Applied Soft Computing Journal 96 (2020) 106624

Contents lists available at ScienceDirect

Applied Soft Computing Journal

journal homepage: www.elsevier.com/locate/asoc

A unified framework of deep networks for genre classification using movie trailer

Ashima Yadav, Dinesh Kumar Vishwakarma *

Biometric Research Laboratory, Department of Information Technology, Delhi Technological University, Delhi, India

ARTICLE INFO

Article history:
Received 30 March 2020
Received in revised form 4 June 2020
Accepted 1 August 2020
Available online 24 August 2020

Keywords:
Affective computing
Deep learning
Emotions
Inception
Sentiments
Video classification

ABSTRACT

Affective video content analysis has emerged as one of the most challenging and essential research tasks as it aims to analyze the emotions elicited by videos automatically. However, little progress has been achieved in this field due to the enigmatic nature of emotions. This widens the gap between the human affective state and the structure of the video. In this paper, we propose a novel deep affect-based movie trailer classification framework. We also develop an EmoGDB dataset, which contains 100 Bollywood movie trailers annotated with popular movie genres: Action, Comedy, Drama, Horror, Romance, Thriller, and six different types of induced emotions: Anger, Fear, Happy, Neutral, Sad, Surprise. The affect-based features are learned via ILDNet architecture trained on the EmoGDB dataset. Our work aims to analyze the relationship between the emotions elicited by the movie trailers and how they contribute in solving the multi-label genre classification problem. The proposed novel framework is validated by performing cross-dataset testing on three large scale datasets, namely LMTD-9, MMTF-14K, and ML-25M datasets. Extensive experiments show that the proposed algorithm outperforms all the state-of-the-art methods significantly, as reported by the precision, recall, F1 score, precision–recall curves (PRC), and area under the PRC evaluation metrics.

1. Introduction

Movies are one of the most popular means of entertainment for the audience as they elicit a wide range of emotions from humans [1]. Emotion recognition is one of the primary tasks in affective computing. It has been observed that humans get more fascinated and comprehend to the visual data, thus making visual emotion or sentiment detection as a crucial domain for research. However, video-based applications are quite challenging as many established machine learning-based techniques face difficulty in handling them. In this work, we have focused on an affect-based video classification problem called movie genre classification in which the genre of the movie is predicted by analyzing the emotions evoked in the viewer while watching a trailer. Earlier literature has focused on deriving the relationship between media and emotion [2–4]. They have shown that the elements of a film have an emotional impact on the viewers. Since the movie trailers are specially constructed to engage the viewers quickly in less amount of time, hence it can be assumed that movie trailers are an amalgamation of different types of emotions [5]. Despite some efforts made in this domain, the affect-based movie genre classification has certain challenges, which include: (1) The predicted classes (genres) are not explicitly present in the frames of the video. This is because genres are abstract and elusive features which cannot be identified in the movie frame sequence. (2) Further, movie genre classification is a multi-label problem because a trailer can belong to multiple genres at the same time. (3) The enigmatic nature of emotions, and the difficulty in bridging the gap between human affective state and structure of the video, poses yet another challenge in this area. The viewer's emotion after watching a trailer, known as induced emotions are highly subjective and prone to cognitive bias. They may not necessarily be the same as the emotion carried out in the video. Thus, a proper approach needs to be designed, which could map the human emotions to the structure of the movie trailers.

Therefore, we propose a deep affect-based movie genre classification framework which aims to develop a relationship between the induced emotions and movie genre by applying ILDNet architecture. The framework is divided into three stages. In Stage 1, we preprocess the video segments by cropping those frames which have faces in them and ignoring the rest. For this, we develop the *EmoGDB* dataset, which contains 100 Bollywood movie trailers belonging to six popular Bollywood genres. In Stage 2, the ILDNet extracts prominent features from the movie trailers by learning the high-level spatial features via Inception V4 network. To obtain a robust feature vector, we add the layers of Bi-LSTM and LSTM units, which helped in generating an effective feature vector mapping for each of the six emotions corresponding to

* Corresponding author.
E-mail addresses: ashimayadavdtu@gmail.com (A. Yadav), dvishwakarma@gmail.com (D.K. Vishwakarma).

https://doi.org/10.1016/j.asoc.2020.106624
1568-4946/

all the genres. The final feature vector is passed to the softmax classifier to obtain the final probabilities. Lastly, in Stage 3, we use these probabilities to model the relationship between the emotions and the movie genre. In this way, we train our ILDNet architecture for each genre and finally create a stacked ensemble model for classifying new movie trailers into multiple genres. The significant contributions of this paper are as follows:

- We propose a novel deep affect-based movie genre classification framework to obtain discriminative and comprehensive high-level features with a unique combination of Inception V4, Bi-LSTM, and LSTM layers.
- To the best of our knowledge, none of the previous literature has focused on classifying the genre of Bollywood movies related to the Indian cinema. Hence, we make considerable efforts in developing *EmoGDB* (Emotion-based Genre Detection for Bollywood) dataset, which contains 100 Bollywood movie trailers in six popular and distinct genres: Action, Comedy, Drama, Horror, Romance, Thriller. The entire dataset is labeled with six induced emotions: Anger, Fear, Happy, Neutral, Sad, Surprise corresponding to every movie genre.
- This work proposes a novel idea in the field of affect-based video classification, where we contribute by successfully establishing and validating the relationship between psychology and cinematography. The prime advantage of our work is that without watching the entire movie trailer, the architecture can classify the trailer into multiple movie genres.
- The proposed architecture is systematically evaluated by performing cross-datasets testing on three large scale datasets, namely LMTD-9, MMTF-14K, and ML-25M. Experimental results show the proposed framework outperforms all the state-of-the-art and alternative methods, which successfully demonstrates the superior performance and efficacy of our framework.

The remainder of the manuscript is organized as follows: In Section 2, we highlight the background on emotion analysis, review the past literature on movie genre classification and highlight the motivation behind our work. In Section 3, we discuss our proposed EmoGDB dataset. In Section 4, we detail the proposed architecture, and in Section 5, we validate the proposed work by conducting experiments and reporting results on several datasets. Finally, Section 6 concludes the paper.

2. Background and related work

This section gives a background on emotion analysis and briefly review the progress related to emotion content analysis in videos. We further discuss the past literature focused on movie genre classification and highlight the motivation behind our work.

2.1. Emotion analysis

The contents of a video are analyzed broadly at two levels [6]: Cognitive level and affective level. A significant amount of research has already focused on analyzing the video contents at the cognitive level by investigating the structure of a movie [7], capturing different objects of a movie [8], identifying the components of a scene [9]. With the growing popularity of fields like sentiment analysis [10,11] and emotion analysis, researchers are trying the fill the "affective gap" which refers to the discontinuity between the low-level visual features and high-level affective components like human emotions [12,13]. Hence, to fill this gap, psychologists have developed emotion-based models for analyzing the emotional content in the videos.

