models by combining several models, resulting in one reliable model. For image classification, both occluded and nonoccluded facial videos are used.

The online social networks produce billions of visual information, which are useful to recognise sentiment. A proverb in many languages goes, "A picture is worth a thousand words," which means that a single image can convey many ideas more effectively than spoken description. Visual sentiment analysis on social network content helps us to understand the user behaviour and provides useful information for related data analysis. The majority of users of social networking platforms prefer to use images and emoticons other than typing long sentences. The Twitter platform encourages the communication between users using short texts or images. The main objective of the work is to classify the sentiment behind the messages represented in the form of image in social networks, using state of art machine learning and deep learning methods.

(i) To propose new techniques for face detection

(ii) To develop a system with high accuracy of face emotion detection

(iii) To propose an ensemble convolution neural network for face emotion classification.

The primary goal of this paper is to examine social media post in the form of image data to identify user attitudes regarding a particular topic of discussion. Utilizing attitudes toward a topic of discussion on social media can help to identify and predict the sentiments. Additionally, it aids in assess personality traits in interviews, product testing and client feedback, enhance customer service to quickly adapt to constantly changing needs. This paper first proposes an ensemble deep learning classification algorithm, namely, $CNN_{ENSEMBLE,}$ which is proposed in this work by combining the outcomes of $CNN_{HOG\text{-}KLT}$, $CNN_{Haar\text{-}SVM}$, and CNN_{PATCH} for the analysis of sentiments. This proposed ensemble deep learning classification algorithm is employed in this work for performing classification over occluded and nonoccluded social media postimages to improve the accuracy of emotion classification.

Therefore, the main contributions of this paper are:

(i) The first contribution of this work is that it proposes three different face detection methods such as $_{HOG\text{-}KLT}$, $_{Haar\text{-}SVM,}$ and $_{PATCH}$ are used parallel in this work for emotion analysis, which effectively identifies the occluded and nonoccluded faces

(ii) Second, an efficient CNN based emotion classification model that highlights the impact of handling occlusion and nonocclusion for improvising the classification of emotions

(iii) Finally, an ensemble deep learning algorithm named, $CNN_{ENSEMBLE}$ is proposed in this work for performing effective emotion recognition.

These sentiment classification algorithms have been evaluated with extended CK+ for emotion recognition. The results obtained from this work show that $CNN_{ENSEMBLE}$ emerges as the highly accurate model for emotion recognition for occluded and nonoccluded faces.

The rest of this article is organized as follows: Section 2 provides a survey of related work that are existing in the literature on emotion classification and they are compared with the proposed work. Section 3 explains the methods used and the algorithms proposed in this paper. Section 4 discusses about the results obtained from this work and performs a comparison with existing work. Section 5 gives the conclusions arrived from this works and lists some future enhancements.

2. Related Works

In this section, we surveyed the identification of facial expression recognition, occlusion-aware facial expression recognition, and the techniques used for facial emotion recognition.

2.1. Facial Expression Recognition. Nowadays, distance learning and e-classrooms are the part of our life, in e-learning scenario, by analysing the facial expressions, the teachers can understand the engagement of students in learning [10]. Viola–Jones and HAAR cascade algorithms are used for object detection and feature extraction. CNN is used for expression classification. Facial expressions have been analysed for various applications. However, not all the facial regions contribute to expression detection, and a few areas do not change with varied terms [11, 12]. Determining human behaviour from facial expressions is an excellent application in healthcare, tourism and hospitality, and the retail industry. Sajjad et al. [13] proposed a framework for analysing human behaviour via facial expressions from video. The face is initially detected using the Viola–Jones algorithm and then tracked via KLT. Viola–Jones has various stages, including Haar features selection, which selects the most important facial features [14, 15], AdaBoost training, and cascading classifiers.

The localisation of the eyes and nose is detected by Haar cascades [16, 17]. The algorithm first identifies the rectangular area of the eyes and nose position. Later, the eye centres are computed. If there is a mismatch in detection, anthropometric statistics are used. Face alignment is achieved using the position of eyes since eyes do not move with expressions. After the nose position is extracted, the mouth region is identified using the nose region as a reference. The curves in the upper lips shall be detected using horizontal edge detection techniques. The position of the eyes is also used for identifying the eyebrow region of interest.

Second, the detected face (if not priorly registered, is registered into the database and) is recognised using the SVM classifier, followed by facial expression recognition using CNN. The proposed work is constructed along these lines with additional semantic feature extraction using CNN. Bounding box approaches have been combined with confidence score and class prediction parameters within layers of CNN to achieve improved detection accuracy in video surveillance. The proposed work also uses bounding box approaches for improved face detection accuracy.

2.2. Occlusion-Aware Facial Expression Recognition. Occlusion is the primary issue in handling real-time videos. Partial facial occlusion has been widely addressed in the literature. However, real-life occlusion detection is essential for applications in the healthcare and hospitality industries.

In handling occlusions, patch-based approaches [18, 19] have emerged as the state-of-the-art in real-time. VGGNet is used to represent the input image as feature maps. ACNN decomposes the input image feature maps into subfeature maps. This disintegration into multiple subfeature maps results in the identification of local patches. The feature maps are sent to the Gg unit to identify the facial occlusion's location.

Patch-based ACNN (pACNN) performs region decomposition and occlusion perception. Region decomposition uses an exclusive approach [20] to select 24 out of 69 facial landmarks. The local feature maps are cropped and fed to the respective convolution layers using selected landmarks without compromising spatial resolution. After sufficient learning, the feature maps are converted to vector-shaped local features, provided to the attention layer. The attention net determines the scalar weight as a means of quantifying the importance of the identified local patch. Global-local-based ACNN (gACNN) takes care of the later stages of processing. It takes the full-face region and extracts the local details of the patches and their respective global context cues.

2.3. Datasets for Facial Emotion Recognition. Table 1 summarises publicly available video datasets and the addressed emotion categories.

The proposed work assumes CK+ and ISED databases and proceeds to facial expression detection using deep learning models. Additionally, the proposed work concentrates on extracting emotions about basic categories.

2.4. Machine and Deep Learning Approaches for Facial Emotion Recognition. Emotions related to e-learning, like boredom, confusion, contempt, curiosity, disgust, eureka, delight, and frustration were mainly identified in recent literature [32–39]. Deep learning models, mainly convolutional neural networks, are used for emotion classification. Different deep learning models such as VGGNet [34, 39] and ResNet [35] are used for the implementation. A variant of CNN, DCFA-CNN [36], is tested with different image datasets and got excellent classification result. Yolcu et al. [40] presents a deep learning-based system for customer behavior monitoring applications. The system uses 3-cascade CNN for head pose estimation, facial components segmentation, and expression classification. GoogLeNet and AlexNet, which consists of 2 consecutive CNN layers, are widely used in facial expression recognition [41]. Table 2 presents various classifiers used for facial emotion recognition.

3. Ensemble Framework for Facial Emotion Recognition

The framework in Figure 1 discusses the facial emotion recognition from videos using ensemble CNN classifiers. It also highlights the ensemble CNN for robust video emotion detection with late multiple feature fusion using deep semantic facial features. The video is the input used for the recognition of emotion. The video may contain a single person or multiple people, and emotion can be identified for both occluded and nonoccluded faces. The input video contains sequence frames. Initially, the edges will be

Table 1: Facial expression video datasets and emotion categories.

Video dataset	# Videos	Addressed emotions
MMI [21]	2900	Sadness, surprise, fear, happiness, anger, and disgust
CK+ [22]	593	Neutral, sadness, surprise, happiness, fear, anger, contempt, and disgust
DEAP [23]	880	Valence, arousal, and dominance
Belfast [24]	1400	Disgust, fear, amusement, frustration, surprise, anger, and sadness
DISFA [25]	130000	Action unit intensities
RECOLA [26]	3.8 hrs	Valence and arousal
ISED [27]	328	Sadness, surprise, happiness, and disgust
Aff-Wild [[28]	428	Valence and arousal
BP4D [29]	298	Sadness, surprise, fear, anger, embarrassment, physical pain, happiness, and disgust
RAVDESS [30]	7356	Neutral, calm, happiness, sadness, anger, fear, surprise, and disgust
DEFE [31]	164	Neutral, happiness, anger, valence, arousal, and dominance

Table 2: Classifiers used for facial emotion recognition.

Name of classifier	Addressed emotions
DCNN + RLPS [42]	Anger, disgust, fear, happiness, sadness, surprise, and neutral
3DCLS [43]	Angry, sad, happy, and neutral
AdaBoost, relevance feedback [44]	Fear, sadness, and joy
HMM [45]	Sadness, happiness, anger, fear, disgust, and surprise
LDA [46]	Joy, acceptance, fear, surprise, sadness, disgust, anger, and anticipation
Inception V3 [47]	Boredom, confusion, engagement, and frustration
Modified CNN [48]	Ekman's basic emotions
LSTM [49]	Boredom, confusion, engagement, and frustration
Facial action coding system [50]	Ekman's basic emotions
Decision trees, Kinect V2 [51]	Engagement and frustration
WEKA and OpenFace [52]	Boredom, confusion, engagement, and frustration
Linear SVMs [53]	Engagement, liking, and familiarity
PSO, kNN [54]	Happiness, sadness, anger, and fear
Information aggregation, decision fusion [55]	A neutral, slight smile, large smile, small laugh, giant laugh, and thrilled
Hybrid CNNs [56]	Engaged, bored, and neutral
ResNet-18 model and transformers [57]	Fear, angry, sad, happy, contempt, disgust, and surprise
Multitask CNN [58]	Happiness, fear, disgust, and surprise

extracted from videos. The structures may have faces or nonfaces. By using a keyframe extraction method, the keyframes are identified. The idea is to create a model that ensembles the inherent emotional information within the video frames. In this work, three different methods are used for emotion recognition and for improving accuracy. All the methods are fused based on an ensemble strategy. Before emotion recognition, the first step is identifying the faces in the video frame.

