Received July 18, 2021, accepted July 26, 2021, date of publication August 3, 2021, date of current version August 12, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3102042

An Automatic Multimedia Likability Prediction System Based on Facial Expression of Observer

VIVEK SINGH BAWA[1], SHAILZA SHARMA[2], MOHAMMED USMAN[3], (Senior Member, IEEE), ABHIMAT GUPTA[4], AND VINAY KUMAR[2]

[1]Visual Artificial Intelligence Laboratory, Oxford Brookes University, Oxford OX3 0BP, U.K.
[2]Department of Electronics and Communication Engineering, Thapar Institute of Engineering and Technology, Patiala 147004, India
[3]Department of Electrical Engineering, King Khalid University, Abha 61411, Saudi Arabia
[4]Department of Computer Science and Engineering, Thapar Institute of Engineering and Technology, Patiala 147004, India

Corresponding author: Vinay Kumar (vinay.kumar@thapar.edu)

The authors extend their appreciation to the Deanship of Scientific Research at King Khalid University for funding this work through General Research Project under grant number (RGP.1/376/42).

ABSTRACT Every individual's perception of multimedia content varies based on their interpretation. Therefore, it is quite challenging to predict likability of any multimedia just based on its content. This paper presents a novel system for analysis of facial expressions of subject against the multimedia content to be evaluated. First, we developed a dataset by recording facial expressions of subjects under uncontrolled environment. These subjects are volunteers recruited to watch the videos of different genre, and provide their feedback in terms of likability. Subject responses are divided into three categories: Like, Neutral and Dislike. A novel multimodal system is developed using the developed dataset. The model learns feature representation from data based on the three provided categories. The proposed system contains ensemble of time distributed convolutional neural network, 3D convolutional neural network, and long short term memory networks. All the modalities in proposed architecture are evaluated independently as well as in distinct combinations. The paper also provides detailed insight into learning behavior of the proposed system.

INDEX TERMS Affective computing, deep neural architecture, facial expression analysis, multimedia evaluation system, representation learning.

I. INTRODUCTION

Facial expressions are the most powerful non-verbal tools used by human beings to share various types of information amongst themselves. Facial expressions have played significant role in the evolution of our species [1]. Unless one is trained or proficient to conceal, generally expressions (micro or macro) convey true human response to an applied stimuli [1]. Study of facial expression finds application in wide range of areas, such as mental health diagnosis [2], gaming experience [3], customer services [4], automotive safety [5], and human machine interaction [6]. In the last decade, facial expression analysis and emotion recognition have attracted substantial amount of interest from computer vision research community [7]–[11].

Most common application for study of facial expression is in human emotion recognition. According to Paul Ekman [12], emotions can be divided into six basic categories: happy, sad, surprise, disgust, fear, and anger. Although, few more categories of emotions has been included by other researchers; such as, boredom, contempt and engagement etc. [13], [14]. These emotions are used to understand subject response and infer decisions. But, humans exhibit very complex behavior and sometimes response of subject to the provided stimuli can be a hybrid emotion; i.e., combination of multiple emotions. Hence, approaches based on expression to emotion mapping can fail in complex scenarios.

Earlier emotion recognition and expression analysis methods used to rely on static images [7], [15], [16]. These methods use local spatial structures to identify the existing emotion. In other scenarios, where facial expression changes with time, these static methods remain inadequate. In such cases, dynamics (video) based methods are preferred, which can extract spatial as well as temporal information from video of face [17], [18]. Sometimes multiple modalities are also used to understand the true emotion of subject [19], [20].

The associate editor coordinating the review of this manuscript and approving it for publication was Shen Yin.

Proposed work focuses on the commercial aspect of the facial expression analysis. In this paper, we present a novel framework for automatic evaluation of likability of multimedia content by observing the facial expression of viewer. The system produces output in three classes: like, neutral and dislike. The following reasons make this task very intricate in nature:

1) The expression effectuated on face of subject are stimulated by the content of observed multimedia. Hence, elicited expressions are natural instead of acted. Since natural expressions create very insignificant muscle movement, compared to the acted ones, they are difficult to recognize correctly.
2) The multimedia can have a wide variety of content leading to wide range of expressions. For example, two different videos liked by a subject can have very different content and cause completely unique set of expressions. In other words, collected dataset will have very high intra-class variance.
3) Even in a single observed video, the subject can produce wide range of expressions depending on the flow of content.
4) Finally, we propose a novel method that used combination of three submodules to learn different appearance features. Model also presents a novel approach to avoid overfitting to training dataset in limited data scenario.

The system proposed in this paper uses multiple modalities to capture different set of features from subject face. The ensemble based architecture is designed to learn spatio-temporal features from input video sequence. The system is trained from end to end in single training cycle.

Rest of the paper is organized in following sections: state of the art is discussed in Section II, motivation for the presented work follows in Section III, data collection and experimental setup are discussed in Section IV, details of the implementation of proposed framework are presented in Section V, section VI discusses the results on proposed architecture, and their analysis followed by conclusion in Section VIII.

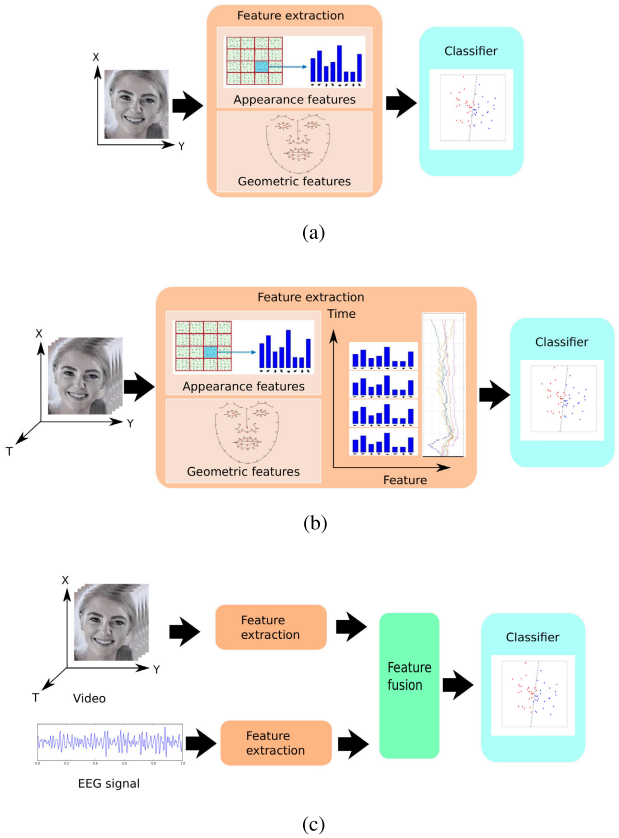

FIGURE 1. Different types of methods used in affective computing (a) static image based spatial feature methods (b) video based spatio-temporal feature extraction methods (c) multi-modal feature extraction methods based on multiple types of signal.

II. RELATED WORK

The present work falls under the ambit of affective computing, which tries to recognize underlying emotion from different formats of data. However, the present work attempts to infer the user's response to the applied stimuli instead of recognizing effect. Response of user to a certain stimuli depends on underlying complex psychophysical phenomenons. The approaches for affective computing can be divided into three major categories based on the input data type: statics (image) based, dynamics (video) based and multi-modal methods.

A. STATIC IMAGE BASED METHODS

Static image based approaches utilize geometry based or appearance based techniques to extract the features from current image input (shown in fig. 1a).

Involvement of different facial muscles makes facial expression analysis a very complex task [21], [22]. The relative movement of facial muscles can also vary for each individual. Two different approaches are used to analyze facial expressions: appearance based methods [23], [24] and geometry based methods [25], [26]. Appearance based methods rely on various feature extraction techniques to capture the variation in facial features. Geometry based methods use fiducial facial points to model the movement of different facial muscles.

[7]–[9] use facial landmarks to identify the geometric structure of the face. Geometric characteristic variation on face are modelled to link underlying emotion. [27] used dense SIFT (Scale-Invariant Feature Transform) features to generate Bag of Words (BOW) to model the appearance based facial features. [28] extracted facial appearance features using Local Fisher Discriminant Analysis (LFDA). Appearance features were extracted by [16] from a set of specific facial patches to enhance discriminative power of model.

[29] developed a deep Convolutional Neural Network (CNN) based method for affect recognition in SFEW dataset [10] containing static images from the movie scenes. The task requires mapping of emotional state of the main subject to one of the basic seven categories of emotions. CNN was pre-trained on FER dataset [11] due to lack of

training samples in SFEW dataset. Multiple CNN models were averaged at the end to calculate the final output. [30] uses static images to develop an automatic pain assessment system, which used combination of both geometric as well as appearance features.