The psychological theory related to the emotion-based models, which are prevalent in videos were categorized into the *dimensional approach* and *categorical approach* [14]. In the dimensional approach, human emotions were represented by defining their positions in two or three-dimensional spaces such as the valence-arousal model [15] or Plutchik's model [16]. On the other hand, the categorical approach focuses on those human emotions which are different from each other and can be distinguished by facial expressions or biological processes. The famous Ekman's model [17] falls under this category. Many theorists criticized the dimensional approach model for having less predictive power [3] and insufficiency in capturing the differences in emotions [18]. Hence, we have focused on the categorical approach for affective analysis of video in this paper.

The affective content of a video is related to the feeling or emotion embraced in the video, which impacts the affective domain of the viewers. The video affective content analysis is further categorized into direct approaches and indirect approaches [13]. The direct approaches impact the affective domain of the viewers directly from the audio–visual cues of the videos. Majority of previous work has focused on direct approaches. The first research in this field was conducted by Hanjalic et al. [6], who developed a framework to link the 2D valence-arousal model with the low-level features extracted from videos. The results gave a major boost to researchers for developing direct approaches. The low level video features commonly used for movie emotion classification includes motion [19], shot [20], lighting [21], color [21], and camera distance [4]. In order to reduce the semantic gap created by low-level features, researchers were motivated towards mid-level representation for affective analysis of videos. This representation captures audio features like sound related to laughter, horror, and keywords from textual data [22]. Canini et al. [21] developed an approach to generate the relationship between user's affective preferences and mid-level connotative features of the film.

The indirect or implicit video affective analysis focuses on analyzing the viewer's reaction to deduce the emotional content of the video. Past studies have majorly concentrated on physiological signals [23] like heart rate, respiration rate, eye blinking rate, skin temperature, etc. to analyze the emotional experience of the viewers. Still, all these methods require wearable devices, which makes the end user uncomfortable and limits the application of these approaches. Thus, researchers shifted towards analyzing the implicit contents of human behavior like facial expressions and eye movements [24]. However, the emotions displayed by the user may not express their true feelings always. Thus, research needs to focus on developing a proper relationship between the user's expressions and their true feelings.

Recently, multimodal data has gained attention for video emotion content analysis. Yi et al. [25] proposed a novel framework for assigning weights to different modalities and temporal input. The framework contains three layers: statistical layer for robustness, temporal based fusion layers to fuse time-based data, and multimodal fusion to combine different modalities. Noroozi et al. [26] proposed a multimodal emotion recognition system to extract emotions from audio and video modality. The video features were extracted from facial expressions, and audio features include prosodic features. However, the system failed to discriminate against the fear and happy emotion adequately.

2.2. Review of movie genre classification

Movie genre classification also serves as a major factor in building movie based recommender systems. Bansal et al. [27]

applied Latent Semantic Indexing and SVD on the movie tweets extracted from Twitter. The movies were recommended to the user according to the predicted genre. The proposed approach was able to model the semantic features in the movies. Several approaches were proposed, which fused multiple modalities to predict the final genre. Ghaemmaghami et al. [28] performed four class genre classification based on brain signals by showing the correlation between brain signals and audio–visual features in movies. Huang et al. [29] utilized Harmony Search optimization algorithm to select spatial and temporal visual features along with time and frequency domain based audio features for genre classification. Pais et al. [30] performed genre classification on animated movies by fusing text and image modality. Content-based textual features were selected by focusing on crucial words from the synopsis of the film, which provided rich information about its genre. Low-level visual features like color, activity information was extracted from the images. Finally, both the features were fused to get the final predictions. Choroś et al. [31] performed video genre classification by analyzing the sequences of shot lengths. The major drawback of the above approaches is that they are not able to capture the semantics and high-order features of the movie genres precisely.

The growing popularity of deep learning has motivated researchers to develop deep based approaches for movie genre classification. Simoes et al. [32] developed a novel method named CNN-MoTion based on 2D-CNN. The motion features capture the active pixels overtime for every scene in the trailer. However, the method tends to categorize some unsure or generic frames into any random genre, for which the features were not easy to extract. This proves that CNN was unable to capture the context of the images. Wehrmann et al. [33] trained multiple CNN networks to develop the CoNNecT approach, which learns different characteristics of frames. The downside of the approach was its incapability to handle the "Horror" genre. Both the above methods trained their network on single-label classification loss functions, which cannot be scaled up for multi-label classification. Since in the real-world, a movie might belong to multiple genres; hence Wehrmann et al. [34] applied CNN with residual connections for multi-label genre classification of Hollywood movie trailers. The proposed model could extract the temporal relationships between the movie frames by considering '*l*' consecutive frames over time. The model also learned audio-based features by extracting the features from 3-sec audio clips.

Apart from classifying movie genres from their trailers, several researchers have focused on estimating genres from various data like movie plot summaries, synopsis, and posters using deep learning. Ertugrul et al. [35] extracted plot summaries from movies and estimated the genres from each sentence separately. The individual predictions were fused to get the final class label corresponding to each plot summary. Also, bi-directional LSTM was applied to capture the forward and backward context of the input text. The major limitation of plot summaries was that they focused on the starting portion of the main plot and did not consider the entire content of the movie. Similarly, Wehrmann et al. [36] utilized synopsis of the movie for multi-label genre classification by applying attention mechanisms on the word embeddings. These works indicate that deep learning models have shown tremendous growth in extracting the complex, high-level features from the videos. Wi et al. [37] proposed a gram layer on CNN, which extracts the style feature information from the posters. They identified the correlation between the style features of the posters and the movie genre. The major problem faced with poster based genre classification includes the lack of images for many movies that can be regarded as posters. Further, the text-based approaches like plot summaries or synopsis have not been able to take sufficient advantage of the powerful deep learning networks like RNN for modeling the sequential data.

Thus, from the above survey, we find sufficient gaps that need to be addressed for developing an efficient genre classification framework. These gaps serve as the primary motivation behind our work which are as follows:

- The trailers contain rich and diverse features of the movies. They are specially crafted to evoke different types of emotional expressions to engage the viewers. Hence, our proposed framework captures the correlation between the movie genre and the evoked emotions, which will provide feedback to the directors about the viewer's perspective.
- Mostly, all the previous work has developed a single-label, or multi-label genre classification approaches with prime focus on single cinema only. Moreover, no significant research has been done on genre classification for Indian movies. This motivates us to develop a robust framework that could handle the movie genre across different cinemas like Bollywood, Hollywood, Cinema of Japan, Denmark, South Korea, etc. efficiently.
- The literature lacks a proper dataset that could strongly annotate the frames of the movie trailer corresponding to the emotion evoked by it. This would reduce the "affective gap" between the low-level visual features and high-level affective contents like human feelings or emotions. Hence, we develop an *EmoGDB,* which focuses on studying the relationship between the field of psychology and cinematography.
- We believe that the proposed framework will encourage to develop movie recommender system, categorization of movies, TV series, web series, etc. Additionally, this literature may provide new state-of-the-art algorithms to the researchers, academia, and technocrats working in multimedia data analytics and film theory.