In the first method, the face is detected through the Haar cascade algorithm and tracked using KLT tracking. The detected face image will be fed into CNN for further emotion classification. Similarly, face detection is achieved using HOG features and SVM in the second method. The images classified as face images will be the input of the CNN emotion classifier. The third template-matching method is used to detect the faces in the frame. Then, using a patch-based CNN, the emotion of the image will be recognised. This proposed end-to-end trainable Patch-Gated Convolution Neutral Network (PG-CNN) can automatically detect and focus on the most discriminative non-occluded areas of the face. After identifying emotions by three distinct methods, ensemble max rule-based emotion recognition accurately classifies the feelings.

3.1. Keyframe Extraction. Keyframe extraction is used to reduce redundant frames that lead to the dimensionality reduction of the feature vector for classification. The input video is processed for keyframe extraction, where multiple keyframes are extracted in this module. This work uses the histogram difference method for keyframe extraction. The difference is calculated between each frame, and the threshold value is obtained. Consider two frames f_1 and f_2. If there are any changes or differences found in f_2 from f_1, then f_2 is taken into account. If there are no changes next subsequent frame is taken for examination, and the process is continued till f_n frame.

The process has two main phases. In the first phase, the threshold (TD) value will be computed using the mean and standard deviation of the histogram of the absolute difference of successive image frames. In the second phase, compare the threshold (TD) extracted from keyframes against the fundamental difference of consecutive image frames.

The video frames will be extracted one by one at first. The histogram difference between two successive frames will be calculated for each video frame. To determine a threshold point, the mean (M) and standard deviation (SD) of the absolute difference of the histogram are calculated. The

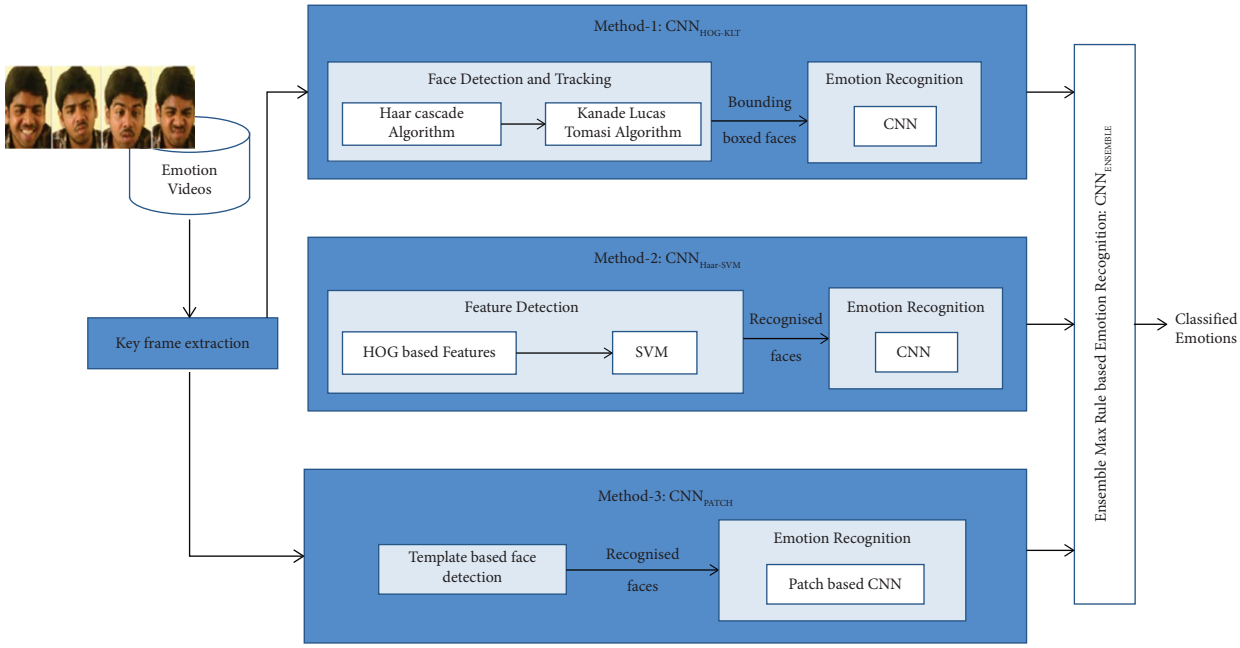

Figure 1: Ensemble framework for facial emotion recognition.

following equation can be used to calculate the threshold (TH), where M_d is the mean of absolute difference and SD_d is the standard deviation of fundamental difference. After obtaining the threshold, the next phase determines the keyframes by comparing the absolute difference of the histogram to the point. The process of histogram difference-based essential frame selection was described in Algorithm 1.

$$TH = M_d + SD_d. \tag{1}$$

3.2. Method 1: $CNN_{HOG\text{-}KLT}$. Method 1, developed inside the ensemble framework, consists of two main steps such as (1) face detection and tracking and (2) emotion recognition.

3.2.1. Face Detection and Tracking. It used the Haar cascade and Kanade Lucas Tomasi algorithm to detect and track faces in the frames extracted from videos. The steps used to compute face detection and tracking are mentioned below.

Step 1: Input the keyframes

Step 2: Identify the relevant features RLF_i using the Haar cascade algorithm, which locates the face. It requires the identification of located feature points needs to be reliably tracked

Step 3: Use the KLT method, which computes the displacement of the tracked points from one keyframe to another. It finds the traceable feature points in the first frame and then follows the detected features in the succeeding frames using the calculated displacement.

(1) Haar Cascade Algorithm. The Haar cascade is used to recognise faces in keyframes. It essentially identifies adjacent rectangular regions in a detection window at a specific location. The calculation entails adding the pixel intensities in each area and subtracting the sums. These features can be challenging to determine for a large image. To overcome the difficulty, integral photos are used, reducing the number of operations compared to larger original images. Necessary images return the pixel value at any (a, b) location is the sum of all pixel values present before the current pixel. Instead of computing at each pixel, it creates subrectangles and array references for each subrectangle. The Haar features are then computed using these. Haar is primarily used to extract three distinct parts: line, edge, and rectangle features. The representation of line, edge, and rectangle features is represented in Figure 2.

The Haar features are applied to determine the facial features in which the line feature, edge feature, and rectangle features are denoted by L_f, E_f, and R_f. The value of the feature VF_i it is calculated by identifying the sum of pixel values in the black area minus the sum of pixel values in the white space. A threshold T_i it is set for each feature. Initially, the average sum of each feature is calculated. After that, compute the difference and check with T_i. If the value is above or matches with T_i Then, it is detected as a relevant feature RLF_i. During the creation of integral image $Im(a, b)$, it identifies the sum of pixel values in an image or rectangular part of a painting by (2) in which $I(a', b')$ is the intensity of the original image.

$$Im(a,b) = \sum_{a' \le a, b' \le b} i\left(a', b'\right). \tag{2}$$

The integral image can be calculated in a single pass using the following equations, in which $csum(a, b)$ is the cumulative row sum

(1) Extract the video frames ($f_1 \ldots f_n$)
(2) Find the histogram difference between two adjacent frames (f_j, f_{j+1})
(3) Calculate M and SD of absolute difference
(4) Compute threshold, TH
(5) Compare the difference (d) with TH
    if the $d >$ TH
      select it as a keyframe (f_k)
    Else
      go to step no. 2.
(6) Continue the process till the end of the video

Algorithm 1: Histogram difference for keyframe selection.

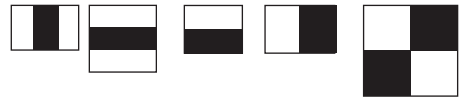

Figure 2: Haar cascade features. (a) Line features. (b) Edge features. (c) Rectangle features.

$$csum(a,b) = s(a,b-1) + i(a,b), \tag{3}$$

$$Im(a,b) = Im(a-1,b) + s(a,b). \tag{4}$$

In the integral image, csum $(a, 1) = 0$ and $Im(1, b) = 0$. After generating the integral image, each feature can be calculated at a constant time.

(2) Kanade Lucas Tomasi Algorithm. Face detection requires tracking of faces on keyframes. Kanade Lucas Tomasi (KLT) is an effective feature-based face-tracking algorithm. It continuously tracks human faces in a strong frame extracted from videos. This method finds the parameters that reduce dissimilarity measurements between feature points related to the original translational model. For tracking the face, it finds the traceable feature points in the first keyframe and then follows the detected features in the succeeding keyframes based on computed displacement value.

Let us assume that initially, one of the corner points is (a, b). If (a, b) is displaced by some variable vector $(v_1, v_2, \ldots v_n)$ in next frame, the displaced corner point (dc) can be calculated by equation

$$dc = (a,b) + v_i. \tag{5}$$

The coordinates of the new point will be $a' = a + v1$ and $b' = b + v2$. It uses warp function $W(a; d) = (a + v1; a + v2)$ to calculate the coordinates. The alignment is calculated by the following equation:

$$\sum_a [I(W(a;d)) - T(a)]\,[I(W(a;d)) - T(a)]^{\hat{2}}, \tag{6}$$

where d is the displacement parameter. Assume an initial estimate of d as a known parameter and find Δd using the following equation:

$$\sum_a [I(W(a;d+\Delta d)) - T(a)]^{\hat{2}}. \tag{7}$$

The displacement Δd is calculated by finding the Taylor series and then differentiating it concerning Δd using equation (8), in which H is called the Hessian matrix.

$$\Delta d = H^{-1} \sum_x \nabla I \left[\nabla I \frac{\partial W}{\partial d}\right]^T . [T(a) - I(W(a;d))]. \tag{8}$$

3.2.2. Emotion Recognition by CNN. This work uses CNN to achieve high precision in emotion recognition. There are two primary functions of CNN; Feature Extraction and classification. CNN has multiple layers in which each layer performs a specific transformation function. The goal of CNN is to reduce the images so that it would be easier to process without losing valuable features for accurate prediction.