These methods have inferior performance due to their reliance on only current state of subject's expressions, but they require very low computational resources.

B. DYNAMICS (VIDEO) BASED METHODS

Dynamic approaches use temporal information in addition to spatial information for recognizing emotions (refer fig. 1b). These methods generally perform better than static methods but require much higher memory and computational power.

[18] developed Bayesian network to extract continual set of features to capture temporal variation in features. [31] used Independent Component Analysis (ICA) to extract the discriminative components from video sequence. [32] proposed LGBP-TOP features (Local Gabor Binary Pattern from Three Orthogonal Planes) to simultaneously model spatial as well as temporal structures corresponding to certain emotions. [17] presented STLMBP (SpatioTemporal Local Monogenic Binary Pattern) features, which are also good at modeling spatio-temporal features in video sequence.

Fan *et al.* [33] developed a hybrid network for emotion recognition in videos. Network contained two input pipelines, first pipeline was 3D convolutional neural network (3D-CNN) with a capability to extract spatial and temporal features simultaneously. Second pipeline had CNN paired RNN. The CNN extracts spatial features of each frame while RNN learns the temporal relationship between the features of each frame.

Specialist model approach, where each model learns a specific modality, was developed by [34]. It generates the final score by combining the outcome of all the models. Kaya *et al.* [35] developed a multi-modal system for emotion recognition in wild. CNN along with conventional feature extractor (SIFT-FUN, LPQ-TOP, LGBP-TOP) is used to extract the visual feature from image. Audio feature are extracted using openSMILE library [36]. Extreme Learning Machine (ELM) and Partial Least Square (PLS) are trained independently on these features and later weighted averaging is used for score fusion.

C. MULTIPLE MODALITY BASED METHODS

Multi-modal approaches use different types of signals; such as, audio, EEG etc., to produce a better hypothesis (shown in fig. 1c). Han *et al.* [37] developed a multi-modal regression model to predict the arousal and valence levels. They used Bidirectional Long Short Term Memory (BLSTM) network to learn the features from image and audio data. Support Vector Regression (SVR) model was used on top of these extracted features to produce the arousal and valance activity of a face.

[38] uses only EEG signals to classify the emotions. The authors use mutual information based feature selection procedure along with kernel based classifier to improve the classification accuracy. Combination of textual information and video data was used by [39] to correctly classify the emotional sentiments in social media content. Zheng *et al.* [40] developed a deep belief network based system to classify the EEG signal into three emotions: positive, negative, and neutral. EEG signal for each class was generated by providing external stimuli to subject through different types of videos.

Turker *et al.* [41] tried to identify the laughter in a natural environment using audio and video features. They manually annotated the dataset for laughter audio and other environmental noises. Head movement and facial features are extracted from frames of videos, and audio features are extracted from audio data. These features are fused together and classified using SVM classifier. Additionally, time delayed neural network is also deployed to improve the detection accuracy.

D. MEDIA ANALYSIS WITH AFFECTIVE STATE

Different techniques have been proposed to measure true subject response against the observed media. [42] used combination of EEG signal, gaze, and pupillary activity to extract user response against the given video. Micu *et al.* [43] recorded electromyography (EMG) data from subjects while they watch different types of advertisements. They linked the activity in EMG signal to self reported facts. [44], [45] developed a system to label content of video based on physiological data of viewer and content of the movie scene. [46], [47] showed that extent of smile feature can be used to identify if viewer likes the product presented in the advertisement. [48] developed a system to analyze the level of engagement of viewer based on head pose and activity of various action units. [49] used the facial activity and body motion of subjects to develop a system to predict movie ratings.

Most of the discussed multi-modal approaches may not be practically suitable for proposed application. Proposed task needs to produce like/dislike sentiment for a video irrespective of location, device, and environment of viewer. It is not always possible to get the physiological signals from the subject. The video data is the most easily accessible and suitable choice for the proposed system.

III. MOTIVATION

With the advent of internet, online multimedia content has shown an exponentially increasing trend in the last decade. Popularity of social media sites like Youtube, Facebook, Twitter etc. are also one of the root causes of this growth. Facebook has around 2.23 billion monthly active users [50]. Similarly, Youtube hosts more than 1.5 billion users every month [51]. Approximately, 30 million users use Youtube on daily basis out of its total user bank [52]. Over 300, 000 user are paying Youtube for its services and 50 million users have uploaded a total of more than 5 billion videos.

This vast outreach of social media platforms has made them biggest commercial platform of this era. A large number of people these platforms are their main source of income.

The corporations use these platforms to promote their products through advertisements. Additionally, online streaming services like Netflix, Amazon prime and Hulu etc. have also become mainstream sources of entertainment.

All these platforms are great way to get real-time feedback from the user and improve the quality of product. But current feedback systems ask user to manually provide feedback in the form of comments or through like/dislike button. It has been observed that even a well performing content gets only around 4% view to feedback ratio. Whereas, maximum of the content receives less than 1% feedback. These statistics are too low and susceptible to deception. The feedback received by these methods can vary from true response, and is very small to be of any significance at early stage.

An automated feedback system based on facial expression of the viewer is a much better and faster strategy to collect the feedback of your target audience with 100% feedback rate. As the expression of the user are largely immune to deception, it is a good technique to collect true response.

IV. DATA ACQUISITION

A. EXPERIMENTAL SETUP

Data collection is very critical step for the development of a reliable system in any type of application. Any kind of biases and constraints in collected data can limit the scope of developed system. Hence, we tried to make the data collection process as realistic as possible.

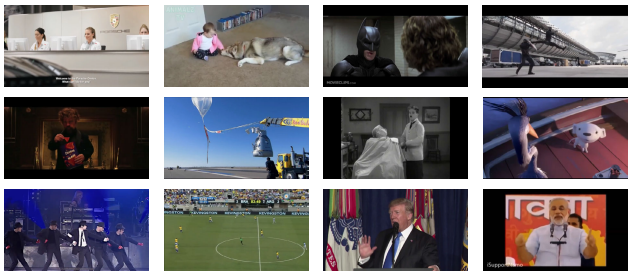

FIGURE 2. Some random samples from the videos used to stimulate the subject response.

Data collection process requires collection of two types of data. First data is the video samples, which are used to stimulate the natural facial responses of the subjects. We selected more than 150 videos from a wide variety genres, such as, movie, music, sports, fights, war, speech, advertisements etc. Some screenshots of the sample videos are shown in fig. 2, and number of samples in each genre are presented in Table 1. Duration of these video clips varies between 2 mins and 5 mins. Second data collection is the response of the viewers. Length of recorded response is very important in this case. We observed that response of a subject varies a lot throughout the recording session. Hence, it is not feasible to identify the correct viewer sentiment based on a short length clip (e.g. 5-10 sec).

TABLE 1. Table showing the number of videos belonging to each genre.

Genre	Number of sample
Movie	25
Music	16
Speech	13
Advertisement	50
War/fight	18
Horror	11
Comedy	21

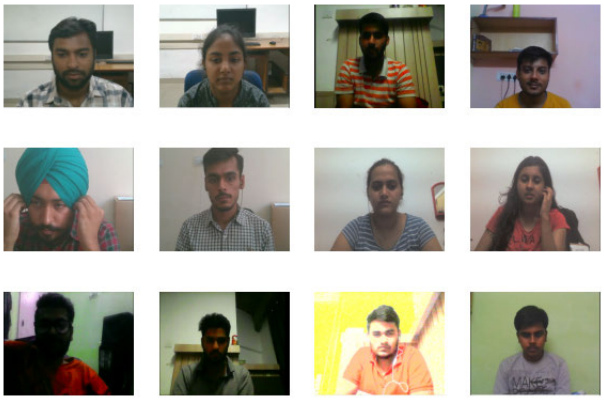

FIGURE 3. Samples from the recorded videos of subjects while watching the multimedia content.

B. SAMPLE RECORDING

All recorded samples have a frame rate of 15 fps and frame size of 640 × 480. In total, 73 volunteers are recruited for the task of response recording. All the recordings are conducted in uncontrolled environment. Different locations and lighting conditions can be observed in all the samples shown in fig 3.

TABLE 2. Details of the volunteer recruited as subjects for data acquisition.