3. EmoGDB dataset

To further enhance the research in the area of movie genre detection, especially for Indian cinema, we developed an *EmoGDB* (Emotion-based Genre Detection for Bollywood) dataset, which is specifically related to detecting the genre of Hindi Bollywood movies. The major work which has contributed to study the relationship between the field of psychology and cinematography is [4]. Following their work, we adopt six emotion categories, namely: Happy, Surprise, Anger, Sad, Fear, and Neutral. These are the major emotions that are evoked while watching any movie. The prime advantage of this dataset lies in the fact that it is labeled with five-movie genres along with the six different types of emotions (which are elicited while watching a movie trailer) corresponding to each genre. To the best of our knowledge, no dataset has been developed in the literature that provides such rigorous affective-based information for different movie genres. The trailers have a broad range of release dates (1996 to 2020). Fig. 1 shows some sample images from our dataset (*movie name: genre)*. As seen in the figure, each genre is provoking a wide range of emotions.

We browsed and collected a list of famous Bollywood movies of different genres from four popular film libraries: IMDB, NetFlix, Hotstar, and Amazon Prime. We only focused on those movie trailers which have a common genre on each of these libraries. Our dataset is created and structured to allow the research community to use it with ease. We created one folder per movie. The file structure of the dataset per movie is illustrated in Fig. 2.

Each movie folder contains two sub-folders of uncropped facial frames and cropped facial frames, along with the corresponding movie trailer. The naming format of each folder and sub-folder is shown in Fig. 3. The cropped frame information is stored in the

Fig. 1. Sample images from EmoGDB dataset (a) 1920: Horror (b) Ae Dil Hai Mushkil : Romance (c) Gameover : Thriller (d) Behen Hogi Teri : Comedy (e) Chhichhore : Drama (f) Baahubali 2_The Conclusion : Action.

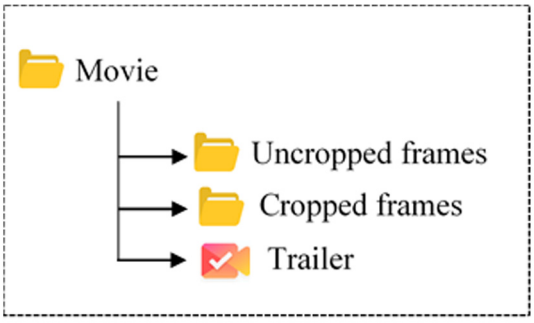

Fig. 2. Folder structure of the proposed dataset.

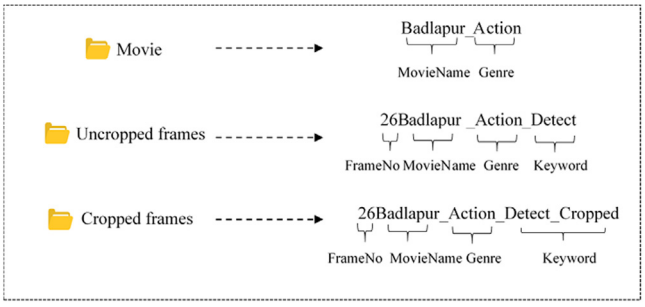

Fig. 3. Naming format rules for folders and sub-folders.

CSV file together with the output labels. The CSV contains the following information: Frame_Name, Movie_Name, Genre, and Emotion. The final trailers belong to the following genres, namely: Action (17), Comedy (16), Drama (17), Horror (17), Romance (16), and Thriller (17). The reason to limit our work to these genres is that mostly all the movies can roughly be classified into at least one of these genres. Secondly, we propose that these genres elicit strong induced emotions, which can be crucial for classifying these movie trailers into multiple genres. EmoGDB dataset consists of roughly over 1,00,000 frames corresponding to 100 Bollywood movie trailers. Since the length of a Bollywood film is very long (up to 2-3 h), hence we concentrated on extracting the features from the movie trailer, which typically has a duration of around 2–4 min.

There may be only one emotion that dominates the entire video, but other emotions could also make interspersed appearances [38], which makes it crucial to analyze the segments of the videos to engender the overall emotions from them. Hence, we start by extracting the frames from the videos. The facial expressions directly depict certain emotions, so we focused on those frames which have faces in them to learn the emotions from the facial expressions. The rest of the frames were ignored. We used OpenCV based Deep learning network, which uses ResNet [39] architecture as a backbone. It is based on the Single Shot Multibox detector [40], which handles the objects of different sizes. The input frame is preprocessed before passing it to the deep network. First, we perform mean subtraction and scaling on the input frames to combat the illumination changes. We have used mean values $(\overline{X})$ of Red (R), Green (G), Blue (B) channels as 124.96, 115.97, and 106.13, respectively. The scaling factor σ defaults to 1. The mean is subtracted and scaling factor is divided from each input channel to get the final R, G, B values as follows:

$$R = R - \overline{X_r} \Rightarrow R = (R - 124.96)/\ 1 \quad (1)$$

$$G = G - \overline{X_g} \Rightarrow G = (G - 115.97)/\ 1 \quad (2)$$

$$B = B - \overline{X_b} \Rightarrow B = (B - 106.13)\ /\ 1 \quad (3)$$

The input is then passed to our OpenCV based deep network which extracts the facial images by calculating the confidence of detection (c_r), to filter out the weak detections. If $c_r > 50\%$, the corresponding face is extracted from the image, and further detections are checked on the same image. The same process is repeated for all the input frames. The advantage of using this method for facial detection is that it works for tiny faces and can handle substantial occlusion in the images. It can detect different face orientations more accurately as compared to other face detection algorithms, as shown in Fig. 4. Tables 1 and 2 gives the list of abbreviations and symbols, respectively, along with their meanings, which have been used in the manuscript.

4. Proposed methodology

In this section, we discuss the proposed deep affect-based movie trailer classification framework in detail. Fig. 5 shows the pipeline of the proposed framework. The major steps are: (i) fetching and preprocessing the frames (ii) extracting the high-level features from the frames with the help of ILDNet architecture (iii) incorporating several emotions to develop a novel

Fig. 4. Different face orientations and occluded images are captured by our face detection algorithm.