The first layer to extract features from the input image is convolutional. Convolution can perform edge detection, blur, and sharpening operations by applying filters to an embodiment. When the image is too large, the pooling layer function is used to reduce the number of parameters. Spatial pooling, like average pooling, reduces the size of each map while retaining important information. The fully connected layer has flattened the matrix into a vector and feeds it into a neural network-like fully connected layer.

During emotion recognition, the convolutional layer recognises features in pixels. Then, pooling layers are responsible for making these features more abstract. Finally, the fully-connected layer is accountable for the classifications of emotions. The first layer is convolutional with a kernel size of 5×5 pixels and 16 output channels. The second layer is a max pooling layer with a 2×2 kernel size. In

the same manner, nine convolution layers are used in this work. The following three layers are fully connected neural layers with 100, 50, and 5 neurons in each layer.

Image pixels are directly used as input to standard feed-forward neural networks for emotion recognition in the convolution layer. In emotion classification, one or more 2D matrices are fed into the convolutional layer, and multiple 2D matrices are generated as output using equation

$$O_k = af\sum_{j=1}^{n} Im_j * Kn_{j,k} + C_k. \quad (9)$$

Each input matrix Im_j is convoluted with a corresponding kernel matrix $Kn_{j,k}$. Then the sum of all convoluted matrices is computed, and a bias value C_k is added to each element of the resulting matrix. Finally, a nonlinear activation function af is applied to each aspect of the previous matrix to produce one output matrix O_k. Each set of kernel matrices represents a local feature extractor that extracts regional features from the input matrices. The learning procedure aims to find groups of kernel matrices $Kn_{j,k}$ That extracts good discriminative features to be used for emotion recognition. Backpropagation, a neural network connection weight optimisation algorithm, can train kernel matrices and biases as shared neuron connection weights.

The pooling layer is used to reduce feature dimension. It reduces the number of output neurons in the convolutional layer, and pooling algorithms should be used to combine the convolution output matrices' neighbouring elements. Max pooling is used to reduce dimensionality. The Max pooling layer with a 2×2 kernel size chooses the highest value from four adjacent input matrix elements to generate one component of the output matrix. During the error backpropagation process, the gradient signal must be only routed back to the neurons that contribute to the pooling output. In our CNN model, the ReLU activation function $f(x) = \max(0, x)$ is used in the convolutional layer, which significantly improves both learning speed and emotion recognition performance in CNN.

Batch learning is used to accelerate and improve learning speed and accuracy. Instead of updating the connection weights after each backpropagation, we process 128 input samples in a batch and update the entire set with a single update. To further speed up the learning, momentum weight combined with weight decay is applied. The weight $\Delta\omega_i$ is updated by utilising equation

$$\Delta\omega_i(t+1) = \omega_i(t) - \eta\frac{\delta E}{\delta\omega_i} + \alpha\Delta\omega_i(t) - \lambda\eta\omega_i. \quad (10)$$

The $\omega_i(t) - \eta\delta E/\delta\omega_i$ part is the backpropagation, where $\omega_i(t)$ is the current weight vector. $\delta E/\delta\omega_i$ is the error gradient concerning the weight vector, and η is the learning rate. The $\alpha\Delta\omega_i(t)$ is the momentum part, where α is the momentum rate. The momentum weight update will speed up learning. The $\lambda\eta\omega_i$ is the weight decay part, where λ is the weight delay rate. It slightly reduces the weight vector towards zero in each learning iteration, which helps stabilise the learning process. The working process of CNN during emotion recognition is shown in Figure 3. In this, it detects the type of emotion after analysing the features by CNN from the face image.

3.3. Method 2: $CNN_{Haar\text{-}SVM}$. Method 2, developed inside the ensemble framework, consists of two main steps such as (1) face detection and (2) emotion recognition.

3.3.1. Face Detection. Face recognition is a two-step process. Initially, HOG-based normalised face features are extracted. After generating normalised feature vectors, all the features are given as the input to the SVM classifier for face recognition.

(1) Histogram of Oriented Gradients (HOG) Face Feature Extraction. The HOG feature descriptor is used for emphasising face structures or shapes. The magnitude and gradient angle are used to compute the features in this feature descriptor. It outperforms other edge descriptors. It generates histograms for the areas of the face image based on the magnitude and direction of the gradient. Using HOG, each face image is first divided into small square cells. It then computes a histogram of oriented gradients for each cell. The result is then normalised using a block-wise pattern, and a description for each cell is output. Figure 4 depicts the flow of HOG feature computation from the input face image.

Seven significant steps are used to compute HOG features from the input face image. It is explained below.

Step 1: *Input Face image and Perform Preprocessing*. Consider the input face image. Initially, the images are preprocessed to reduce the width-to-height ratio to 1 : 2. Most preferably, the input face image size should be resized to a size of 64×128. Then the resized images are considered for further processing for the optimal extraction of features.

Step 2: *Compute Gradients*. Combining the image's magnitude and angle yields the gradient such as gradient_x and gradient_y. It is calculated for each pixel value in the input image. The gradient_x is computed by equation (11) and gradient_y is computed by equation (12) in which R and C represent the row and column of each image matrix A. After the gradient value computation of each pixel in the image, the magnitude and angle values are computed by equations (13) and (14).

$$\text{gradient}_x(R,C) = A(R,C+1) - A(R,C-1), \quad (11)$$

$$\text{gradient}_y(R,C) = A(R-1,C) - A(R+1,C), \quad (12)$$

$$\text{Mag}(\mu) = \sqrt{\text{gradient}_x^2 + \text{gradient}_y^2}, \quad (13)$$

$$\text{Ang}(\theta) = \left|\tan^{-1}\left(\frac{\text{gradient}_y}{\text{gradient}_x}\right)\right|. \quad (14)$$

Step 3: Dividing gradient image into 8×8 cells

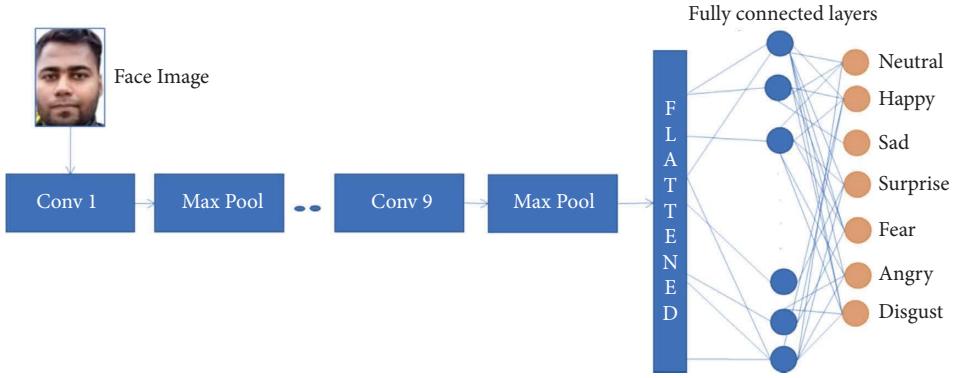

Figure 3: The process of CNN during emotion recognition.

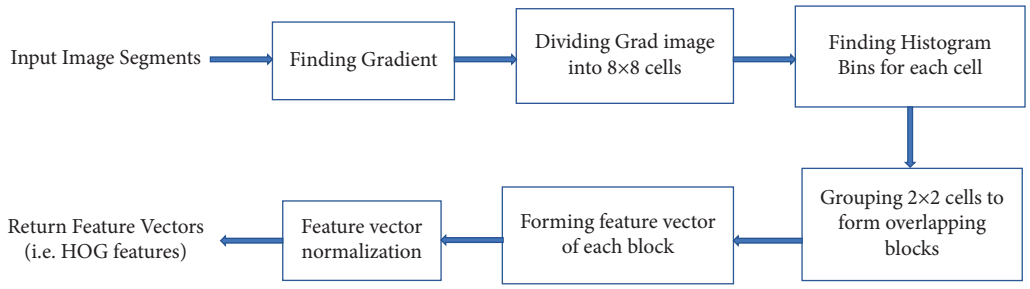

Figure 4: The flow of computation of HOG features from the face image.

Initially, the image is divided into 8×8 cells. After that, the feature descriptors of the histogram of oriented gradients are computed for each 8×8 cell in the image

Step 4: Identify the Histogram Bins for each 8×8 cell

In this step, a 9-point histogram is computed for each 8×8 cell block with 20° angle range. For that, assume Total number of histogram bins = 9 that is ranging from angles 0° to 180°.

Step 5: Construct overlapping blocks by grouping 2×2 cells.

Four blocks from the nine-point histogram matrix are combined to create a new block after the histogram computation for all of the blocks is complete (2×2). This clubbing is carried out in an overlapping fashion with an 8-pixel stride

Step 6: Generate a feature vector for each cell block

Generate 36 feature vectors by concatenating constructed 9-point histograms for each cell block

Step 7: Perform Feature vector normalisation

The gradients of the image are sensitive to the overall lighting. But the generated HOG features for the image's 8×8 cells offered noticeably brighter than the other portions. This maintains visibility in the image. Normalising the gradients by taking 16×16 blocks may lead to fluctuation in lighting. Therefore, a 16×16 block will be formed by joining four 8×8 cells. In step 4, a histogram comprises a 9×1 matrix for each 8×8 cell. Therefore, a single 36×1 matrix or four 9×1 matrices are there while creating a 16×16 block. Mathematically, for given vectors V are 36 rows there, which is represented in equation

$$V = [r_1, r_2, \ldots, r_{36}]. \tag{15}$$

For normalising, the matrix is constructed based on equation (16). It divides each value by the square root of the sum of squares of the values (k) as per equation (17).

$$\text{Vector}_{\text{normalized}} = \left(\frac{r_1}{k}, \frac{r_2}{k}, \ldots, \frac{r_{36}}{k}\right), \tag{16}$$

$$k = \sqrt{(r_1)^2 + (r_2)^2 + \ldots + (r_{36})^2}. \tag{17}$$

The process of creating HOG features for the image is completed. For the 16×16 blocks of the image, features for the complete image are built by integrating the features. Now the generated features are given as the input to the SVM classifier for detecting faces.