	Male	Female	Total
Video count	254	143	397
Subject count	29	44	73

Minimum age	18
Maximum age	28
Median of age	20
Mean of age	21.4

Subjects are selected to maximize the diversity in gender and facial characteristics (shown in fig. 3). Subjects have wide range of facial features, such as, beard, moustache, turban, glasses etc. All the subject have age between 17 and 58 years. More details on subjects are given in Table 2. Each subject is shown a video to stimulate the facial expressions which are recorded using webcam. Only 3 to 6 samples are recorded from single subject in one sitting to minimize the effects of physical fatigue. To get a uniform number of samples for all categories of videos, each subject is encouraged to

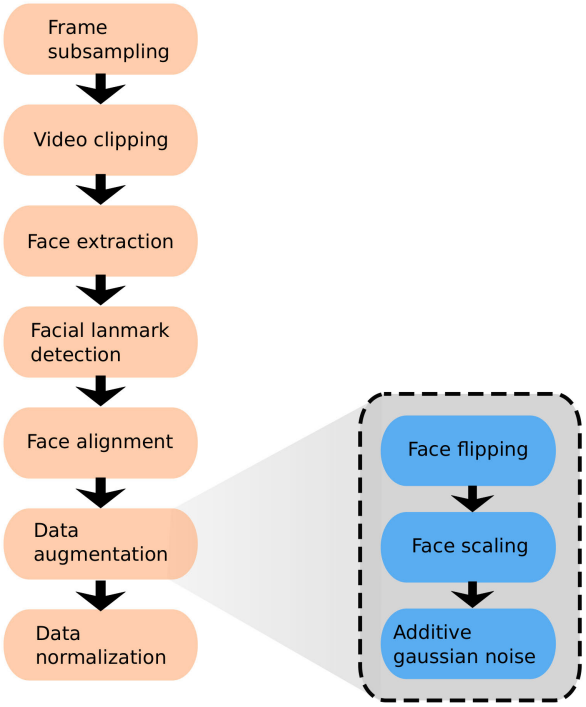

FIGURE 4. Block diagram of video preprocessing pipeline for the proposed system.

watch videos from different genres. Subjects are asked to rate the video in three classes: "Like", "Neutral", or "Dislike".

V. METHODOLOGY

A. PREPROCESSING OF DATA

Deep neural networks are prone to overfitting and thus require large amount of data to achieve a better generalization on test dataset. Hence, multiple preprocessing methods are used on original videos to increase sample count for the training process. Block diagram of the data preprocessing pipeline is shown in fig. 4.

As mentioned earlier, the original videos of subjects are recorded at 15 fps. Using high frame rate can be computationally very expensive and do not hold any substantial advantage. Hence, first step of preprocessing is frame subsampling. We reduce the frame rate to 1 fps, which is neither too low to loose significant information nor too high to increase computational requirements of framework. In next step, recorded videos are split into 1 min segments to increase the sample count for training data. One minute duration is long enough to provide adequate arousal activity and correctly identify the response from subject's face. This process enables us to increase sample count from 397 to 1372. The process of frame rate reduction and video splitting is also shown in fig. 5. Total number of samples for each category, after the application of discussed procedure, are shown in Table 3.

Next stage in preprocessing is segmentation of faces and extraction of facial landmarks in each frame. Dlib library [53] is used to extract the faces as well as facial landmarks from

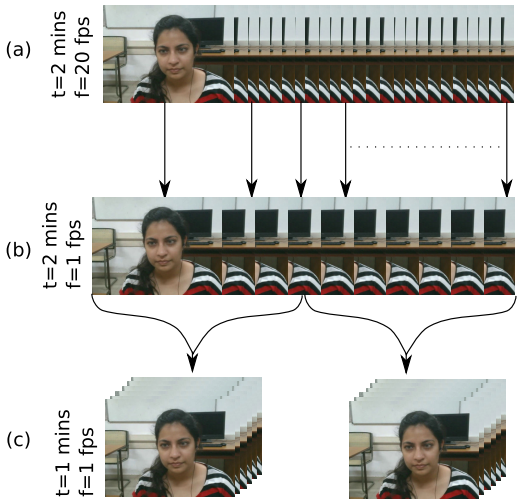

FIGURE 5. The frame rate reduction and video splitting procedure (a) Original video sample (b) Video sample of same duration but with reduced frame rate (c) Video split into multiple sample of 1 min duration.

TABLE 3. Count of samples belonging to each of the three categories.

Categories	Number of samples
Like	773
Neutral	317
Dislike	282
Total sample	1372

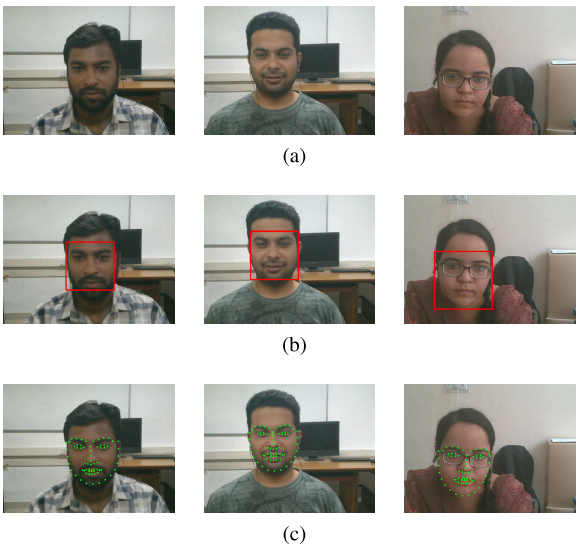

FIGURE 6. Results of segment faces and detected facial landmarks in video sequence (a) Frame from actual recorded video (b) Extracted faces from one frame of video sequence (c) Extracted facial landmarks from one frame of video sequence.

the input frames. The faces are extracted with frame size of (100×100) in RGB color space. Then, all 68 facial landmarks are extracted from each of the frames (shown in fig 6c). Each landmark is depicted by x & y coordinates, hence the output array of landmarks for each frame has (2×68) size.

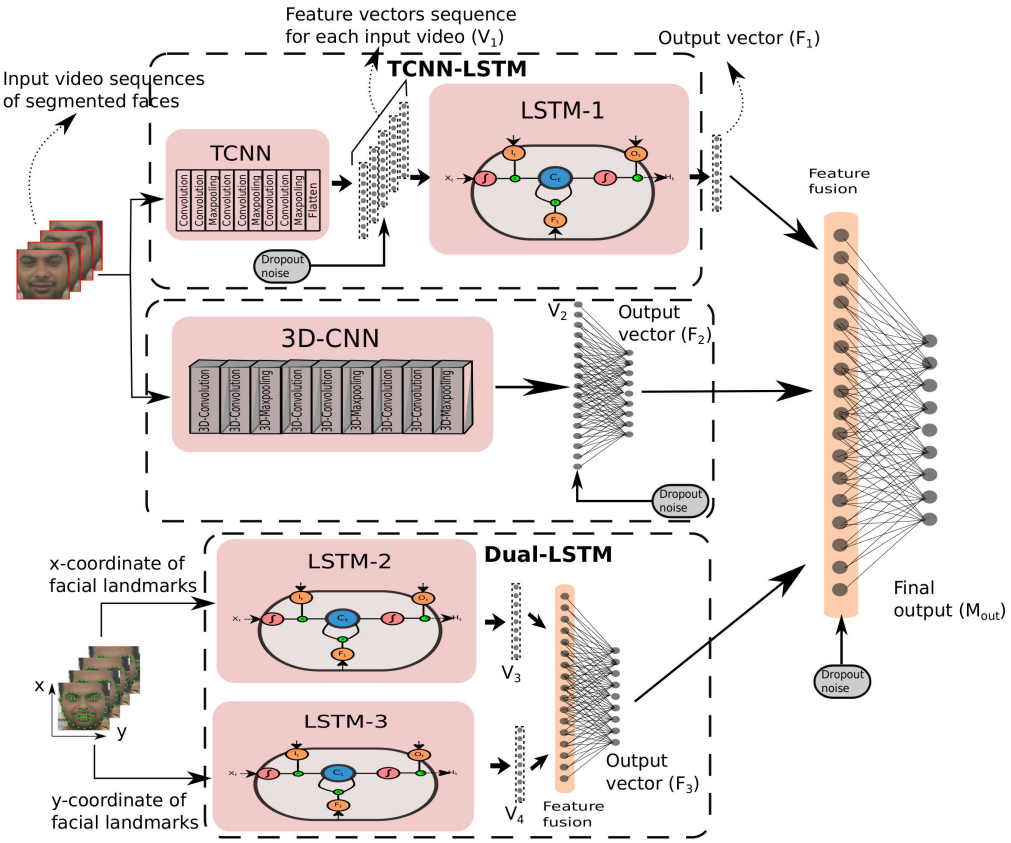

FIGURE 7. Proposed Deep Composite Neural Architecture (DCNA) with ensemble of three submodules to learn different modalities from input video sequence. The model is a single cycle end to end trainable system.

Some resulting samples of face segmentation and landmark detection are shown in fig. 6.