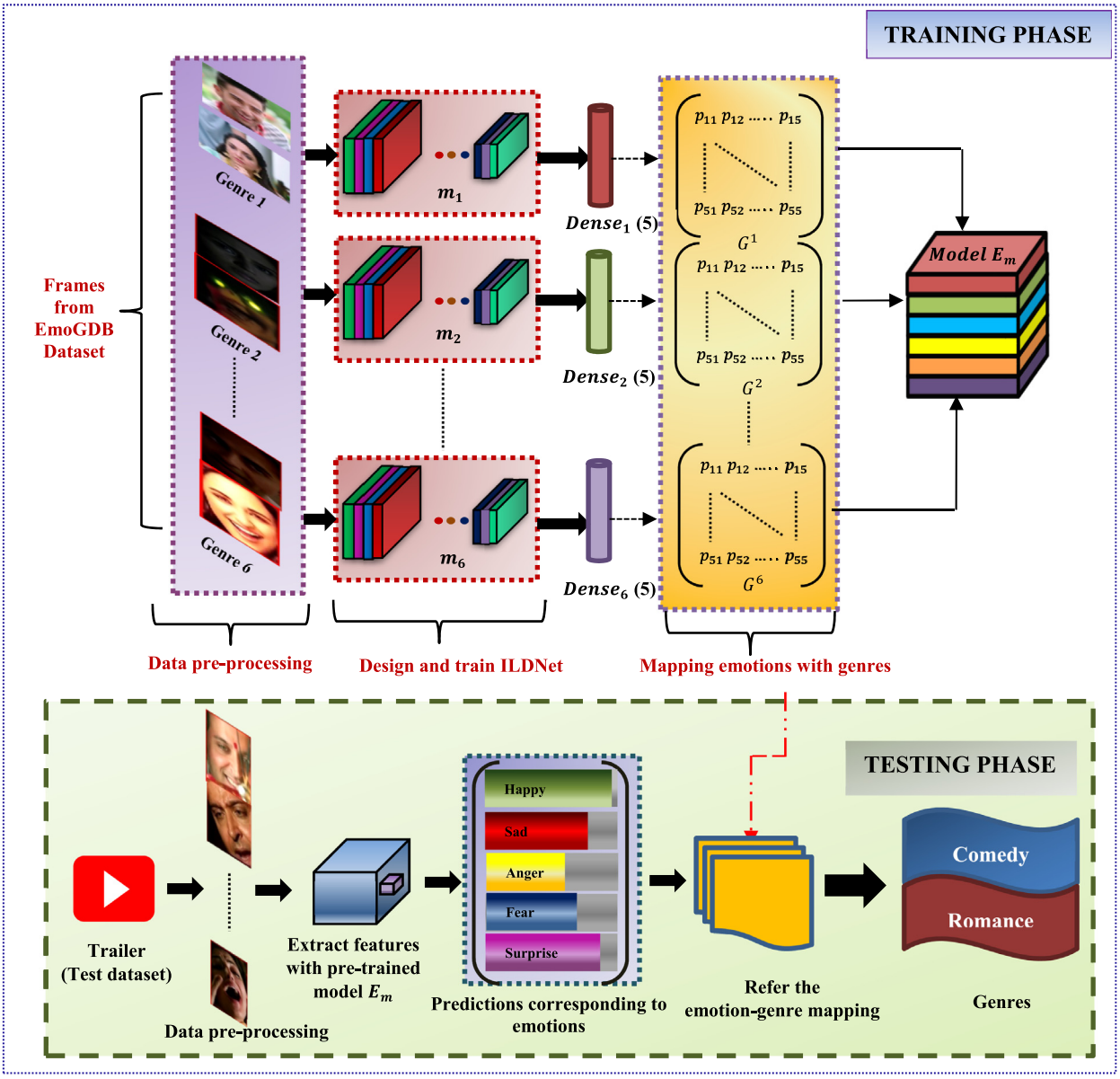

Fig. 5. Pipeline of the proposed framework.

multi-label genre detection theory. The upcoming sections discuss the training and testing phases of the proposed framework with respect to the above steps. To the best of our knowledge, this is the first work that incorporates several emotions for predicting the genre of a movie trailer.

4.1. Data pre-processing

Initially, we process a sequence of video frames from EmoGDB dataset, defined as $X_i^1 = \{x_1^1, x_2^1, x_3^1, \ldots, x_{i+n-1}^1\}$ where x_1^1 denotes the first frame of genre 1. As discussed in Section 3, the face detection algorithm extracts the facial expressions from the input frames and discards the remaining structures. Let $S = \{x_2^1, x_4^1, \ldots, x_{i+n-4}^1, x_{i+n-2}^1\}$ denotes the set of discarded frames from genre 1, then the set of final input sequence $X_f^1 = \{X_i^1 - S\}$. The same steps are repeated for all the remaining five genres, which gives us the training set of $\sum_j X_f^j = \{X_f^1, X_f^2, X_f^3, X_f^4, X_f^5\}$ frames. Moreover, to increase the training set and feed the model with different variants of an image, we perform data augmentation techniques, which include horizontal flipping, zooming,

Table 1
List of Abbreviations with their meanings.

Abbreviations	Meanings
EmoGDB	Emotion-based Genre Detection for Bollywood
ILDNet	Inception-LSTM based Deep Network
LMTD	Labeled Movie Trailer Dataset
MMTF	Multifaceted Movie Trailer Feature
ML	MovieLens
ReLU	Rectified Linear unit
CNN	Convolutional Neural Network
LSTM	Long short term memory
Bi-LSTM	Bi-directional LSTM
PRC	Precision–Recall curve
AU(PRC)	Area under the Precision–Recall curve

Table 2
List of symbols with their meanings.

Symbols	Meanings
X_r, X_g, X_b	Mean of red, green, blue channel, respectively
c_r	Confidence of detection
σ	Scaling factor
x_1^1	First frame of first genre
S	Set of discarded frames
X_f^1	Set of final input frames of first genre
$\sum_j X_f^j$	Complete Training set
$\overrightarrow{h}_t^{(i)}$, $\overleftarrow{h}_t^{(i)}$	Hidden layers for the forward and backward pass, respectively
$\overrightarrow{W}^{(i)}$, $\overleftarrow{W}^{(i)}$	Weights for the forward and backward pass, respectively
$\overrightarrow{b}^{(i)}$, $\overleftarrow{b}^{(i)}$	Bias for the forward and backward pass, respectively
$\hat{y}_t$	Final classification score of LSTM model
p_i	Probability of an ith sample
G^1	Matrix showing the relation between video frames and emotions elicited by them for first genre
$A_{:,:}^{g_1}{}_{e_1}$	Vertical cross-section of all frames in genre g_1 for emotion e_1
E_m	Stacked ensemble model

rescaling, and shearing. The same pre-processing steps are repeated for the test set. The input frames from the test set are fetched to perform normalization, followed by the extraction of facial frames by calculating the confidence of detection, as discussed in Section 3.

4.2. Inception-LSTM based deep network (ILDNet) architecture

The training phase discusses the proposed ILDNet architecture, which extracts the spatial and temporal features from the video frames, and the testing phase applies this pre-trained architecture on the test set. The following section discusses both these phases.

4.2.1. Design and train ILDNet architecture

To design an affect-based theory for genre detection, the crucial features from the input frames $\sum_j X_f^j$, needs to be learned for each genre. This is achieved by developing a deep learning-based feature extraction ILDNet architecture for each genre, which is a combination of Inception V4 [41], Bi-LSTM, and LSTM layers for extracting the spatial and temporal information from the frames. The classification layers of ILDNet architecture are shown in Table 3.