(2) Support Vector Machine (SVM) for Face Recognition. A classic two-class recognition problem is solved using a support vector machine (SVM). It transforms the data using a kernel trick and then finds an optimal boundary between the possible outputs based on the transformations. This work uses SVM for face recognition by modifying the interpretation of an SVM classifier's output and devising a representation of facial images concordant with a two-class problem. It selects the decision boundary that minimises the distance between the classes' closest data points. The maximum margin classifier or maximum margin hyperplane is the decision boundary generated by SVMs.

Let P_j be the HOG features and Q_j be the class labels of training data. The images may be labelled as face images as +1 and nonface images as −1. The SVM algorithm considers the input (P_j, Q_j) during training. After that, it finds the optimal decision surface with T_n the number of support vectors. Then, the linear surface can be calculated by equation (8), in which $\propto_j$ is the coefficient weight, Q_j is the class label of the support vector SV_j and the weighted summation (w). The computation of w in equation (18) is calculated using equation (19).

$$w.A + b_i = 0, \tag{18}$$

$$w = \sum_{j=1}^{T_n} \propto_j Q_j SV_j. \tag{19}$$

$A \in F^n$ is a facial image representation vector, where F^n is referred to as face space. Face space can be another feature space or the vectorised original pixel values. The equation's function calculates the SVM classifier function (18).

Feed a training set with two classes—one of the nonfacial photos and the other of facial images to build a classifier for the image "A." An SVM algorithm creates a linear decision surface to determine if the face image "A" is a face or not. The following equation states that image "A" is a face image if the input picture A meets the requirement.

$$w.A + b_i > 0. \tag{20}$$

If the input image A satisfies the condition given in the following equation, then image "A" is a nonface image.

$$w.A + b_i \leq 0. \tag{21}$$

After detecting the frames with faces, all the face images are passed to the CNN for emotion recognition.

3.3.2. Emotion Recognition by CNN. The emotion recognition is carried out by convolution with different filter sizes and pooling layers of CNN. The flow of work detecting emotion from the face in CNN is mentioned in Figure 5. The working process of CNN is already discussed in Section 3.2.2.

3.4. Method 3: CNN_{PATCH}. In this method, the recognition of the face is done by a template-based method, and the emotion is detected by analysing the patches from the face by patch-based CNN.

3.4.1. Template-Based Face Detection. Using the correlation between the templates and the input photos, template matching locates faces using predefined face templates. For instance: A human face can initially be broken down into its eyes, face contour, nose, and mouth. The edge detection technique can then create an edge-rich face model. It is a method of looking for and finding a template within a bigger picture. It determines whether the input face and template images are similar (training images). The presence of full-face features can then be ascertained by analysing the correlation between the input face photos and the standard patterns stored in the full-face parts. It looked at the input photos at various scales to achieve the shape and scale invariance. Algorithm 2 explains the process of template-based face detection in keyframes.

Let $K_f(x, y)$, $t(x, y)$ denotes the keyframe and template image, respectively. During the matching of t and K_f, the correlation value (cv) is calculated using equation (22). Then, normalise the correlation value using equation (23). The correlation threshold (T) is computed by adding the mean with an arbitrary number of standard deviations. After that, compare cv and T. If the value of cv exceeds T, then that segment is marked as the face.

$$R(x, y) = \sum_{x', y'} \left(\text{tem}\left(x', y'\right) - kf\left(x + x', y + y'\right) \right)^2, \tag{22}$$

$$R(x, y) = \frac{\sum_{x', y'} \left(\text{tem}\left(x', y'\right) - kf\left(x + x', y + y'\right) \right)^2}{\sqrt{\sum_{x', y'} \text{tem}\left(x', y'\right)^2 . \sum_{x', y'} kf\left(x + x', y + y'\right)^2}}. \tag{23}$$

3.4.2. Patch-Based CNN for Emotion Recognition. For the effective handling of occlusion in face images, A Patch-Gated CNN (PG-CNN) [59] is used in this work. The primary reason to use patches instead of the entire face is to increase the number of training samples for effective and optimal feature-based CNN learning. The second reason is that traditional CNN needs to resize faces when using full-face images as input. It significantly reduces the discriminative information. Using local patches maintains the native resolution of original face images, which increases discriminative ability. The framework of patch-based CNN during emotion recognition is shown in Figure 5.

This approach uses facial landmarks for region decomposition to generate the input image patches. In this, an end-to-end trainable Patch-Gated Convolution Neutral Network (PG-CNN) [59, 60] automatically percepts the occluded region of the face and focuses on the most discriminative unoccluded areas. According to the locations of facial landmarks, PG-CNN divides an intermediate feature map into 24 patches to identify potential regions of interest on the face. After that, a suggested Patch-Gated Unit in PG-CNN is computed from the patch itself and reweights each patch according to relevance. The working of the Patch-Gated Unit, followed by CNN during partial facial occlusion is represented in Figure 6. The algorithm for occlusion detection from the patched image is described in Algorithm 3.

The keyframes from the keyframe extraction phase will be considered as the input image of the network. The network receives the information and displays it as feature maps. The feature maps of the entire face are then divided

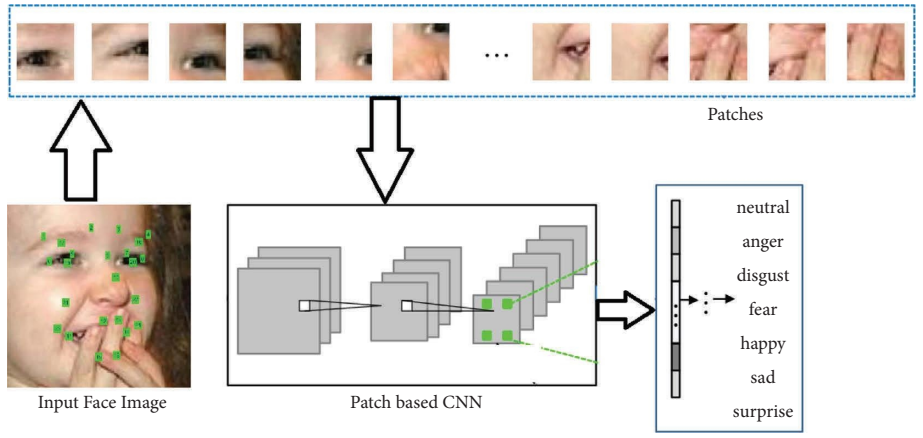

Figure 5: Framework for patch-based CNN during emotion recognition.

Input: the keyframe(K_f) and template image (t)
Output: The keyframe with face
Read the keyframes K_f
Read the template image t
Apply template matching to detect face image
  Slide the t over K_f
  Compare t and K_f and find the correlation value (cv) using equation (1)
  Normalise the correlation value using (2)
  Compute the correlation threshold (T)
    Add mean with an arbitrary number of standard deviation
  Compare cv and T
  If the cv > T the
    The segment is marked as a face
  Else
    The segment is marked as nonface
  End If

Algorithm 2:Template-based face detection in keyframes.

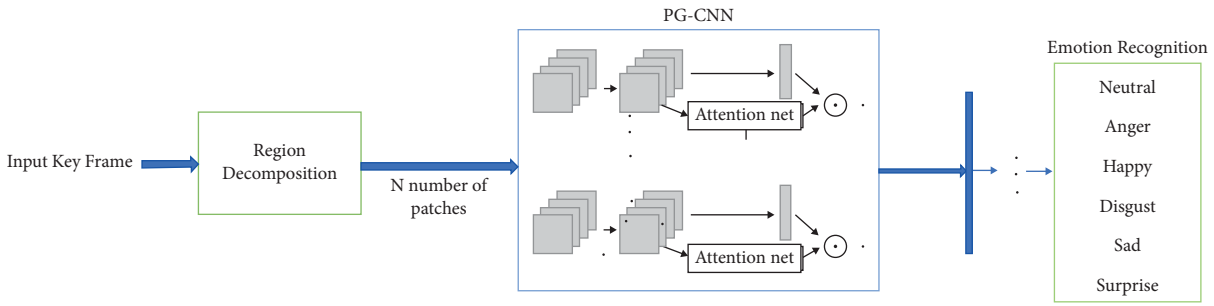

Figure 6: The working CNN with of Patch-Gated Unit during partial facial occlusion.

into 24 subfeature maps for 24 local patches via PG-CNN. A Patch-Gated Unit encodes each local patch as a weighted vector of local features (PG-Unit). By taking into account each patch's obstructed-ness, or how much of the patch is blocked, PG-Unit determines the weight of each patch by an Attention Net. The occluded face is finally represented by concatenating the weighted local features. The face is assigned to one of the emotional categories through three completely connected layers. The soft-max loss function value is minimised to optimise the PG-CNN. For handling the occlusion issue, PG-CNN used two key schemes Region decomposition and Occlusion perception.

(1) Region Decomposition. Extract the patches based on the locations of the facial landmarks for each individual to identify the usual facial regions associated with expression.

Input: Keyframes
Output: Representation of occluded face
Input the extracted keyframe KF_i as a face image
Generate a feature map (FM) from each keyframe
  Return 24 local patches (P_1, $P_2 \ldots P_{24}$)
  For each local patch
    Decomposes the feature map into 24 subfeature-maps ($SFM_1 \ldots SFM_{24}$)
    Encode a weighted vector (wv) of local feature (lf) by a PG-Unit
    PG-Unit computes the weight by an attention net based on its obstructed-ness
    Concatenate the weighted local features
    Return the representation of the occluded face.
  End For

Algorithm 3: Occlusion detection from patched image.