The extracted facial landmarks are used for face alignment. Affine transformation is used to align faces based on the location of selected landmarks. Only aligned face videos are used for model training. Data augmentation is also incorporated in the data preprocessing pipeline to further increase the variability in the our dataset. We can not use all of the commonly used augmentation techniques on facial images due to special geometry of faces. Three augmentation methods are used in this work: face flipping, face scaling, and additive Gaussian noise. For face scaling, a scaling factor of 1.1 is used. Higher and lower scaling factors can eliminate the important facial information. Additive Gaussian noise, generated with the help of equation 1, is also added to the frames of input video to achieve better generalization behavior from the model.

$$N_g(z) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(z-\mu)^2}{2\sigma^2}} \tag{1}$$

here, $N_g(z)$ represents the probability density function for Gaussian noise for given pixel value (z). μ and σ are mean and standard deviation for the noise, and can be adjusted to change the amount of noise added into input sample.

The last stage of pipeline is data normalization. Data normalization, given in equation 2, is used to limit the range of data values so that all the videos have similar sample structure.

$$X_{norm}(i,j) = \frac{x(i,j) - X_{min}}{X_{max} - X_{min}} \tag{2}$$

here, X_{norm} is the output pixel value for input pixel value x at coordinates (i, j). X_{max} and X_{min} are maximum and minimum value present in matrix for a given batch, respectively.

B. PROPOSED ARCHITECTURE

We propose a novel Deep Composite Neural Architecture (DCNA) to learn the facial arousal activity of subjects. DCNA is an ensemble of three models: TCNN-LSTM, 3D-CNN and Dual-LSTM. These three models, discussed next, learn different modalities from the input video sequence (shown in fig. 7).

The first model (TCNN-LSTM) is a combination of time distributed CNN (TCNN) and LSTM (represented as LSTM-1 in fig. 7). The second model, 3D convolutional neural network (3D-CNN), has three dimensional kernels to extract spatio-temporal features from the frames of video sequence. The third model does not take video sequence as

TABLE 4. Table showing layer wise kernel size and count for TCNN architecture.

Layer	Kernel size	kernel count
Conv2D	3×3	12
Conv2D	3×3	12
Maxpool2D	2×2	12
Conv2D	3×3	16
Conv2D	3×3	16
Maxpool2D	2×2	16
Conv2D	3×3	20
Conv2D	3×3	20
Maxpool2D	2×2	20

TABLE 5. Table showing layer wise kernel size and count for 3D-CNN architecture.

Layer	Kernel size	kernel count
Conv3D	$3 \times 3 \times 3$	12
Conv3D	$3 \times 3 \times 3$	12
Maxpool3D	$2 \times 2 \times 2$	12
Conv3D	$3 \times 3 \times 3$	16
Conv3D	$3 \times 3 \times 3$	16
Maxpool3D	$2 \times 2 \times 2$	16
Conv3D	$3 \times 3 \times 3$	20
Conv3D	$3 \times 3 \times 3$	20
Maxpool3D	$1 \times 2 \times 2$	20

input but uses coordinates of all 68 facial landmarks to map the arousal patterns of different landmarks in the given time sequence. This model contains two LSTMs (LSTM-2 and LSTM-3, refer fig. 7) to extract temporal patterns in both x and y coordinates independently. The architectural details, like kernel size and kernel count for TCNN and 3D-CNN, are given in Table 4 and 5, respectively.

One critical drawback of learning representations in neural networks from the dataset is that it is not possible to control what kind of feature model learns. Additionally, using a single style of feature extraction units (e.g. 2D convolutional kernel or 3D convolutional kernel) limits the feature representation power of the network. Hence, we use two different modules TCNN-LSTM and 3D-CNN to learn the appearance features from the original facial images. The third module (Dual-LSTM) is employed to learn the motion of the facial landmarks. This helps in reducing effect of the noise in the face images.

The TCNN model is a 6 convolutional layer based CNN model (Table 4). It fetches one frame at a time and produces reduced feature vector per frame. Convolutional neural network extracts feature from the each frame and produces a 2304 dimensional feature vector for each image. Features extracted at each layer can be represented by eq. 3.

$$F_l(i,j) = \sum_m \sum_n \sum_d F_{l-1}(i+m, j+n, d) \times K_l(m,n,d) \tag{3}$$

here, F_l is the output feature map for l^{th} convolutional layer, and F_{l-1} is the output feature map of previous layer. K_l is the convolutional kernel for current layer with size $(m \times n)$, and d represents the depth (number of kernels in previous layer). The final output of a time distributed CNN (TCNN) can be represented by eq 4.

$$V_{1i} = f_\beta(w_\beta, b_\beta, I_i) \quad \forall\, i = (1, 2, 3, \cdots, 60) \tag{4}$$

V_{1i} is output feature vector for input image frame (I_i) of a video, i varies from 1 to 60 as each video has 60 frames. f_β is the function representing CNN model with weights w_β and biases b_β.

Proposed model uses multi level noise induction at different stages of feature learning. This noise propels model to learn true data distribution and helps to avoid overfitting. Dropout is used as noise with a random dropout value of 0.2. Output after multiplying with binary noise vector becomes:

$$X_{1i} = V_{1i} \times \tilde{N}_{dropout} \tag{5}$$

here, $\tilde{N}_{dropout}$ represents the random binary vector for dropout of feature vector with a probability of 0.2. The output feature sequence (X_{1i}) is passed through LSTM-1 to learn the temporal dependencies between features of different time frames of input video.

Long Short-Term Memory (LSTM) networks are the most commonly used neural networks for sequential relationship modeling. LSTM does not suffer from vanishing gradients, unlike conventional Recurrent Neural Networks (RNN), and are easy to train. The LSTM-1 contains 128 memory cells and takes 60 sequential spatial feature vectors of size 2880. The output of the model is 128 dimensional vector.

Hidden state of any memory cell (H_s) of LSTM can be defined as a function of previous hidden state (H_{s-1}) and input at the current time step (X_s) (refer eq. 6).

$$H_s = f_\alpha(H_{s-1}, X_s) \quad \forall\ X_s \in \mathbb{R}^{L \times T} \tag{6}$$

here, X_s is a real valued vector of dimensions $(L \times T)$, L represents the length of a single vector and T represents the number of time steps in a given sequence. H_s can also be represented mathematically in term of parameters $(w_\alpha, u_\alpha, b_\alpha)$ and input data (X_s) as shown in eq. 7.

$$H_s = H_f(w_\alpha X_s + u_\alpha H_{s-1} + b_\alpha) \quad \forall\, s \in [1, T] \tag{7}$$

and input to LSTM is the output of TCNN, hence:

$$X_s = X_{1i} \tag{8}$$

Output feature vector (F_1), given in eq 9 and generated by LSTM-1, represents the probability of occurrence of feature vector (X_{t+1}) given the feature vector for previous frames $(X_t, X_{t-1}, \ldots\ldots)$.

$$F_1 = P(X_{t+1} | X_t, X_{t-1}, \ldots\ldots X_2, X_1) \tag{9}$$

By chain rule of probability, we can decompose eq 9 as:

$$F_1 = P(X_{t+1}|X_t) \times P(X_t|X_{t-1}) \times P(X_{t-1}|X_{t-2}) \ldots\ldots \tag{10}$$

The above equation can also be rewritten as:

$$F_1 = \prod_{t=1}^{T} P(X_{t+1}|X_t) \tag{11}$$

Similar to the first model of ensemble, second model (3D-CNN) also uses convolution based operation to extract spatial as well as temporal features from input video sequence. The operation can be mathematically represented as:

$$F_l(i,j,k) = \sum_m \sum_n \sum_o \sum_d F_{l-1}(i+m, j+n, k+o, d) \times K_l(m,n,o,d) \tag{12}$$

here, F_l represents the output feature map after convolution with 3 dimensional kernel (K_l) of size ($m \times n \times o$). d representer the depth (number of kernels in previous layer). It can be observed from eq. 12 and eq. 3 that both, output features and convolutional kernel, have an extra dimension in case of 3D-CNN. The final output of the 3D-CNN model can be represented as:

$$V_{2i} = f_\gamma(w_\gamma, b_\gamma, I_{video}) \tag{13}$$

here, V_{2i} is output feature vector for complete input video sequence (I_{video}). f_γ is the function representing 3D-CNN model with weights w_γ and biases b_γ. Architectural details of 3D-CNN model are given in Table 5.

Dropout is also used at the output of 3D-CNN model to learn the true data distribution for a particular category with dropout probability of 0.2 (shown in eq. 14).

$$X_{2i} = V_{2i} \times \tilde{N}_{dropout} \tag{14}$$

The noisy output (X_{2i}) is then passed through a fully connected layer to compress the feature vector. Final output (F_2) can be represented as a function of weight (w_θ), biases (b_θ) and input (X_{2i}) of fully connected layer (eq. 15).

$$F_2 = f_\theta(w_\theta, b_\theta, X_{2i}) \tag{15}$$

The third model in ensemble is used to model effect of arousal activity on the facial landmarks. Sequence of x landmark coordinates is given to LSTM-2, and sequence of y coordinates is given to LSTM-3. Output features of LSTM-2 and LSTM-3 are concatenated and passed through a fully connected layer for feature compression.