The input shape of 299*299*3 is passed to the Inception V4 block to obtain a high dimensional representation of input in the form of [1*1001] convolution features. The prime motivation of using Inception V4 model is that it processes the information from different scales, which can capture large size variations of the spatial features from the images. This is achieved by performing parallel filter operations on the input from previous layers by using multiple receptive field sizes for convolutions (1*1, 3*3, 5*5). Moreover, the depth of the network is preserved by using 1*1 convolutions before the 3*3 and 5*5 convolutions which prevents to increase the computational cost of the network. This can be visualized in Fig. 6.

As seen in Fig. 6, an image with 112*112*64 *(h*w*d)* dimensions is convolved using 1*1 convolutions 32 times for projecting the depth to lower sizes resulting in 112*112*32 image. This reduces the depth of an image from 64 to 32, without changing the spatial dimensions. In this way, the inception module can utilize multiple smaller convolution kernels, which could capture the local information in the image without blowing up the computational complexity significantly. The final feature vector of [1*1001] is transformed linearly and passed into the Bi-LSTM layer with 512-dimensional hidden state to model the correlation of images over time. It exploits the past and future dependencies for a given prediction.

LSTM belongs to the category of Recurrent Neural Network (RNN), which are popular deep networks for modeling the sequential data. However, they suffer from the vanishing gradient problem [42]. Hence, [43] addressed this problem by developing the LSTM architecture. Thereafter, several variants were proposed by refining the classical LSTM architecture. The core idea behind these variants is that they contain a memory cell that is capable of maintaining the cell's state over time, along with a non-linear gating mechanism to control the flow of information to and fro of the cell. We discuss some of the popular variants of LSTM by highlighting their fundamental properties. For detailed knowledge regarding these networks, we refer [44] to the interested reader, which provides a comparative analysis on all the networks.

(1) Vanilla LSTM [43]: This classic version of LSTM includes only the input and output gates. The input gate decides what information should be added to the cell state based on the current input. The output gate decides which part of the cell state should be passed on as the output. Hence, these networks can remember information over a long period of time by passing the information from each gate's input state to the next state.

(2) LSTM with forget gate [45]: The next popular variant of LSTM includes the forget gate, which controls the extent to which the value of the old cell's state will be discarded based on the current input.

(3) LSTM with peephole connections [46]: In this variant, peephole connections were added by which the cell can control the gates. This is achieved by adding some direct connections from the cell's state to all the different gates.

(4) Gated Recurrent Unit (GRU) [47]: They combine the input gate and forget gate into a single update gate. They did not use

Table 3
Parameters of ILDNet architecture.

Layer Name : Type	Description	Input Shape	Output Shape	Parameters #
Input_1 : Input Layer	Input image of 299*299	(299, 299, 3)	(299, 299, 3)	0
Inception_v4 : CNN	Inception_v4 block	(299, 299, 3)	(1,1001)	1538537
bidirectional_1 : Bidirectional LSTM	Bidirectional LSTM with 512 units	(1, 1001)	(1,1024)	6201344
lstm_2 : LSTM	LSTM layer with 128 units	(1, 1024)	(1,128)	590336
dropout_2 : Dropout	Dropout with 0.5	(1,128)	(1,128)	0
activation_150 : Activation	ReLu activation function	(1,128)	(1,128)	0
my_dense_1 : Dense	Dense Layer with 5 neurons	(1,128)	(1,5)	1161

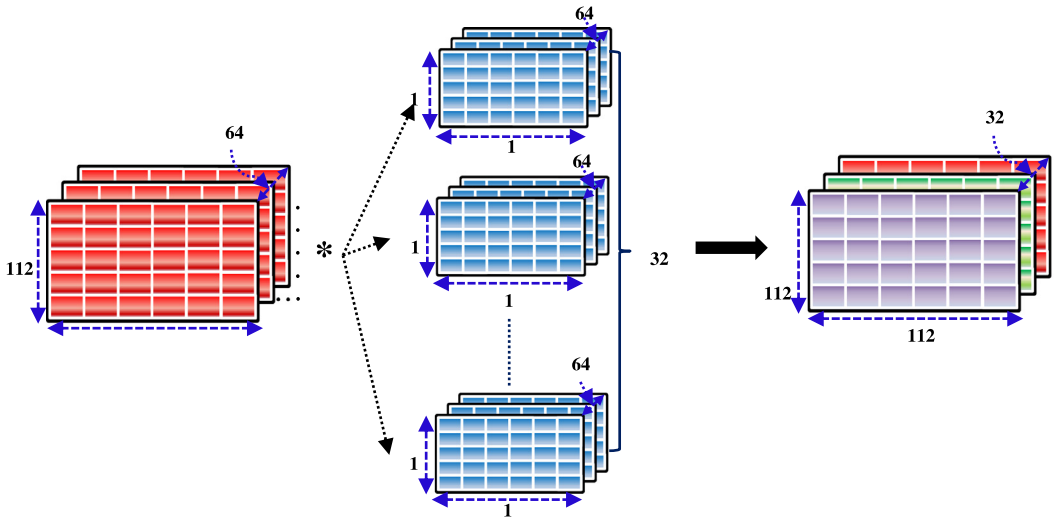

Fig. 6. Applying 1*1 convolutions for reducing the depth of an image without altering the spatial dimensions.

the peephole connections or an output activation function. Additionally, their output gate was changed into the reset gate, which decides how much past information the network will forget.

(5) Bi-directional RNN (Bi-RNN or BRNN) [48]: They are made up of two RNN units connected in the opposite directions, i.e., going from left to right (forward) and other going from right to left (backward), simultaneously. In this way, these networks can easily preserve the information from both the past and future. The units in Bi-RNN could be standard RNN, LSTM, or GRU cells.

Apart from these architectures, other variants of LSTM includes Depth based gated LSTM [49], clockwise RNN [50], full gradient version LSTM [51], etc. We have used Bi-directional LSTM (Bi-LSTM) in our work. As discussed, in Bi-LSTM, the network maintains two hidden states, one for the left to right (forward) propagation and other for the right to left (backward) propagation. The output from layer $(i-1)$ at timestep t becomes the input to the next neuron at layer i. Let x_t^i be the input to the ith layer at instance t, the flow of information from the Bi-LSTM layer is depicted in Eq. (4) to (6):

$$\overrightarrow{h}_t^{(i)} = \sigma\left(\overrightarrow{W}^{(i)} x_t^{(i-1)} + \overrightarrow{U}^{(i)} h_{t-1}^{(i)} + \overrightarrow{b}^{(i)}\right) \tag{4}$$

$$\overleftarrow{h}_t^{(i)} = \sigma\left(\overleftarrow{W}^{(i)} x_t^{(i-1)} + \overleftarrow{U}^{(i)} h_{t+1}^{(i)} + \overleftarrow{b}^{(i)}\right) \tag{5}$$