Then, depending on the n points discovered, it finds n facial landmark points. The informative facial area, consisting of two eyes, a nose, a mouth, a cheek, and eyebrows, is then covered by a new computation of m points. The selected patches are then defined as the regions by treating each of the m points as the centre. Following are the procedures used in this study to compute region decomposition for creating feature maps.

Step 1: Detects 68 facial landmark points

Step 2: Select or re-compute 24 points. It must hold informative regions of the face such as eyes, nose, mouth, cheek, and eyebrow.

Step 3: Consider each of the 24 points as the centre and define 24 patches (P_1, $P_2 \ldots P_{24}$)

Step 4: Based on the feature maps of size $512 \times 28 \times 28$ and 24 local region centres, 24 local regions or Patches ($512 \times 6 \times 6$) are obtained.

(2) Occlusion Perception with PG-Unit. The PG-Unit embedded in the PG-CNN automatically percept the blocked facial patch and pay attention mainly to the unblocked and informative patches. In each patch-specific PG-Unit, the cropped local feature maps are fed to two convolution layers without decreasing the spatial resolution. This is more effective in preserving more information when learning region-specific patterns. The last $512 \times \text{six} \times \text{six}$ feature maps are processed in two branches. The first branch encodes the input feature maps as the vector-shaped local feature. The second branch consists of an attention net that estimates a scaler weight to denote the importance of the local patch. The computed weight then weights the local feature.

Each local patch is encoded as a weighted vector of local features by a Patch-Gated Unit (PG-Unit). PG-Unit computes the weight of each patch by an attention net, considering its obstructed-ness (to what extent the patch is occluded). Finally, the weighted local features are concatenated and serve as a representation of the occluded face. Three fully connected layers are followed to assign the face to one of the emotional categories. PG-CNN is optimised by minimising the soft-max loss. The steps are used to identify occlusion with PG-unit and further emotion recognition.

Step 1: Input the feature map SFM_i of patch $P_i toPG\,\text{unit}_i$

Step 2: $PG\,\text{unit}_i$ calculated the weighted feature φ_i as per equation (24), Importance or unobstructed-nessα_i based on equation (25) and feature vector ψ_i by equation (26), in which $P_i = \varphi(P_i)$ is the last feature map ahead of the two branches, (Θ) denotes production, and (.) denotes the attention net operations: pooling, convolution, inner productions, and a sigmoid activation.

$$\varphi_i = I_i(\widetilde{P}_i)\,\Theta\,\psi(\widetilde{P}_i), \tag{24}$$

$$\alpha_i = I_i(\widetilde{P}_i), \tag{25}$$

$$\psi_i = \psi(\widetilde{P}_i).\,\psi(\widetilde{P}_i). \tag{26}$$

Step 3: The sigmoid activation forces the output αi ranges in [0, 1], where 1 indicates the most salient unobstructed patch and 0 indicates the completely blocked patch.

3.5. Ensemble Max Rule Method for Emotion Recognition: $CNN_{ENSEMBLE}$. The idea is to create a model that ensembles the inherent emotional information within the video frames. Two basic approaches are proposed to achieve this purpose: (1) maximum emotion ensembles (MSE) (2) late multiple feature fusion (LMFF).

In maximum emotion ensembles, three models are explored: (1) Max. Emotions (2) Max. Emotion Intensity (3) Max. Emotion Sustenance [19, 61]. All three models work on extracted keyframes of the given video. Maximum Emotions count on the maximum probability related to each emotion across the separated keyframes and depicts it as the final emotion. The maximum Emotion Intensity model measures the intensity of emotions for every keyframe and recommends the most intensified emotion. The maximum emotion sustenance model is more accurate than the above two models [19, 61]. This model measures the emotion in every keyframe and looks globally at the emotion that had occurred repeatedly for the more extended sequence of keyframes.

Late multiple-feature fusion operates independently in 3 ways. The first method ($Method_1$) performs face detection and tracking from input video and then uses CNN for emotion detection over the face boundary box. The other two methods perform video emotion recognition via image-based approaches. The input video sequence is split into multiple keyframes. The keyframes are then fed to the face recognition module, which identifies the sample. The corresponding face set from the database is identified, features extracted from the input face and matched with the trained sample features using SVM, and then fed to CNN for emotion detection by $Method_2$. The last approach ($Method_3$) uses patch-based ACNN for occlusion-aware emotion recognition. All three emotion recommendations are later fused in the ensembling setting to recommend the emotion at the output. The pseudocode for ensemble classification is stated in Algorithm 4.

4. Experimental Results and Discussion

This section discussed the dataset used in this work and the evaluation results of emotion recognition with the ensemble method.

4.1. Video Emotion Dataset. The primary data set used in this work is the Extended Cohn-Kanade (CK+) dataset, which contains 593 video sequences from a total of 123 subjects. In this, 327 videos are labelled with anger, contempt, disgust, fear, happiness, sadness, and surprise. A detailed description of the dataset is tabulated in Table 3. The CK+ database is one of the most widely used facial expression classification databases. We have included some camera-recorded participants' facial expressions without disturbing their natural emotion outlay. The videos are recorded for 1–10 seconds. Each video contains an average of 10–15 keyframes. A total of 1830 videos were taken for the experiment and among which 80% were taken for training and 20% for testing.

4.2. Keyframe Extraction. A keyframe extraction approach [61] uses the histogram with deep learning to extract the pertinent keyframe from the video sequence. The keyframe extraction gets the highest recall and precision values for all the video sequences. In most cases, a metric's highest value is insufficient. The precision metric assesses a method's capacity to obtain the most accurate outcomes. A high accuracy number indicates more substantial keyframe relevance. However, a high-precision number can be obtained by choosing just a few keyframes from a video sequence. The keyframe extraction algorithm depends heavily on the accuracy and speed of both parameters. If the algorithm is slow, then the throughput of the system gets affected. It is also necessary that extracted keyframes are the relevant and accurate. Further, it will affect other processes, such as object detection, classification, and object description. The Precision in equation (27) and Recall in equation (28) are evaluated during keyframe extraction and tabulated in Table 4.

$$\text{Precision } (P) = \frac{\text{number of correctly detected keyframes}}{\text{number of keyframes}}, \tag{27}$$

$$\text{Recall } (R) = \frac{\text{Number of correctly detected keyframes}}{\text{number of detected keyframes, + number of undetected keyframes}}. \tag{28}$$

The Precision and Recall value achieved using the keyframe extraction method is high, so the model gives unique frames without replica. The result also calculates the CPU time (0.50) to extract the keyframes; it shows that the extraction speed is good.

4.3. Face Detection. Facial detection plays a significant role in facial identification and emotion recognition. The method of face detection in photographs is complicated due to the variability across human faces, including pose, expression, position and orientation, skin colour, glasses or facial hair, differences in camera gain, lighting conditions, and image resolution. This method's strength is to concentrate computational resources on the area of an image holding a face.

One of the computer technologies involved in image processing and computer vision is called object detection, and it deals with finding instances of objects like people, cars, buildings, and trees. Finding out if there is a face in the image is the primary goal of face detection algorithms. In this paper, we employ two face-detection techniques. Face detection allows us to gather the data required for emotion analysis.

4.3.1. Face Detection Using Haar Cascade and KLT Algorithm. The video may contain a single person or multiple people, and emotion can be identified for both occluded and nonoccluded faces. Initially, the face is detected through the Haar cascade algorithm and tracked using KLT tracking, which accurately tracks the detected face. The sample results of face detection using Haar and KLT are displayed in Figure 7 and tabulated in Table 5.

In Method 1, the keyframes with faces are identified and tracked by Haar cascading and the KLT algorithm. The model was accurately detecting the faces in the image. But partially occluded or side-angled faces are missing in the model. In our experiment, we got an accuracy of 92.6% for the model.

Data: Training Set
[Data = $\{(s_a, r_a)|s_a \in R, r_a \in \{\text{Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral}\}\}$]
BC: Number of base classifiers
SR: Ratio of samples that need replacement
λ : Parameter used to reduce the distance between training and synthetic data
a_i : i^{th} attribute
σ_i : standard deviation
r: normal distribution's sampling value $N(0, 1)$
Training Phase:
For $a = 1$: BC
  Copy the original dataset i.e. $\text{Data}_a \leftarrow$ Data
  Identify the number of training samples that need replacement, i.e., $TS = \text{round}(n \times SR)$
  For $b = 1$: TS
    Randomly pick "z" samples from Data_a
      If x is a majority class sample, then
        Generate a neighborhood of z based on $a_i'(b) = a_i(b) - \lambda r \sigma_i$ and replace z exists in Data_a
      Else if
      Check z is a minority sample, then compute $m = \text{Round}\ ((\text{Imbalanced}_{\text{ratio}} - 1)/(SR + 1))$
      Replace m neighbourhoods of z in Data_a
    End For
    Build base classifier BC_a from Data_a
End For
Classification Phase:
For a given z
  evaluate ensemble $\{\text{Method}_1, \text{Method}_2, \text{Method}_3\}$ to classify the sample z based on the majority voting strategy

Algorithm 4: Ensemble classification.

Table 3: Video emotion dataset description.

	Emotion categories						
	Neutral	Anger	Disgust	Fear	Happy	Sad	Surprise
Videos	190	240	210	280	310	270	230

Table 4: Recall, Precision, and CPU time for keyframe extraction.