Coordinate of all 68 landmarks for 60 frames can be represented as $\mathscr{L}(x,y,n,f)$, where x and y represent the coordinate value in x and y direction. $n(\in [1,68])$ represents the landmark number, and f is the frame number for the given set of coordinates. LSTM can not model two dimensional temporal sequences, therefore we separately model the activity in x and y directions for each of the landmark. Each of these LSTMs (LSTM-2 and LSTM-3) have 128 memory cells and generates 128 dimensional feature vector which are concatenated to generate 256 dimensions. LSTM-2 models the sequence of x coordinates of landmarks ($\mathscr{L}_t^x$) which can be mathematically represented as (used directly from previously derived eq. 11):

$$V_3 = \prod_{t=1}^{T} P(\mathscr{L}_{t+1}^x|\mathscr{L}_t^x) \tag{16}$$

here, V_3 is conditionally dependent feature vector of size 128 for x coordinates.

LSTM-3 models the sequence of y coordinates of landmarks ($\mathscr{L}_t^y$), which can be mathematically represented as:

$$V_4 = \prod_{t=1}^{T} P(\mathscr{L}_{t+1}^y|\mathscr{L}_t^y) \tag{17}$$

here, V_4 is conditionally dependent feature vector of size 128 for y coordinates.

Final output, after feature compression through a fully connected layer, is a function of weights (w_δ), biases (b_δ), and inputs (V_3 and V_4) of fully connected layer (eq. 18).

$$F_3 = f_\delta(w_\delta, b_\delta, V_3, V_4) \tag{18}$$

The final outputs of three models are fused and passed trough a fully connected model to compute the combined class score. This final output (M_{out}) of the composite architecture is represented by the following equation:

$$M_{out} = f_\eta(w_\eta, b_\eta, F_1, F_2, F_3) \tag{19}$$

C. MODEL OPTIMIZATION

Categorical cross-entropy (shown in eq. 20) is used as loss function to train the end to end system.

$$L(y', y) = -\frac{1}{N}\sum_{i=0}^{N}\left[y_i log(y'_i) + (1-y_i) log(1-y'_i)\right] \tag{20}$$

here, $L(y', y)$ is the loss value, y'_i and y_i denote predicted and target labels, respectively; and N represents single mini-batch sample count. In proposed work, mini-batch size of 10 is used to train the model.

We use RMSprop [54] optimizer to update model parameter according to the value of loss function. If L_t is the loss value for a given mini-batch and w_t are the model parameters, then, gradient (g_t) for current mini-batch is given by eq 21.

$$g_t = (1-\gamma)\left[L'_t\right]^2 + \gamma g_{t-1} \tag{21}$$

here, γ represents the momentum, which can be used to control the contribution of present and past gradients in the current updation of parameter. The parameter updation value (Δw_t) is given by following equation:

$$\Delta w_t = -\frac{\eta}{\sqrt{g_t + \epsilon}} \times L'_t \tag{22}$$

here, η is the learning rate, and ϵ is a mathematical constant used to avoid non-integer value for parameter update. The final value of model parameter is given by eq. 23.

$$w_{t+1} = w_t + \Delta w_t \tag{23}$$

Out of total 1372 video samples, generated after pre-processing of data, we use 1200 samples for training and 172 samples for testing, each mini-batch is of 10 samples. All the models are trained for 150 epochs without mini-batch learning. Batch normalization, which helps in avoiding exploding and vanishing gradient problems, is used on output of all the activation layers for all the models.

TABLE 6. Accuracy achieved by proposed composite neural architecture on test data.

Model	Parameters	Test Accuracy	Train accuracy
TCNN-LSTM	1,553,223	55.23	56.50
3D-CNN	5,198,479	73.83	99.25
Dual-LSTM	218,371	53.23	56.50

VI. RESULTS AND DISCUSSION

Each model in the ensemble of deep composite neural architecture (DCNA) is of paramount importance. Each modality is tested independently to study its performance for the proposed problem. Table 6 shows test and training accuracies achieved by all of three basic models individually. Each of these models are trained using same training data pipeline described in Section V-A. 3D-CNN model is able to achieve the highest accuracy out of the three with test accuracy of 73.83. Both TCNN-LSTM and Dual-LSTM achieve around 55.23% and 53.23% accuracy on validation set, respectively. The reason for significantly lower performance is a combination of multiple factors. In our experiments we sound that if we increase model capacity (parameters), models start to overfit to training data. Hence, parameters can not be increased beyond a certain limit. Furthermore, presence of only one module pushes these models to focus on one specific type of features leading to lower performance. Additionally in case of Dual-LSTM, only facial landmarks are used for training the model. Hence it can not extract any other appearance features from the face image which leads to lowest performance of all.

Both the combinations of vision based models with landmark based model are able to significantly improve the performance (shown in Table 7). Combination of TCNN-LSTM and Dual-LSTM models achieves same test accuracy as combination of 3D-CNN and Dual-LSTM models. Although, combination of TCNN-LSTM and Dual-LSTM achieve lower accuracy on 'Neutral' class, and higher accuracy on both 'Like' and 'Dislike' categories. Both of these models achieve significantly lower performance on neutral class because lack of arousal activity on face make it difficult to correctly identify the class. The 3D-CNN and TCNN-LSTM combination achieves significantly lower accuracy than other two. This reduction is due to the fact that we are using similar type of models which use same images to learn the class probabilities. Over parameterization and data overfiting can be be considered as direct cause of the lowered performance.

TABLE 7. Analysis of performance of different combination basic models on final test performance against final DCNA.

Model	Parameters	Class accuracy			Average accuracy	Train accuracy
		Like	Neutral	Dislike		
TCNN-LSTM+ Dual-LSTM	1,796,103	91.57	43.90	72.22	76.16	92.16
3D-CNN+ Dual-LSTM	5,457,359	88.42	53.66	69.44	76.16	99.41
3D-CNN+ TCNN-LSTM	6,802,003	74.12	60.13	65.03	61.10	95.76
DCNA (without dropout)	7,028,371	86.31	39.02	41.67	65.69	99.33
DCNA (with multi-level dropout)	7,028,371	85.26	73.17	77.77	80.81	96.91

Full DCNA has significantly higher performance when the architecture uses multi-level dropout (refer Table 7). Elimination of dropout noise from model reduces test accuracy by 15.12%. The variation in the performance is due to the overfitting of DCNA when no dropout is applied. The proposed architecture contains multiple modules and the dataset is smaller in comparison to datasets in other deep learning applications. Hence, when no dropout is applied, the model overfits to the training data and validation performance does not match the training performance. The proposed multi-level dropout helps in efficiently reducing the

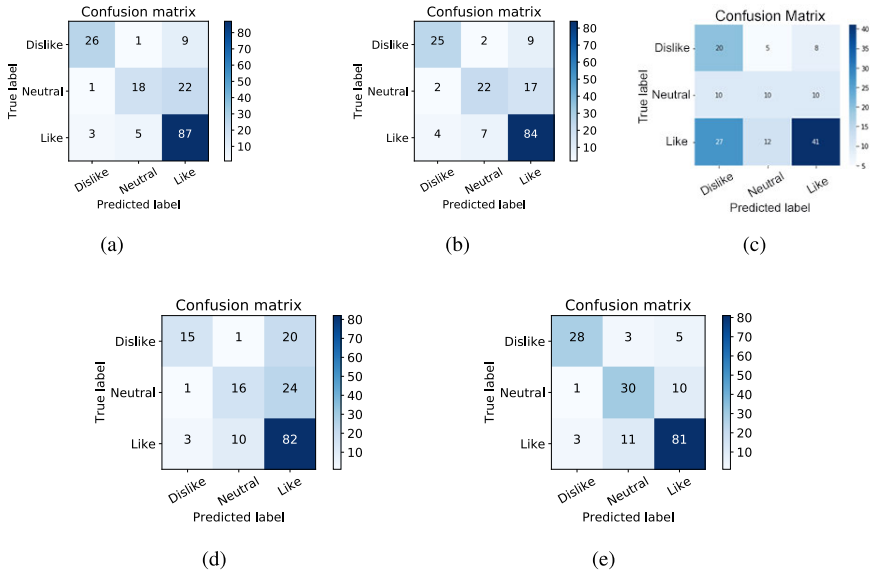

FIGURE 8. Confusion matrices for different models shown in Table 7 (a) Confusion matrix for TCNN-LSTM + Dual-LSTM (b) Confusion matrix for 3D-CNN + Dual-LSTM (c) Confusion matrix for DCNA (without dropout) (d) Confusion matrix for DCNA (with multi-level dropout).