$$\hat{y}_t = g(W_y\left[\overrightarrow{h}_t^{(i)} + \overleftarrow{h}_t^{(i)}\right] + b_y) \tag{6}$$

where $\overrightarrow{h}_t^{(i)}$, $\overleftarrow{h}_t^{(i)}$ are the hidden layers for the forward and backward pass, respectively. $\overrightarrow{W}^{(i)}$, $\overrightarrow{U}^{(i)}$, $\overleftarrow{W}^{(i)}$, $\overleftarrow{U}^{(i)}$ are the weights and $\overrightarrow{b}^{(i)}$, $\overleftarrow{b}^{(i)}$ are bias. The final classification score $\hat{y}_t$ is calculated by combining the scores of both the hidden layers. The obtained feature vector of [1*1024] from Bi-LSTM is passed into the LSTM unit with 128-dimensional hidden state. In this way, the temporal information of the images is modeled over time, and the model can learn the temporal relations among the frames. Finally, we add a combination of dropout layer (0.5 rate) and ReLU (Rectified Linear unit) activation to overcome the overfitting problem, followed by a dense layer with same number of neurons as the number of classes in the datasets. The softmax layer is attached in the end, which gives the probability distribution of size [1 * 5]. The probability of an ith sample is given in Eq. (7):

$$p_i = \frac{e^{s_i}}{\sum_{j=1}^{n} e^{s_j}} \tag{7}$$

Where, s_i denotes the score = $f(x_i, w)$ for ith sample. Thus, in the training phase, we train six different genre models namely m_1, m_2, m_3, m_4, m_5, m_6 with their corresponding images from the *EmoGDB* dataset. Finally, we concatenate these models to develop a stacked ensemble model E_m, as shown in Fig. 5 (Training Phase). We use categorical cross-entropy loss as our objective function.

4.2.2. Applying the pre-trained model E_m on test set

The pre-trained stacked ensemble model E_m is used for testing the new movie trailers and generate the probability predictions for different emotions evoked from the test set frames.

4.3. Emotion-genre based theory

This section discusses how the emotion-genre based theory is developed in the training phase and referred by the test set for multi-label genre classification.

4.3.1. Mapping emotions with movie genres

As discussed in Section 4.2.1, the softmax classifier outputs the probability distributions for each emotion class corresponding to every genre. Thus, if the trailer of first genre (e.g., action) is passed to our ILDNet architecture which is trained on EmoGDB dataset, the softmax classifier will give a probability distribution $p_1, p_2, \ldots, p_5$ for each of the five emotion classes. The same process will be repeated for every input frame $f_1, f_2, \ldots, f_n$ to yield the total emotions expressed by the movie trailer. This generates a 2*2 matrix, which shows the relationship between the frames of the video and emotions elicited by them. Mathematically, this can be represented in the form of a matrix for the first genre, as given in Eq. (8):

$$G^1 = \begin{array}{c} \\ f_1 \\ f_2 \\ \\ f_n \\ \\ \\ \end{array} \begin{array}{c} \begin{array}{ccccccc} e_1 & & e_2 & \ldots & \ldots & \ldots & e_5 \end{array} \\ \begin{pmatrix} p_{11} & p_{12} & \ldots & \ldots & \ldots & \ldots & p_{15} \\ p_{21} & p_{22} & \ldots & \ldots & \ldots & \ldots & p_{25} \\ \vdots & & & & & & \\ p_{n1} & p_{n2} & \ldots & \ldots & \ldots & \ldots & p_{n5} \end{pmatrix} \\ \begin{array}{cc} A_{:,e_1}^{g_1} & \qquad A_{:,e_5}^{g_1} \end{array} \end{array} \tag{8}$$

where e_1, e_2, e_3, e_4, e_5 represents the five emotions (Anger, Fear, Happy, Joy, and Surprise), $f_1, f_2, \ldots, f_n$ are the processed

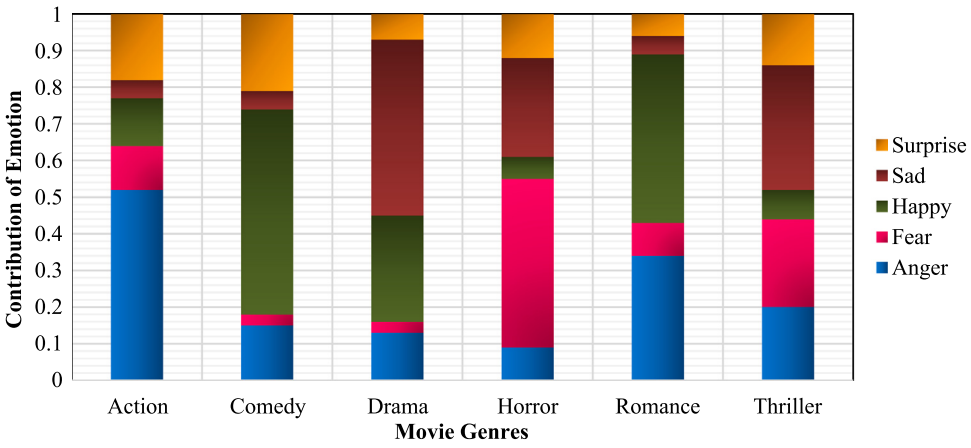

Fig. 7. Emotion-genre mapping for classifying the movie-trailer into multiple genres.

frames from the *EmoGDB* dataset, $A^{g_1}_{:,\,e_1}$ represents the vertical cross-section of g_1 for emotion e_1. Thus, p_{11} indicates the amount of emotion e_1 (let us say, surprise) expressed in frame f_1 and so on. Similarly, $A^{g_1}_{:,\,e_1}$ indicates the combined prediction of surprise emotion for all the frames of the action movie trailer.

We have removed the neutral emotion for training the ILDNet architecture because experimental results in Section 5.2.2 show that it increases the misclassification rate. Hence, the mean probabilities (predictions) of the five emotions for kth genre is represented as in Eq. (9):

$$G^k = \{A^{g_k}_{:,\,e_1},\ A^{g_k}_{:,\,e_2}, A^{g_k}_{:,\,e_3}, A^{g_k}_{:,\,e_4}, A^{g_k}_{:,\,e_5}\} \tag{9}$$

In the training phase, the same process is repeated for the remaining five genres to get the probabilities corresponding to each emotion. Based on the above predictions (G^k), we establish a relationship between emotions evoked while watching a trailer and its corresponding genre. Thus, for action genre Eq. (10) will be represented as:

$$G^{action} = \{A^{g_{action}}_{:,\,surprise},\ A^{g_{action}}_{:,\,sad}, A^{g_{action}}_{:,\,happy}, A^{g_{action}}_{:,\,fear}, A^{g_{action}}_{:,\,anger}\} \tag{10}$$

or

$$G^{action} = \{0.18,\ 0.05,\ 0.13,\ 0.12,\ 0.52\} \tag{11}$$

Similar procedure is adopted for the remaining genres. These results are visualized in Fig. 7, which shows how much each of the above five emotions contributes to different movie genres. Hence, from Fig. 7, we can conclude that emotions play a crucial rule in detecting the multiple genres of a movie trailer. The dominant emotions in action, comedy, drama, horror, romance, and thriller genres are anger, happy, sad, fear, happy, and sad, respectively. The results are quite intuitive. However, we observe that the thriller genre mostly involves sad and fear emotions.