Recall	Precision	CPU time (ms)
0.95	0.92	0.50

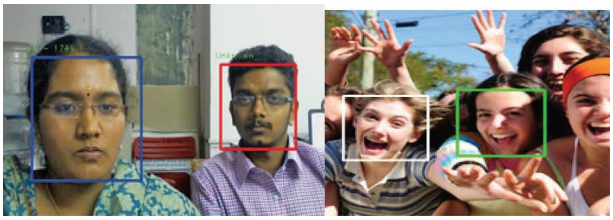

Figure 7: Sample result for face detection in method1.

Table 5: Sample result for face detection in method1 for two different keyframes.

Keyframe	The number of faces detected	Number of nonfaces detected	Number of faces not detected
1	2	0	0
2	2	0	2

4.3.2. Face Detection Using HOG and SVM. Feature extraction for facial emotion recognition is performed through HOG features, where the video is fed to keyframe extraction and the particular image has all the information and details about the face (i.e.) face directions, edges, intensities, and colour are extracted and saved in a separate file. This information is fed to the SVM classifier for accurate face detection. The sample results of face detection using HOG and SVM are displayed in Figure 8 and tabulated in Table 6.

The effectiveness of the face detection model is often evaluated based on Precision in equation (29), Recall in equation (30), and Accuracy by equation (31). The Precision and Recall are in Table 7, and Accuracy in Table 8 proves that more accurate face detection is possible when using the HOG-SVM model.

$$\text{Precision } (P) = \frac{\text{number of correctly detected faces}}{\text{number of detected faces}}, \tag{29}$$

$$\text{Recall } (R) = \frac{\text{Number of correctly detected faces}}{\text{number of detected faces, + number of undetected faces}}, \tag{30}$$

$$\text{Accuracy} = \frac{\text{number of correct face recognition}}{\text{Total number of keyframes used for face recognition}}. \tag{31}$$

HOG achieves face detection, and SVM offers more accurate results than Haar KLT because it detects faces with angle change or partially covered to some extent.

4.4. Emotion Recognition. After identifying the face, the emotions are detected using CNN and patch-based CNN. The performance of emotion recognition is discussed below.

4.4.1. Patch-Based CNN for Emotion Recognition. In this study, CNN is used as the base classifier by PG-CNN. The straightforward structure and unique item categorisation performance are the cause. Attach 24 PG-Units after selecting the first nine convolution layers as the feature map for region decomposition. The model was initialised using the pretrained model based on the ImageNet dataset. For each dataset, both the train and test corpus are mixed with occluded images with a ratio of 1 : 1. We adopt a batch-based stochastic gradient descent method to optimise the model. The base learning rate was set as 0.001 and was reduced by the polynomial policy with a gamma of 0.1. The momentum was set as 0.9, and the weight decay was set as 0.0005. The training of models was completed on a Titan-X GPU with 12 GB memory. During the training stage, we set the actual batch size as 128 and the maximum iterations as 50 K. It took about 1.5 days to finish optimising the model.

CNN disintegrates the feature maps as multiple subfeature maps. The region decomposition the feature maps are divided by CNN into many subfeature maps. The facial picture is aligned by fixing the 68 facial landmarks around the face, and the region is decomposed by splitting the facial landmark into 24 patches that span the entire informative area. Then patches are extracted based on the locations of the landmarks on each subject's face. The following procedure is used to choose the facial patches:

(i) Sixteen points are picked from the original 68 facial landmarks to cover eyebrows, eyes, nose, and mouth. The selected points are indexed as 19, 22, 23, 26, 39, 37, 44, 46, 28, 30, 49, 51, 53, 55, 59, and 57

(ii) In addition, 4 points are picked around the eyes and eyebrows, and then the midpoint of each point pair is computed as the delegation.

Based on the $512 \times 28 \times 28$ feature maps and the 24 local region centres, a total of 24 provincial regions are obtained, each with a size of $512 \times 6 \times 6$.

The inbuilt PG-CNN detects blocked face patches automatically and focuses mainly on unblocked and informative patches. The cropped local feature maps are given to two convolution layers, the attention layer and the encoding layer, in each patch-specific PG unit. Figure 9 depicts the regional features.

Table9 shows the performance of both nonoccluded and occluded images during emotion recognition. For both occluded and nonoccluded scenarios, the overall accuracy on seven facial expression categories is evaluated by performing a 10-fold evaluation.

A 10-fold test accuracy test has been performed on CK+, ISED Dataset with synthetic occlusions. The size of occlusion are 8×8, 16×16, and 24×24, represented by R8, R16, and R24, respectively. The full image size is 48×48. The input images (size 48×48) without occlusion have high accuracy of 97.02%. In the same image set, synthetic occlusion was applied on a different scale (S8, S16, S24). The accuracy of occluded images varies with the amount of occlusion. But Table 10 shows that the accuracy of the occluded images is also high.

4.4.2. Performance of CNN vs. Patch-Based CNN vs. Ensemble. Table 9 shows the different sets of experiments conducted for emotion detection from facial expressions. In the first method $CNN_{HOG\text{-}KLT,}$ face detection is performed by HOG and KLT methods, and detected CNN classifies frames. Similarly, in $CNN_{Haar\text{-}SVM,}$ Haar cascade and SVM techniques are used for face detection in extracted frames. Again CNN is

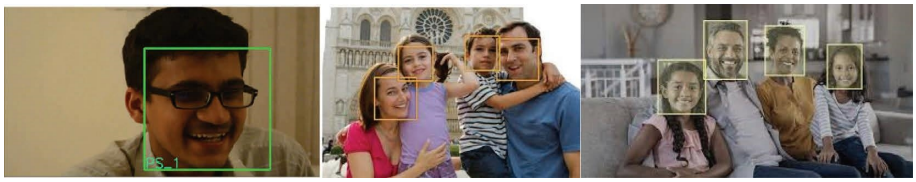

Figure 8: Sample result for face detection in method 2.

Table 6: Sample result for face detection in method 2.

Keyframe	The number of faces detected	Number of nonfaces detected	Number of faces not detected
1	1	0	0
2	4	0	0
3	4	0	0

Table 7: Precision and Recall for face detection using HOG-SVM and Haar KLT.

No. of images	*P*		*R*	
	HOG-SVM (%)	Haar KLT (%)	HOG-SVM (%)	Haar KLT (%)
700	98.13	87.71	98.13	88.71

Table 8: Accuracy for face detection using HOG-SVM and Haar KLT.

No. of images	Accuracy	
	HOG-SVM (%)	Haar KLT (%)
700	98.18	92.64

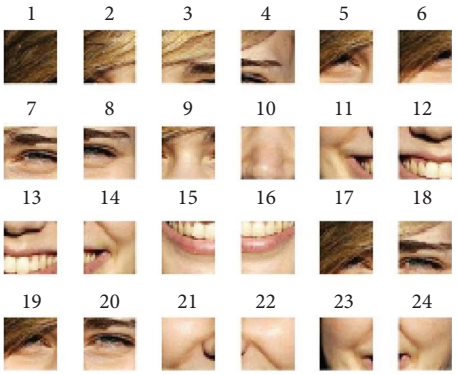

Figure 9: The sample patches generated from the input face image.

Table 9: Experimental methods and their description.

Methods	Description
$CNN_{HOG\text{-}KLT}$	Emotion recognition (ER) using CNN after HOG and KLT face detection method
$CNN_{Haar\text{-}SVM}$	ER using CNN after face detection by Haar and SVM
CNN_{PATCH}	ER using CNN after face detection by patch-based method
$CNN_{ENSEMBLE}$	ER by CNN ESEMBLE

Table 10: Accuracy of PG-CNN under different amounts of synthetic occlusions.

Images with different scales	PG-CNN accuracy (%)
Occlusion: S8	96.43
Occlusion: S16	95.15
Occlusion: S24	92.46
Images without occlusion	97.02

Table 11: Confusion matrix for $CNN_{HOG\text{-}KLT}$.

		Predicted label-$CNN_{HOG\text{-}KLT}$						
		Angry	Disgust	Fear	Happy	Sad	Surprise	Neutral
Actual label	Angry	92	2	1	1	1	1	1
	Disgust	2	90	1	2	2	2	2
	Fear	1	1	91	2	1	2	2
	Happy	1	1	1	93	2	1	1
	Sad	1	2	1	1	92	1	1
	Surprise	2	1	2	2	2	89	2
	Neutral	1	2	1	1	1	1	90

Table 12: Confusion matrix for $CNN_{Haar\text{-}SVM}$.

		Predicted label-$CNN_{Haar\text{-}SVM}$						
		Angry	Disgust	Fear	Happy	Sad	Surprise	Neutral
Actual label	Angry	91	1	2	2	2	1	1
	Disgust	1	91	1	2	2	2	1
	Fear	2	2	90	1	1	2	2
	Happy	1	1	1	93	2	1	1
	Sad	2	2	1	1	92	1	1
	Surprise	2	2	1	1	2	91	1
	Neutral	2	1	2	2	1	2	90

Table 13: Confusion matrix for CNN_{PATCH}.

		Predicted label-CNN_{PATCH}						
		Angry	Disgust	Fear	Happy	Sad	Surprise	Neutral
Actual label	Angry	91	2	1	2	1	1	2
	Disgust	1	91	1	2	2	2	1
	Fear	2	2	89	1	2	2	2
	Happy	1	2	1	92	1	2	1
	Sad	2	2	1	2	91	1	1
	Surprise	2	2	1	1	2	91	1
	Neutral	2	2	1	2	1	2	90

Table 14: Accuracy of maximum ensemble of different techniques.