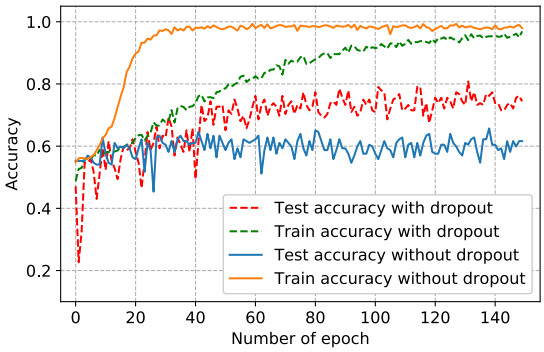

FIGURE 9. Plot of training and validation accuracy for DCNA with and without dropout.

overfitting to training data in limited training data scenarios. Plot of training and test accuracy of DCNA with and without dropout are shown in fig. 9. Plot reveals that model starts to overfit training data in the absence of dropout at early stage of training. It is important to note that data augmentation is used in both of these models.

Confusion matrices for all the four models (shown in Table 7) are shown in fig. 8. Although, both model combinations (TCNN + Dual-LSTM and 3D-CNN + Dual-LSTM) are able to achieve equal test accuracy, confusion matrices of both of these combinations provide in depth information about quality of results achieved. TCNN + Dual-LSTM (refer fig. 8a) has very biased behavior toward 'Like' category, it labels most of the 'Neutral' class samples in positive category along with actual ones. Whereas, 3D-CNN + LSTM model has significantly lower recall rate for 'Neutral' category (refer fig. 8c). Hence, even when test accuracy is same, 3D-CNN based model is able to learn much better feature representation than TCNN based model.

Similarly, confusion matrices for DCNA model with and without multi-level dropout are shown in fig. 8d and fig. 8e, respectively. Without induction of dropout in feature vectors, model is not able to learn the correct feature representations for 'Neutral' and 'Dislike' classes. The higher accuracy on positive class is because of the fact that model classifies any arousal activity on face into positive class, and hence recall rate is very high for 'Neutral' and 'Dislike' classes. Dropout at multiple stage of feature extraction process removes some percentage of compressed feature representation, forcing model to learn more robust features.

We also test proposed model with different kernel counts and optimizers (refer Table 8). The DCNA (low capacity), in Table 8, represents the model with kernel count shown in Table 4 and Table 5 for model TCNN and 3D-CNN, respectively. The model DCNA (medium capacity) and DCNA (high capacity) have 8 and 16 higher kernels in each layer than DCNA (low capacity), respectively. We can observe that increase in capacity of model actually reduces the test accuracy of the model. Intuitive explanation for this behavior is that higher dimensionality of parameter space makes it difficult for gradient descent based optimizers to learn the correct parameter configuration. The proposed model is also tested with different optimizers. RMSprop [54], ADAM [55] and SGD [54] optimizer are able to achieve test accuracy of 80.81%, 75.58% and 72.67%, respectively (refer Table 8).

Plot of validation accuracy of different variants of DCNA without the progress of training is shown in fig. 10. Although all the models show similar pattern in validation accuracy, the empirical selection of different parameters helps

TABLE 8. Study of effect of different architectural variation on performance of DCNA.

Model	parameters	positive	neutral	negative	Average accuracy	train accuracy
DCNA (low capacity)	7,028,371	85.26	73.17	77.77	80.81	96.91
DCNA (medium capacity)	13,443,599	85.26	56.10	75.00	76.16	94.74
DCNA (high capacity)	16,098,383	89.47	56.10	77.78	78.06	96.75
DCNA with ADAM optimizer	7,028,371	83.16	58.54	75.00	75.58	81.58
DCNA with SGD optimizer	7,028,371	84.21	46.34	72.22	72.67	99.08

in improving the model performance. Barchart of the test accuracies achieved by all the test models is shown in fig. 11.

We also analyzed the internal layers of the proposed model to understand the learning characteristics of the model. Activation behavior of different layers of the appearance modeling network (TCNN) and dynamics modeling architecture are shown in fig. 12. Outputs of the 4^{th} layer and 6^{th} layer of 3D-CNN and TCNN models can be observed in the figure. Both the models show higher activation values in the key facial areas like eyes, nose, and lips. However, models are not provided with any specific information to lean the focus on these areas. These results depict that proposed architecture is able to correctly identify the key facial areas and extract facial dynamics corresponding to each category.

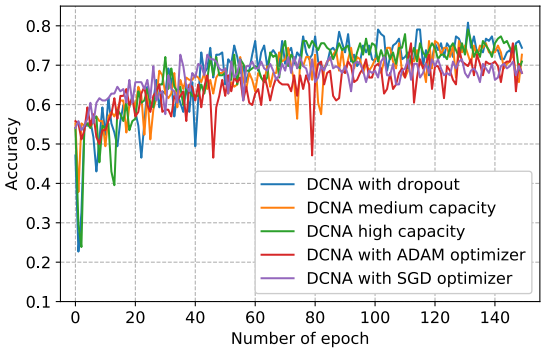

FIGURE 10. Plot of validation accuracy with progress of training for different variants of DCNA.

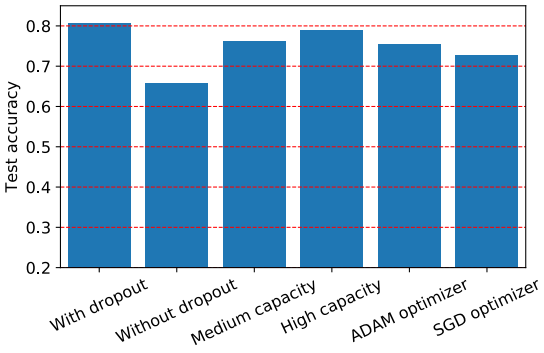

FIGURE 11. Bar chart showing the comparative analysis of test accuracies achieved by different models.

VII. COMPARISON WITH STATE-OF-THE-ART

We used 10 fold cross validation to compare performance of the proposed model against recent state-of-the-art methods. The dataset contains 1395 videos in total. These videos are divided into 10 sets with 139 samples in each. All the models presented in table were trained on nine sets and one evaluated left out set. The final performance of the models was computed by taking the mean over all the sets. The proposed DCNA model achieves mean average validation accuracy of 60.54 with 95% confidence interval between 57.8 and 63.3.

The first state-of-the-art network used for the comparison consists of a CNN based feature extractor network and RNN network [56]. Feature extractor network has three convolution layers. Max Pooling layer is used after first and second convolution layer. For the third convolution layer, gradual pooling is used. This layer is followed by the fully connected layer and the dropout layer. This network is trained separately from the RNN network. The output features obtained from this network are passed as an input to the RNN network to exploit their temporal dependency and obtain the final output. Mean obtained after using this model for training for the validation set is 56.31 which are less as compared to our proposed model. The 95% confidence interval lies between 53.7 and 58.9.

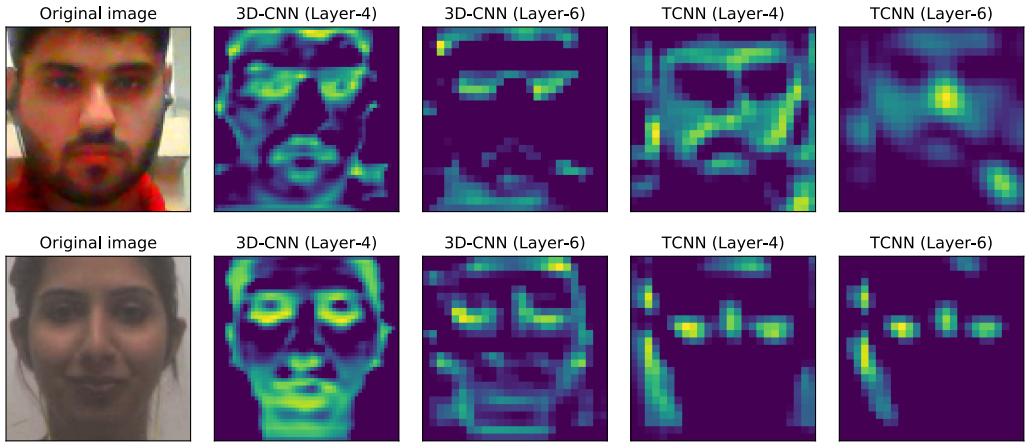

FIGURE 12. Analysis of activity of internal layers of TCNN and 3D-CNN models.