4.3.2. Predicting movie genres

The testing phase involves applying this emotion-genre based mapping theory for calculating the genres corresponding to the emotions predicted in Section 4.2.2. This outputs all the genres for the input trailers. The steps discussed above in the training and testing phase of our proposed framework are summarized in Algorithm 1 and Algorithm 2, respectively.

5. Experiments

In this section, we validate our proposed framework by evaluating it on several datasets and reporting the classification results in terms of several evaluation metrics. Finally, we discuss the computational complexity of our model, followed by implementing the class activation maps to visualize the prominent image regions captured by our architecture.

5.1. Datasets

We validate the proposed ILDNet architecture by performing cross-datasets testing on three publicly available large-scale datasets, namely LMTD-9 (*EmoGDB → LMTD-9),* MMTF-14K (*EmoGDB → MMTF-14K),* and ML-25M (*EmoGDB → ML-25M)* datasets. The complete dataset details are as follows:

5.1.1. LMTD-9 (labeled movie trailer) dataset

LMTD [32,33] is one of the large-scale datasets for movie trailer based genre classification. It consists of 10k movie trailers from 22 different genres. However, for multilabel classification, we consider its subset, LMTD-9 [52], which includes around 4k movie trailers from 9 genres. LMTD-9 removes the trailers which were released before 1980 and contains more than 6500 frames. Since our work focuses on six prominent genres, hence we consider the movie trailers from six genres, namely: Action (853), Comedy (1558), Drama (2023), Horror (435), Romance (649), and Thriller (692). Each of the movie trailers is assigned to at least one and at most three genres. This dataset is challenging because it contains high variability of video features, which includes image quality, aspect ratio, and total length.

5.1.2. MMTF-14K (multifaceted movie trailer feature) dataset

MMTF-14K [53] dataset contains 13,623 Hollywood movie trailers links from 18 different genres. This dataset is primarily used for developing content-based unimodal and multimodal recommender systems. Hence, it addresses three descriptors, namely: Metadata descriptors (Genre features and Tag features), audio descriptors (Block level and I-Vector features), and video descriptors (Aesthetic and AlexNet features). These descriptors help the MMTF-14K dataset to support other multimedia tasks like multilabel genre classification, tag prediction, popularity prediction. We use this dataset for multilabel genre classification tasks and focus on movie trailers from six genres, as discussed above. The Youtube link of the Hollywood movie trailers is parsed to download the trailers. Since some of the trailers are not available on Youtube now, hence we are left with 8674 movie trailers. Each of the movie trailers is assigned to at least one and at most five genres. Hence, this dataset supports the five-class genre classification.

5.1.3. ML (MovieLens) 25M dataset

ML-25M [54] is a benchmark dataset for movie recommender systems. It contains star ratings and text tagging activity from MovieLens.[1] Moreover, this dataset contains 62,423 movie trailer

[1] https://movielens.org.

Algorithm 1 Training phase of deep affect-based movie trailer classification framework

Input: Sequence of movie trailer frames from *EmoGDB* dataset for all the six genres.
Output: Stacked ensemble model (E_m) along with mean probability predictions of each emotion corresponding to the six genres.
1. Data preprocessing:
1.1 Fetch the training frames for Genre 1.
1.2 Perform various data argumentation techniques to increase the training set.
2. Design and train ILDNet architecture:
2.1 Extract the spatial and temporal features from the input frames by ILDNet architecture.
2.2 Apply the softmax classifier to get the final probability predictions.
3. Repeat steps 1 and 2 for all the six genres. This gives us six different models (one for each genre), namely $m_1, m_2, m_3, m_4, m_5, m_6$.
4. Concatenate the above models to generate a stacked ensemble model, E_m.
5. Mapping emotions with movie genres:
5.1 Get the probability predictions from Step 2.2.
5.2 Compute the vertical cross-section of genre 1 with each of the five emotions (neutral emotion is discarded), to generate the affect-based genre predictions.
5.3 Repeat steps 5.1 and 5.2 for rest of the genres to create the affect-based genre mapping..

Algorithm 2 Testing phase of deep affect-based movie trailer classification framework

Input: Sequence of movie trailers from different test datasets.
Output: Multilabel genre predictions corresponding to the movie trailers.
1. Data preprocessing:
1.1 Extract the input frames from the movie trailers.
1.2. Normalize the input frames by performing mean subtraction and scaling.
1.3. Calculate the confidence of detection (c_r) for each input frame.
1.4. If (c_r) > 50%, extract the face from the input frame to build the test set for our model corresponding to the first trailer, else discard the frame.
2. Apply pre-trained E_m on test set:
2.1 Use model E_m (step 4 of Algo 1) to generate the probability predictions for different emotions evoked from the test set frames.
3. Predicting movie genres:
3.1 Apply the affect-based genre theory developed in the training phase for calculating the genres corresponding to the emotions predicted in Step 2.1 above.
3.2 Output the multiple genres for the input trailer.
4. Repeat the above steps for all the trailers in the test set.

IDs, where each trailer is assigned three different ID each from MovieLens, MovieDB,[2] and IMDB.[3] We crawl and download the trailers from the IMDB website only. Since this dataset contains trailers from the year 1900–2019, some of the trailers were not available on all the three platforms (MovieLens, MovieDB, IMDB). Moreover, we focus on multi-label classification for six genres only. Hence, we discard the trailer from other genres. Thus, the total movie trailers available for testing are 18,150. Each of the movie trailers is assigned to at least one and at most five genres. Hence, this dataset also supports the five-class genre classification. Further, this dataset contains a wide range of trailers from different cinemas like the Cinema of Denmark, Cinema of U.S.A, Cinema of India, Cinema of Japan, Cinema of South Korea, etc.

5.2. Experiment setup

In this section, we give the implementation details and discuss the classification results along with baseline comparison on several state-of-the-art and alternate methods.

5.2.1. Implementation details

We build and implement the proposed architecture in Python on popular deep learning framework Keras using Tensorflow backend. All the experiments were performed on Windows 10, 64-bit machine with 128 GB RAM using NVIDIA Titan RTX GPUs. Adam optimizer is used with default parameters of $\beta_1 = 0.9$ and $\beta_2 = 0.999$, learning rate $= 0.001$, and batch size of 64. To feed our model with more amount of data, we perform data augmentation by using different augmentation techniques: horizontal flipping, zooming, rescaling, and shearing. In the training phase, the input samples for each of the six different genre models are split into 80% training, 5% validation, and 15% testing samples. The model is trained for 200 epochs with early-stopping when the validation accuracy does not improve for 20 consecutive epochs. The model converges within 80 epochs. The model achieving the highest validation accuracy is picked as the best-trained model. This model gives the probability distribution for generating the emotion-genre based mapping for each genre from the 15% testing samples of the *EmoGDB dataset.* In the testing phase, we show the generalizability of the proposed architecture by performing cross-datasets testing on the above-discussed test datasets.