Method	Accuracy							
	Angry	Disgust	Fear	Happy	Sad	Surprise	Neutral	Overall
$CNN_{HOG\text{-}KLT}$	92.93	90.91	89.00	92.86	91.20	91.09	91.75	91.39
$CNN_{Haar\text{-}SVM}$	89.24	89.31	90.91	91.12	91.08	91.92	93.72	91.04
CNN_{PATCH}	90.09	88.35	93.68	90.19	91.00	90.09	91.84	90.75
$CNN_{ENSEMBLE}$	**92.93**	**90.91**	**90.91**	**92.86**	**91.20**	**91.92**	**93.75**	**92.07**

Bold values show that the accuracy value of all emotions are more accurate in the CNN-ENSEMBLE model as compared with other models.

used for emotion recognition. In the last method, template matching is used for face recognition, and patch-based CNN is used for emotion classification.

Tables 11–13 present the confusion matrix generated for emotion classification by Method1 ($CNN_{HOG\text{-}KLT}$), Method2 ($CNN_{Haar\text{-}SVM}$), and Method3 (CNN_{PATCH}), respectively. All three methods correctly identified all seven emotions.

From the confusion matrix, the performance such as Precision in equation (32), Recall in equation (33), Accuracy in equation (34), and *F*-measure in equation (35) are evaluated and tabulated in Tables 14–17. From this, the

Table 15: The Precision of maximum ensemble of different techniques.

Method	Precision							
	Angry	Disgust	Fear	Happy	Sad	Surprise	Neutral	Overall
$CNN_{HOG\text{-}KLT}$	0.93	0.90	0.91	0.93	0.93	0.89	0.93	0.92
$CNN_{Haar\text{-}SVM}$	0.91	0.91	0.90	0.93	0.92	0.91	0.90	0.91
CNN_{PATCH}	0.91	0.91	0.89	0.92	0.91	0.91	0.90	0.91
$CNN_{ENSEMBLE}$	**0.93**	**0.91**	**0.91**	**0.93**	**0.93**	**0.91**	**0.93**	**0.92**

Bold values show that the CNN ENSEMBLE model has more precision values as compared with other models.

Table 16: Recall of maximum ensemble of different techniques.

Method	Recall							
	Angry	Disgust	Fear	Happy	Sad	Surprise	Neutral	Overall
$CNN_{HOG\text{-}KLT}$	0.93	0.91	0.93	0.91	0.91	0.92	0.91	0.92
$CNN_{Haar\text{-}SVM}$	0.90	0.91	0.92	0.91	0.90	0.91	0.93	0.91
CNN_{PATCH}	0.90	0.90	0.92	0.90	0.89	0.91	0.93	0.91
$CNN_{ENSEMBLE}$	**0.93**	**0.91**	**0.93**	**0.91**	**0.91**	**0.92**	**0.93**	**0.92**

Bold values show that the CNN ENSEMBLE model has more recall values as compared with other models.

Table 17: *F*-measure of maximum ensemble of different techniques.

Method	*F*-measure							
	Angry	Disgust	Fear	Happy	Sad	Surprise	Neutral	Overall
$CNN_{HOG\text{-}KLT}$	0.93	0.90	0.92	0.92	0.92	0.90	0.92	0.92
$CNN_{Haar\text{-}SVM}$	0.91	0.91	0.91	0.92	0.91	0.91	0.91	0.91
CNN_{PATCH}	0.91	0.91	0.90	0.91	0.90	0.91	0.91	0.91
$CNN_{ENSEMBLE}$	**0.93**	**0.91**	**0.92**	**0.92**	**0.92**	**0.91**	**0.92**	**0.92**

Bold values show that the CNN ENSEMBLE model has more *F*-measure values as compared with other models.

Table 18: Comparison with existing topic embedding with sentiment classification methods.

Year & reference	Dataset used	Emotion recognition	Accuracy (%)
2018 [8]	JAFFE	NN	87.53
2019 [4]	RAFDB	Second-order pooling	RAFDB: 89
	SFEW	CNN	SFEW: 60
2020 [45]	SAVEE	CNN	90
2021 [34]	FER2013	VGG	73.28
2022 [39]	Self, kaggle	VGG16	82
Our work	Extended CK+	ENSEMBLE CNN	92.07

$CNN_{ENSEMBLE}$ method is more accurate during emotion recognition.

$$\text{Precision } (P) = \frac{\text{number of correctly detected emotions}}{\text{number of images used for emotional recognition}}, \tag{32}$$

$$\text{Recall } (R) = \frac{\text{Number of correctly detected emotions}}{\text{number of detected emotions, + number of incorrect emotions}}, \tag{33}$$

$$\text{Accuracy} = \frac{\text{number of correct emotion rescognition}}{\text{Total number of images used for face recognition}}, \tag{34}$$

$$F - \text{measure} = \frac{2 * P * R}{P + R}. \tag{35}$$

4.5. Comparison with Existing Emotion Classification Methods. The comparison of existing emotion classification methods with the proposed model tabulated in Table 18. Based on the comparison, the proposed ensemble CNN ($\mathbf{CNN_{ENSEMBLE}}$) is more suitable for identifying emotion class.

5. Conclusion

An ensemble of CNN methods performs the robust emotion recognition of faces using multiple facial features. This proposed $\mathbf{CNN_{ENSEMBLE}}$ approach is suitable for a single person and multiple persons in the video. Despite partial occlusions, the proposed work responds much better than the previous approaches using CNN. All the faces with emotions within keyframes are initially detected using $CNN_{HOG\text{-}KLT}$, $CNN_{Haar\text{-}SVM,}$ and CNN_{PATCH} methods. After that, the $\mathbf{CNN_{ENSEMBLE}}$ method ensemble the detected emotions by the Max rule and achieved the maximum accuracy of 92.07%. In addition, other performance measures such as Precision, Recall, and *F*-Measure also proved that ensemble increases the emotion recognition rates. This system can detect emotions during occlusion.

The emotion recognition system has to be further improved to handle more partial and complete occlusions. In addition to this, there is a plan to consider contextual information along with facial images to recognise human emotions. Therefore, the extracted features from facial and context regions surrounding that person can be fused to make more labels and classify the emotions in the future.

Data Availability

The emotion datasets used to support the findings of the study are available at https://www.kaggle.com/code/shawon10/ck-facial-expression-detection/data and https://sites.Google.com/site/iseddatabase/download.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Authors' Contributions

All authors contributed equally to this work.

References

[1] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: audio, visual, and spontaneous expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, 2009.

[2] P. Sreeja and G. S. Mahalakshmi, "Emotion recognition from poems by maximum posterior probability," vol. 14 CIC 2016 special issue international journal of computer science and information security (IJCSIS)," in *Proceedings of the International Conference on Advances in Computational Intelligence and Communication (CIC 2016) Pondicherry Engineering College*, pp. 36–43, Puducherry, India, October 2016.

[3] J. Cai, Z. Meng, S. Ahmed, Z. Li, J. O'Reilly, and Y. Tong, "Island loss for learning discriminative features in facial expression recognition," in *Proceedings of the 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pp. 302–309, IEEE, Xi'an, China, May 2018.

[4] X. Tong and S. Sun, "Data augmentation and second-order pooling for facial expression recognition," *IEEE Access*, vol. 7, 2019.

[5] I. Kotsia, I. Buciu, and I. Pitas, "An analysis of facial expression recognition under partial facial image occlusion," *Image and Vision Computing*, vol. 26, no. 7, pp. 1052–1067, 2008.

[6] N. Alsrehin and A. &Mu'tasem, "Face recognition techniques using statistical and artificial neural network: a comparative study," in *Proceedings of the 2020 3rd International Conference on Information and Computer Technologies (ICICT)*, pp. 154–159, IEEE, San Jose, CA, USA, March 2020.

[7] Z. Guo, L. Zhang, and D. Zhang, "A completed modeling of local binary pattern operator for texture classification," *IEEE Transactions on Image Processing*, vol. 19, no. 6, pp. 1657–1663, 2010.

[8] J. Pedro, V. Soto, R. Queiroz Feitosa, V. H. Ayma Quirita, and P. Nigri Happ, "Single sample face recognition from video via stacked supervised auto-encoder," in *Proceedings of the 29th IEEE Conference on Graphics, Patterns and Images*, Sao Paulo, Brazil, October 2016.

[9] D. Huang, C. Shan, M. Ardabilian, Y. Wang, and L. Chen, "Local binary patterns and its application to facial image analysis: a survey," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 41, no. 6, pp. 765–781, 2011.

[10] P. Sharma, "Student engagement detection using emotion analysis, eye tracking and head movement with machine learning," in *Proceedings of the Technology and Innovation in Learning, Teaching and Education: Third International Conference, TECH-EDU 2022*, Lisbon, Portugal, September, 2022.

[11] M. Murugappan and A. Mutawa, "Facial geometric feature extraction based emotional expression classification using machine learning algorithms," *PLoS One*, vol. 16, no. 2, Article ID e0247131, 2021.

[12] E. R. Kimonis, B. Le, G. E. Fleming et al., "Facial reactions to emotional films in young children with conduct problems and varying levels of callous unemotional traits," *Journal of Child Psychology and Psychiatry*, vol. 64, no. 3, pp. 357–366, 2023.

[13] M. Sajjad, S. Zahir, A. Ullah, Z. Akhtar, and K. Muhammad, "Human behavior understanding in big multimedia data using CNN based facial expression recognition," *Mobile Networks and Applications*, vol. 25, no. 4, pp. 1611–1621, 2019.

[14] H. Joseph and B. K. Rajan, "Real time drowsiness detection using Viola jones & KLT," in *Proceedings of the 2020 International Conference on Smart Electronics and Communication (ICOSEC)*, IEEE, Trichy, India, September 2020.

[15] T. H. Obaida, "Real-time face detection in digital video-based on Viola-Jones supported by convolutional neural networks," *International Journal of Electrical and Computer Engineering*, vol. 12, no. 3, pp. 2088–8708, 2022.

[16] S. Happy and A. Routray, "Automatic facial expression recognition using features of salient facial patches," *IEEE transactions on Affective Computing*, vol. 6, no. 1, pp. 1–12, 2015.