TABLE 9. Comparison of performance of the proposed DCNA model against state-of-the-art in 10-fold cross-validation.

S.No.	DCNA	[56]	[57]	[58]	[59]
1	56.83	51.08	51.08	51.08	51.80
2	63.30	56.12	55.40	55.40	56.83
3	60.43	50.36	50.36	52.36	51.80
4	53.95	53.24	51.25	53.24	53.24
5	66.90	61.87	61.87	61.87	61.87
6	62.58	59.71	59.71	59.41	61.15
7	64.74	61.87	60.77	61.87	61.87
8	55.39	51.80	51.80	51.80	51.80
9	55.93	59.71	60.43	59.71	59.71
10	65.34	57.34	58.34	57.34	59.44
Mean	60.54	56.31	56.10	56.40	56.95

Optimum deep 3D-CNN model is proposed by Haddad *et al.* [57] for facial emotion recognition. Deep 3D-CNN architecture is composed of eight consecutive blocks, where each block consists of a 3D convolution layer followed by batch normalization, LeakyReLU, and 3D maxpooling layer. The output feature maps obtained from the eighth block are fed to fully connected layer followed by dropout and other fully connected layer. The final output is obtained by applying softmax activation function on the feature maps of fully connected layer. For this model, the mean is 56.101 and 95 % confidence interval lies between 53.4 to 58.8 for the 10 fold validation set.

Concept of transfer learning is used by Li *et al.* [58] for facial expression analysis. Firstly, visual information is obtained by training a CNN model on source task with large number of images from the labelled dataset. Then, the last layer of CNN model is removed and flattened into single dimension. To obtain the temporal information, these single-dimension feature maps are given as an input to LSTM. For validation set, the mean value obtained using this model is 56.408 and 95% confidence interval lies between 53.9 to 58.9.

3D-ResNet and it's various extensions are examined by Hara *et al.* [59] to simulate advancement in field of computer vision for videos. We used 3D-ResNet 50 model for evaluating the results. The mean average validation accuracy for this model is 56.951 and the 95% confidence interval lies between 54.4 to 59.5.

VIII. CONCLUSION

The paper presents a novel framework to automatically generate users' response to the viewed commercial multimedia based on their facial expressions. For this experiment, volunteers are recruited to observe the videos of different genre. Facial expressions of these observers are recorded (against different genre of videos) to cover various muscular activities. Subjects are also asked to provide their response, divided into 'Like', 'Neutral' and 'Dislike' categories, against the watched multimedia.

A novel deep composite neural architecture is designed to learn the feature representation from the subject's face for this classification task. The proposed model learns different modalities from the input data using ensemble of three different neural architectures. The 3D convolutional neural network is used to extract spatio-temporal features from the input video. Similarly, long short term memory network is used to model the temporal dependencies amongst the features of different frames extracted by time distributed convolutional neural network. The third model (Dual-LSTM) is used to learn the sequential arousal activity observed in facial landmarks against each category.

Different combinations of all three basic models are studied to understand the learning behavior of each model. Final deep composite neural architecture is able to achieve highest accuracy of 80.81% with a multi-level dropout procedure. We presented the activity of internal layers of different models to provide an insight into learning behavior of these models. The proposed model achieves significantly

better performance against state-of-the-art using 10 fold cross-validation.

REFERENCES

[1] P. Ekman, *Darwin and Facial Expression: A Century of Research in Review*. Sacramento, CA, USA: Ishk, 2006.

[2] S. Tokuno, G. Tsumatori, S. Shono, E. Takei, T. Yamamoto, G. Suzuki, S. Mituyoshi, and M. Shimura, "Usage of emotion recognition in military health care," in *Proc. Defense Sci. Res. Conf. Expo (DSR)*, Aug. 2011, pp. 1–5.

[3] E. Hudlicka and J. Broekens, "Foundations for modelling emotions in game characters: Modelling emotion effects on cognition," in *Proc. 3rd Int. Conf. Affect. Comput. Intell. Interact. Workshops*, Sep. 2009, pp. 1–6.

[4] D. Morrison, R. Wang, L. C. De Silva, and W. L. Xu, "Real-time spoken affect classification and its application in call-centres," in *Proc. 3rd Int. Conf. Inf. Technol. Appl. (ICITA)*, Jul. 2005, pp. 483–487.

[5] A. Tawari and M. M. Trivedi, "Audio visual cues in driver affect characterization: Issues and challenges in developing robust approaches," in *Proc. Int. Joint Conf. Neural Netw.*, Jul. 2011, pp. 2997–3002.

[6] M. Turk, "Multimodal interaction: A review," *Pattern Recognit. Lett.*, vol. 36, pp. 189–195, Jan. 2014.

[7] S. Taheri, P. Turaga, and R. Chellappa, "Towards view-invariant expression analysis using analytic shape manifolds," in *Proc. Face Gesture*, Mar. 2011, pp. 306–313.

[8] S. W. Chew, P. Lucey, S. Lucey, J. Saragih, J. F. Cohn, and S. Sridharan, "Person-independent facial expression detection using constrained local models," in *Proc. Face Gesture*, Mar. 2011, pp. 915–920.

[9] 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 *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2010, pp. 94–101.

[10] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Collecting large, richly annotated facial-expression databases from movies," *IEEE MultiMedia*, vol. 19, no. 3, pp. 34–41, Jul./Sep. 2012.

[11] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D. H. Lee, and Y. Zhou, "Challenges in representation learning: A report on three machine learning contests," in *Proc. Int. Conf. Neural Inf. Process.* Berlin, Germany: Springer, 2013, pp. 117–124.

[12] P. Ekman, "Darwin, deception, and facial expression," *Ann. New York Acad. Sci.*, vol. 1000, no. 1, pp. 205–221, 2003.

[13] S. Zhalehpour, O. Onder, Z. Akhtar, and C. E. Erdem, "BAUM-1: A spontaneous audio-visual face database of affective and mental states," *IEEE Trans. Affective Comput.*, vol. 8, no. 3, pp. 300–313, Jul./Sep. 2016.

[14] H. Monkaresi, N. Bosch, R. A. Calvo, and S. K. D'Mello, "Automated detection of engagement using video-based estimation of facial expressions and heart rate," *IEEE Trans. Affective Comput.*, vol. 8, no. 1, pp. 15–28, Jan./Mar. 2017.

[15] A. Dhall, A. Asthana, R. Goecke, and T. Gedeon, "Emotion recognition using PHOG and LPQ features," in *Proc. Face Gesture*, Mar. 2011, pp. 878–883.

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

[17] X. Huang, Q. He, X. Hong, G. Zhao, and M. Pietikainen, "Improved spatiotemporal local monogenic binary pattern for emotion recognition in the wild," in *Proc. 16th Int. Conf. Multimodal Interact.*, Nov. 2014, pp. 514–520.

[18] Y. Li, S. Wang, Y. Zhao, and Q. Ji, "Simultaneous facial feature tracking and facial expression recognition," *IEEE Trans. Image Process.*, vol. 22, no. 7, pp. 2559–2573, Jul. 2013.

[19] M. Liu, R. Wang, S. Li, S. Shan, Z. Huang, and X. Chen, "Combining multiple kernel methods on Riemannian manifold for emotion recognition in the wild," in *Proc. 16th Int. Conf. Multimodal Interact.*, Nov. 2014, pp. 494–501.

[20] J. Chen, Z. Chen, Z. Chi, and H. Fu, "Emotion recognition in the wild with feature fusion and multiple kernel learning," in *Proc. 16th Int. Conf. Multimodal Interact.*, Nov. 2014, pp. 508–513.

[21] Y.-L. Tian, T. Kanade, and J. F. Cohn, "Recognizing action units for facial expression analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 2, pp. 97–115, Feb. 2001.

[22] G. Brown, J. Wyatt, R. Harris, and X. Yao, "Diversity creation methods: A survey and categorisation," *Inf. Fusion*, vol. 6, no. 1, pp. 5–20, 2005.

[23] M. S. Bartlett, G. C. Littlewort, M. G. Frank, C. Lainscsek, I. R. Fasel, and J. R. Movellan, "Automatic recognition of facial actions in spontaneous expressions," *J. Multimedia*, vol. 1, no. 6, pp. 22–35, 2006.

[24] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Image Vis. Comput.*, vol. 27, no. 6, pp. 803–816, 2009.

[25] M. Pantic and I. Patras, "Dynamics of facial expression: Recognition of facial actions and their temporal segments from face profile image sequences," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 36, no. 2, pp. 433–449, Apr. 2006.

[26] M. F. Valstar and M. Pantic, "Combined support vector machines and hidden Markov models for modeling facial action temporal dynamics," in *Proc. Int. Workshop Hum. Comput. Interact.* Berlin, Germany: Springer, 2007, pp. 118–127.

[27] K. Sikka, T. Wu, J. Susskind, and M. Bartlett, "Exploring bag of words architectures in the facial expression domain," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2012, pp. 250–259.

[28] Y. Rahulamathavan, R. C.-W. Phan, J. A. Chambers, and D. J. Parish, "Facial expression recognition in the encrypted domain based on local Fisher discriminant analysis," *IEEE Trans. Affect. Comput.*, vol. 4, no. 1, pp. 83–92, Jan. 2013.

[29] Z. Yu and C. Zhang, "Image based static facial expression recognition with multiple deep network learning," in *Proc. ACM Int. Conf. Multimodal Interact.*, Nov. 2015, pp. 435–442.

[30] P. Werner, A. Al-Hamadi, K. Limbrecht-Ecklundt, S. Walter, S. Gruss, and H. C. Traue, "Automatic pain assessment with facial activity descriptors," *IEEE Trans. Affect. Comput.*, vol. 8, no. 3, pp. 286–299, Jul./Sep. 2017.

[31] F. Long, T. Wu, J. R. Movellan, M. S. Bartlett, and G. Littlewort, "Learning spatiotemporal features by using independent component analysis with application to facial expression recognition," *Neurocomputing*, vol. 93, pp. 126–132, Sep. 2012.

[32] T. R. Almaev and M. F. Valstar, "Local Gabor binary patterns from three orthogonal planes for automatic facial expression recognition," in *Proc. Humaine Assoc. Conf. Affect. Comput. Intell. Interact.*, Sep. 2013, pp. 356–361.

[33] Y. Fan, X. Lu, D. Li, and Y. Liu, "Video-based emotion recognition using CNN-RNN and C3D hybrid networks," in *Proc. 18th ACM Int. Conf. Multimodal Interact.*, Oct. 2016, pp. 445–450.

[34] S. E. Kahou, X. Bouthillier, P. Lamblin, C. Gulcehre, V. Michalski, K. Konda, S. Jean, P. Froumenty, Y. Dauphin, N. Boulanger-Lewandowski, and R. C. Ferrari, "EmoNets: Multimodal deep learning approaches for emotion recognition in video," *J. Multimodal User Interfaces*, vol. 10, no. 2, pp. 99–111, 2016.

[35] H. Kaya, F. Gürpinar, and A. A. Salah, "Video-based emotion recognition in the wild using deep transfer learning and score fusion," *Image Vis. Comput.*, vol. 65, pp. 66–75, Sep. 2017.

[36] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The Munich versatile and fast open-source audio feature extractor," in *Proc. Int. Conf. Multimedia (MM)*, Oct. 2010, pp. 1459–1462.

[37] J. Han, Z. Zhang, N. Cummins, F. Ringeval, and B. Schuller, "Strength modelling for real-worldautomatic continuous affect recognition from audiovisual signals," *Image Vis. Comput.*, vol. 65, pp. 76–86, Sep. 2017.

[38] J. Atkinson and D. Campos, "Improving BCI-based emotion recognition by combining EEG feature selection and kernel classifiers," *Expert Syst. Appl.*, vol. 47, pp. 35–41, Apr. 2016.

[39] S. Poria, I. Chaturvedi, E. Cambria, and A. Hussain, "Convolutional MKL based multimodal emotion recognition and sentiment analysis," in *Proc. IEEE 16th Int. Conf. Data Mining (ICDM)*, Dec. 2016, pp. 439–448.

[40] W.-L. Zheng and B.-L. Lu, "Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks," *IEEE Trans. Auton. Mental Develop.*, vol. 7, no. 3, pp. 162–175, Sep. 2015.

[41] B. B. Turker, Y. Yemez, T. M. Sezgin, and E. Erzin, "Audio-facial laughter detection in naturalistic dyadic conversations," *IEEE Trans. Affect. Comput.*, vol. 8, no. 4, pp. 534–545, Oct. 2017.

[42] M. Soleymani, M. Pantic, and T. Pun, "Multimodal emotion recognition in response to videos," *IEEE Trans. Affect. Comput.*, vol. 3, no. 2, pp. 211–223, Apr. 2012.

[43] A. C. Micu and J. T. Plummer, "Measurable emotions: How television ads really work: Patterns of reactions to commercials can demonstrate advertising effectiveness," *J. Advertising Res.*, vol. 50, no. 2, pp. 137–153, Jun. 2010.
[44] M. Soleymani, G. Chanel, J. J. Kierkels, and T. Pun, "Affective characterization of movie scenes based on multimedia content analysis and user's physiological emotional responses," in *Proc. 10th IEEE Int. Symp. Multimedia (ISM)*, Jan. 2008, pp. 228–235.
[45] M. Soleymani, G. Chanel, J. J. M. Kierkels, and T. Pun, "Affective characterization of movie scenes based on content analysis and physiological changes," *Int. J. Semantic Comput.*, vol. 3, no. 2, pp. 235–254, Jun. 2009.
[46] D. McDuff, R. El Kaliouby, D. Demirdjian, and R. Picard, "Predicting online media effectiveness based on smile responses gathered over the internet," in *Proc. 10th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, Apr. 2013, pp. 1–7.
[47] D. McDuff, R. E. Kaliouby, J. F. Cohn, and R. W. Picard, "Predicting ad liking and purchase intent: Large-scale analysis of facial responses to ads," *IEEE Trans. Affect. Comput.*, vol. 6, no. 3, pp. 223–235, Jul. 2015.
[48] J. Hernandez, Z. Liu, G. Hulten, D. DeBarr, K. Krum, and Z. Zhang, "Measuring the engagement level of TV viewers," in *Proc. 10th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, Apr. 2013, pp. 1–7.
[49] R. Navarathna, P. Lucey, P. Carr, E. Carter, S. Sridharan, and I. Matthews, "Predicting movie ratings from audience behaviors," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, Mar. 2014, pp. 1058–1065.
[50] *Statista*. Accessed: Aug. 4, 2021. [Online]. Available: https://www.statista.com/statistics/264810/number-of-monthly-active-fac% ebook-users-worldwide/
[51] *Statista*. Accessed: Aug. 4, 2021. [Online]. Available: https://www.statista.com/topics/2019/youtube/
[52] *Omnicoreagency*. Accessed: Aug. 4, 2021. [Online]. Available: https://www.omnicoreagency.com/youtube-statistics/
[53] D. E. King, "Dlib-ml: A machine learning toolkit," *J. Mach. Learn. Res.*, vol. 10, pp. 1755–1758, Jan. 2009.
[54] S. Ruder, "An overview of gradient descent optimization algorithms," 2016, *arXiv:1609.04747*. [Online]. Available: http://arxiv.org/abs/1609.04747
[55] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980
[56] D. Rangulov and M. Fahim, "Emotion recognition on large video dataset based on convolutional feature extractor and recurrent neural network," in *Proc. IEEE 4th Int. Conf. Image Process., Appl. Syst. (IPAS)*, Dec. 2020, pp. 14–20.
[57] J. Haddad, O. Lézoray, and P. Hamel, "3D-CNN for facial emotion recognition in videos," in *Proc. Int. Symp. Vis. Comput.* Cham, Switzerland: Springer, 2020, pp. 298–309.
[58] T.-H.-S. Li, P.-H. Kuo, T.-N. Tsai, and P.-C. Luan, "CNN and LSTM based facial expression analysis model for a humanoid robot," *IEEE Access*, vol. 7, pp. 93998–94011, 2019.
[59] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet?" in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6546–6555.

VIVEK SINGH BAWA is currently a Postdoctoral Researcher with the Visual Artificial Intelligence Laboratory (VAIL), Oxford Brookes University, U.K. He also teaches machine vision and artificial intelligence to bachelor's students at Oxford Brookes University. His research interests include semantic scene segmentation and surgeon action recognition and localization.

SHAILZA SHARMA is currently a Ph.D. Scholar with the Department of Computer Science and Engineering, Thapar Institute of Engineering and Technology, India. Her research interests include deep learning and computer vision.

MOHAMMED USMAN (Senior Member, IEEE) received the Ph.D. degree from the University of Strathclyde, Glasgow, U.K. He is currently an Assistant Professor with King Khalid University, Abha, Saudi Arabia. His research interests include machine learning and communication.

ABHIMAT GUPTA is currently pursuing the bachelor's degree with the Department of Computer Science and Engineering, Thapar Institute of Engineering and Technology, Patiala, India. His research interests include deep learning and computer vision.

VINAY KUMAR is currently an Associate Professor with the Department of Electronics and Communication Engineering, Thapar Institute of Engineering and Technology, Patiala, India. He also works in the area of deep learning and computer vision.