[17] M. Ahmadi, W. Ouarda, and A. M. Alimi, "Efficient and fast objects detection technique for intelligent video surveillance

using transfer learning and fine-tuning," *Arabian Journal for Science and Engineering*, vol. 45, no. 3, pp. 1421–1433, 2020.

[18] S. Engoor, S. SendhilKumar, C. H. Sharon, and G. S. Mahalakshmi, "Occlusion-aware dynamic human emotion recognition using landmark detection," in *Proceedings of the 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, pp. 795–799, IEEE, Coimbatore, India, March 2020.

[19] Y. Li, J. Zeng, S. Shan, and X. Chen, "Occlusion aware facial expression recognition using cnn with attention mechanism," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2439–2450, 2019.

[20] Y. Li, J. Zeng, S. Shan, and X. Chen, "Patch-Gated CNN for occlusion-aware facial expression recognition," in *Proceedings of the 2018 24th International Conference on Pattern Recognition (ICPR)*, pp. 2209–2214, IEEE, Beijing, China, August 2018.

[21] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," in *Proceedings of the 2005 IEEE international conference on multimedia and Expo*, p. 5, IEEE, Amsterdam, Netherlands, July 2005.

[22] 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 the 2010 ieee computer society conference on computer vision and pattern recognition-workshops*, pp. 94–101, IEEE, San Francisco, CA, USA, June 2010.

[23] S. Koelstra, C. Muhl, M. Soleymani et al., "Deap: a database for emotion analysis; using physiological signals," *IEEE transactions on affective computing*, vol. 3, no. 1, pp. 18–31, 2012.

[24] I. Sneddon, M. McRorie, G. McKeown, and J. Hanratty, "The belfast induced natural emotion database," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 32–41, 2012.

[25] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn, "Disfa: a spontaneous facial action intensity database," *IEEE Transactions on Affective Computing*, vol. 4, no. 2, pp. 151–160, 2013.

[26] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the recola multimodal corpus of remote collaborative and affective interactions," in *Proceedings of the 2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, pp. 1–8, IEEE, Shanghai, China, April 2013.

[27] X. Zhang, L. Yin, J. F. Cohn et al., "Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database," *Image and Vision Computing*, vol. 32, no. 10, pp. 692–706, 2014.

[28] 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.

[29] S. Zafeiriou, D. Kollias, M. A. Nicolaou, A. Papaioannou, G. Zhao, and I. Kotsia, "Aff-wild: valence and arousal'in-the-wild'challenge," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 34–41, Honolulu, HI, USA, July 2017.

[30] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): a dynamic, multimodal set of facial and vocal expressions in north american English," *PLoS One*, vol. 13, no. 5, Article ID e0196391, 2018.

[31] W. Li, Y. Cui, Y. Ma et al., "A spontaneous driver emotion facial expression (DEFE) dataset for intelligent vehicles," 2020, https://arxiv.org/abs/2005.08626.

[32] M. T. B. Iqbal, M. Abdullah-Al-Wadud, B. Ryu, F. Makhmudkhujaev, and O. Chae, "Facial expression recognition with neighborhood-aware edge directional pattern (NEDP)," *IEEE Transactions on Affective Computing*, vol. 11, no. 1, pp. 125–137, 2020.

[33] I. Tautkute, T. Trzcinski, and A. Bielski, "I know how you feel: emotion recognition with facial landmarks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1878–1880, Honolulu, HI, USA, June 2018.

[34] Y. Khaireddin and Z. Chen, "Facial emotion recognition: state of the art performance on FER2013," 2021, https://arxiv.org/abs/2105.03588.

[35] D. Kamińska, K. Aktas, D. Rizhinashvili et al., "Two-stage recognition and beyond for compound facial emotion recognition," *Electronics*, vol. 10, p. 2847, 2021.

[36] A. H. Reddy, K. Kolli, and Y. L. Kiran, "Deep cross feature adaptive network for facial emotion classification," *Signal, Image and Video Processing*, vol. 16, no. 2, pp. 369–376, 2022.

[37] N. Siddiqui, "A robust framework for deep learning approaches to facial emotion recognition and evaluation," in *Proceedings of the 2022 Asia Conference on Algorithms, Computing and Machine Learning (CACML)*, IEEE, Hangzhou, China, March 2022.

[38] A. Khattak, M. Z. Asghar, M. Ali, and U. Batool, "An efficient deep learning technique for facial emotion recognition," *Multimedia Tools and Applications*, vol. 81, no. 2, pp. 1649–1683, 2022.

[39] S. Dwijayanti, M. Iqbal, and B. Y. Suprapto, "Real-time implementation of face recognition and emotion recognition in a humanoid robot using a convolutional neural network," *IEEE Access*, vol. 10, no. 2022, pp. 89876–89886, 2022.

[40] G. Yolcu, I. Oztel, S. Kazan, C. Oz, and F. Bunyak, "Deep learning-based face analysis system for monitoring customer interest," *Journal of Ambient Intelligence and Humanized Computing*, vol. 11, no. 1, pp. 237–248, 2020.

[41] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," in *Proceedings of the 2016 IEEE Winter conference on applications of computer vision (WACV)*, pp. 1–10, IEEE, Lake Placid, NY, USA, March 2016.

[42] H. Li and H. Xu, "Deep reinforcement learning for robust emotional classification in facial expression recognition," *Knowledge-Based Systems*, vol. 204, Article ID 106172, 2020.

[43] G. Xu, W. Li, and J. Liu, "A social emotion classification approach using multi-model fusion," *Future Generation Computer Systems*, vol. 102, pp. 347–356, 2020.

[44] H. B. Kang, "Affective content detection using hmms," in *Proceedings of the eleventh ACM international conference on multimedia, ACM*, Berkeley CA USA, November 2003.

[45] A. K. Hassan and S. N. Mohammed, "A novel facial emotion recognition scheme based on graph mining," *Defence Technology*, vol. 16, no. 5, pp. 1062–1072, 2020.

[46] G. Irie, T. Satou, A. Kojima, T. Yamasaki, and K. Aizawa, "Affective audio-visual words and latent topic driving model for realizing movie affective scene classification," *IEEE Transactions on Multimedia*, vol. 12, no. 6, pp. 523–535, 2010.

[47] A. Gupta, A. D'Cunha, K. Awasthi, and V. Balasubramanian, "Daisee: towards user engagement recognition in the wild," 2016, https://arxiv.org/abs/1609.01885.

[48] A. T. Lopes, E. de Aguiar, A. F. De Souza, and T. Oliveira-Santos, "Facial expression recognition with convolutional neural networks: coping with few data and the training sample order," *Pattern Recognition*, vol. 61, pp. 610–628, 2017.

[49] T. Huang, Y. Mei, H. Zhang, S. Liu, and H. Yang, "Fine-grained engagement recognition in online learning environment," in *Proceedings of the 2019 IEEE 9th international conference on electronics information and emergency communication (ICEIEC)*, pp. 338–341, IEEE, Beijing, China, July 2019.

[50] Y. Hayashi, "Detecting collaborative learning through emotions: an investigation using facial expression recognition," in *Proceedings of the International conference on intelligent tutoring systems*, pp. 89–98, Springer, Kingston, Jamaica, June 2019.

[51] L. Ramirez, W. Yao, E. Chng et al., "Toward instrumenting makerspaces: using motion sensors to capture students' affective states and social interactions in open-ended learning environments," in *Proceedings of the 12th Int. Conf. Educ. Data Mining (EDM)*, pp. 639–642, Montréal, PQ, Canada, July 2019.

[52] T. J. Tiam-Lee and K. Sumi, "Analysis and prediction of student emotions while doing programming exercises," in *Proceedings of the International conference on intelligent tutoring systems*, pp. 24–33, Springer, Kingston, Jamaica, June 2019.

[53] R. Subramanian, J. Wache, M. K. Abadi, R. L. Vieriu, and S. Winkler, "ASCERTAIN: emotion and personality recognition using commercial sensors," *IEEE Transactions on Affective Computing*, vol. 9, no. 2, pp. 147–160, 2018.

[54] Y. Liu and C. Jiang, "Recognition of shooter's emotions under stress based on affective computing," *IEEE Access*, vol. 7, pp. 62338–62343, 2019.

[55] X. Huang, A. Dhall, R. Goecke, M. Pietikainen, and G. Zhao, "Multimodal framework for analyzing ¨ the affect of a group of people," *IEEE Transactions on Multimedia*, vol. 20, no. 10, pp. 2706–2721, 2018.

[56] T. S. Ashwin and R. M. R. Guddeti, "Automatic detection of students' affective states in classroom environment using hybrid convolutional neural networks," *Education and Information Technologies*, vol. 25, no. 2, pp. 1387–1415, 2020.

[57] A. Chaudhari, C. Bhatt, A. Krishna, and P. L. Mazzeo, "ViTFER: facial emotion recognition with vision transformers," *Applied System Innovation*, vol. 5, no. 4, p. 80, 2022.

[58] P. Foggia, A. Greco, A. Saggese, and M. Vento, "Multi-task learning on the edge for effective gender, age, ethnicity and emotion recognition," *Engineering Applications of Artificial Intelligence*, vol. 118, Article ID 105651, 2023.

[59] K. Prabhu, S. SathishKumar, M. Sivachitra, S. Dineshkumar, and P. Sathiyabama, "Facial expression recognition using enhanced convolution neural network with attention mechanism," *Computer Systems Science and Engineering*, vol. 41, no. 1, pp. 415–426, 2022.

[60] B. Pan, S. Wang, and B. Xia, "Occluded facial expression recognition enhanced through privileged information," in *Proceedings of the 27th ACM International Conference on Multimedia*, pp. 566–573, October 2019.

[61] M. Tan, G. Ni, X. Liu et al., "Bidirectional posture-appearance interaction network for driver behavior recognition," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 8, pp. 13242–13254, 2021.