[2] https://www.themoviedb.org.
[3] http://www.imdb.com.

Table 4
Classification results of ILDNet on LMTD-9, MMTF-14K, and ML-25M datasets (P: Precision, R: Recall, F1: F1 score).

Dataset	Action			Comedy			Drama			Horror			Romance			Thriller		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
LMTD-9	80.7	88.2	84.2	95.0	87.1	90.9	92.1	87.6	89.7	62.0	77.0	68.6	78.6	90.7	84.1	85.0	78.3	81.5
MMTF-14K	69.4	79.6	74.1	92.7	82.5	87.3	89.1	86.3	87.6	76.6	86.2	81.0	76.9	78.3	77.6	78.2	80.8	79.4
ML-25M	57.9	73.6	63.2	81.0	89.8	85.1	83.8	84.8	84.3	89.0	78.2	83.1	48.3	71.9	57.7	69.3	87.0	77.0

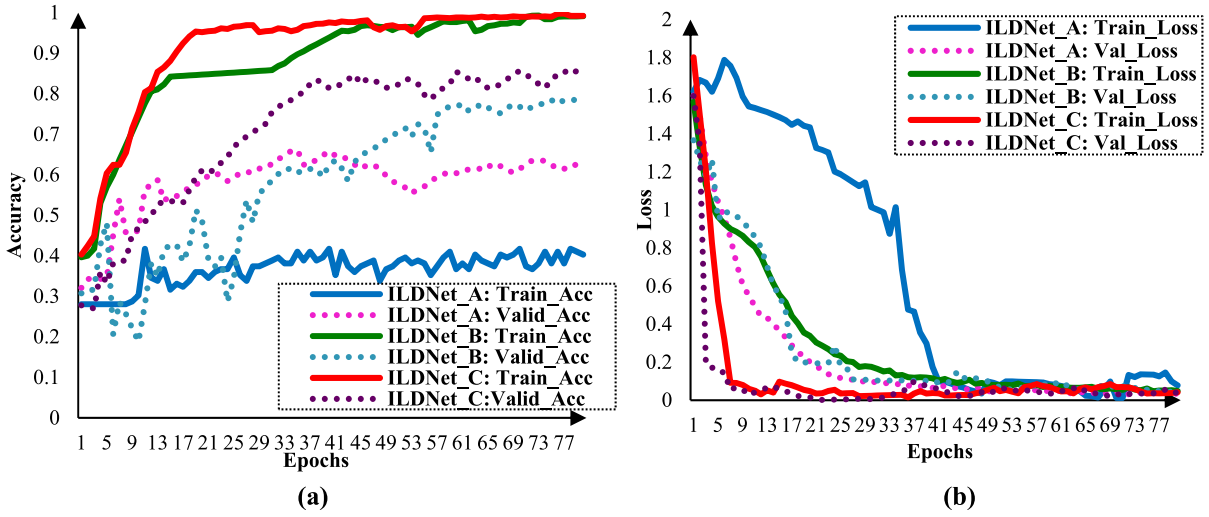

Fig. 8. Evaluating the performance of different variations of ILDNet architecture with (a) Training accuracy + Validation accuracy curves (b) Training Loss + Validation Loss curves.

5.2.2. Classification results

Fig. 8(a) shows the training accuracy + validation accuracy and (b) training loss + validation loss curves to evaluate the training and validation process of ILDNet architecture. We perform experiments with three variations (ILDNet_A, ILDNet_B, and ILDNet_C) within the ILDNet architecture to show the contribution of the major layers in our architecture:

(a) **ILDNet_A**: This model is composed of only the Inception_v4 module.
(b) **ILDNet_B**: This model is composed of Inception_v4 module + Bi-directional LSTM
(c) **ILDNet_C**: This is our proposed model which is composed of Inception_v4 module + Bi-directional LSTM + LSTM + Dropout + Activation.

All the above three models are trained and validated on the proposed*EmoGDB* dataset. The model achieving the best validation results is selected as the final model for cross-dataset testing on other test datasets. From Fig. 8, we see that the ILDNet_A model shows the sign of underfitting as the model is not able to learn the features from the input data (stagnant training accuracy) and still achieves validation accuracy of around 60%. The same behavior can be analyzed from the loss curves of the ILDNet_A model. The second model, ILDNet_B, is more complex than the first one, as the combination of Inception_v4 block and Bi-LSTM can learn the input features properly, thus removes the underfitting issue of ILDNet_A model. Finally, the best results are shown by the ILDNet_C model, which confirms the adequate learning and validating aspect of the model. This shows the importance of LSTM layers for modeling the temporal information in the input.

We perform extensive experiments to evaluate the performance of the proposed ILDNet on LMTD-9, MMTF-14K, and ML-25M datasets. Table 4 shows the classification results in terms of precision, recall, F1 score for each of the output genre classes in the test datasets. The combined accuracy of ILDNet architecture on LMTD-9, MMTF-14, and ML-25M is 86.15%, 83.06%, and 85.3% respectively. From Table 4, we can conclude that the architecture can successfully classify the different genre classes and performs well on all the genres. The best results are seen in Comedy and Drama genres. We also show the confusion matrix in Fig. 9 for all the three test datasets.

We notice that misclassification occurs only between comedy and romance genre along with Horror and Thriller genre. We see many comedy frames were being misclassified as romance frames and vice-versa. Similar behavior is observed in Horror and Thriller genres. The intuition behind this is that the comedy and romance genres are majorly dominated by Happy emotions, whereas the Horror and Thriller genres are dominated by fear and sad emotions. This results in the high inter-class similarity between the respective genre classes as the frame features are not easily distinguishable among them. Despite this, the results indicate that our architecture can correctly recognize and classify different movie genres with great performance. Since our proposed *EmoGDB* dataset has six emotions corresponding to every genre, hence we initially experimented by training ILDNet on all the six emotions, namely: Anger, Fear, Happy, Neutral, Sad, and Surprise. However, we found that a majority of frames were getting misclassified into the neutral category, thus increasing the misclassification results in the test datasets. We also found that this category is not conveying any vital information about the genre of the movie. As an example, we show the confusion matrix for LMTD-9 dataset in Fig. 9(d), which is computed by training ILDNet on all the six emotions. From the figure, it is evident that neutral emotion is increasing the misclassification results and decreases the performance of the classifier. Hence, we remove the neutral emotion for training the ILDNet architecture.

5.2.3. Baseline comparison

We compare the performance of proposed architecture with previous movie genre classification methods on LMTD-9, MMTF-14K, and ML-25M datasets, as shown in Table 5. Since the datasets

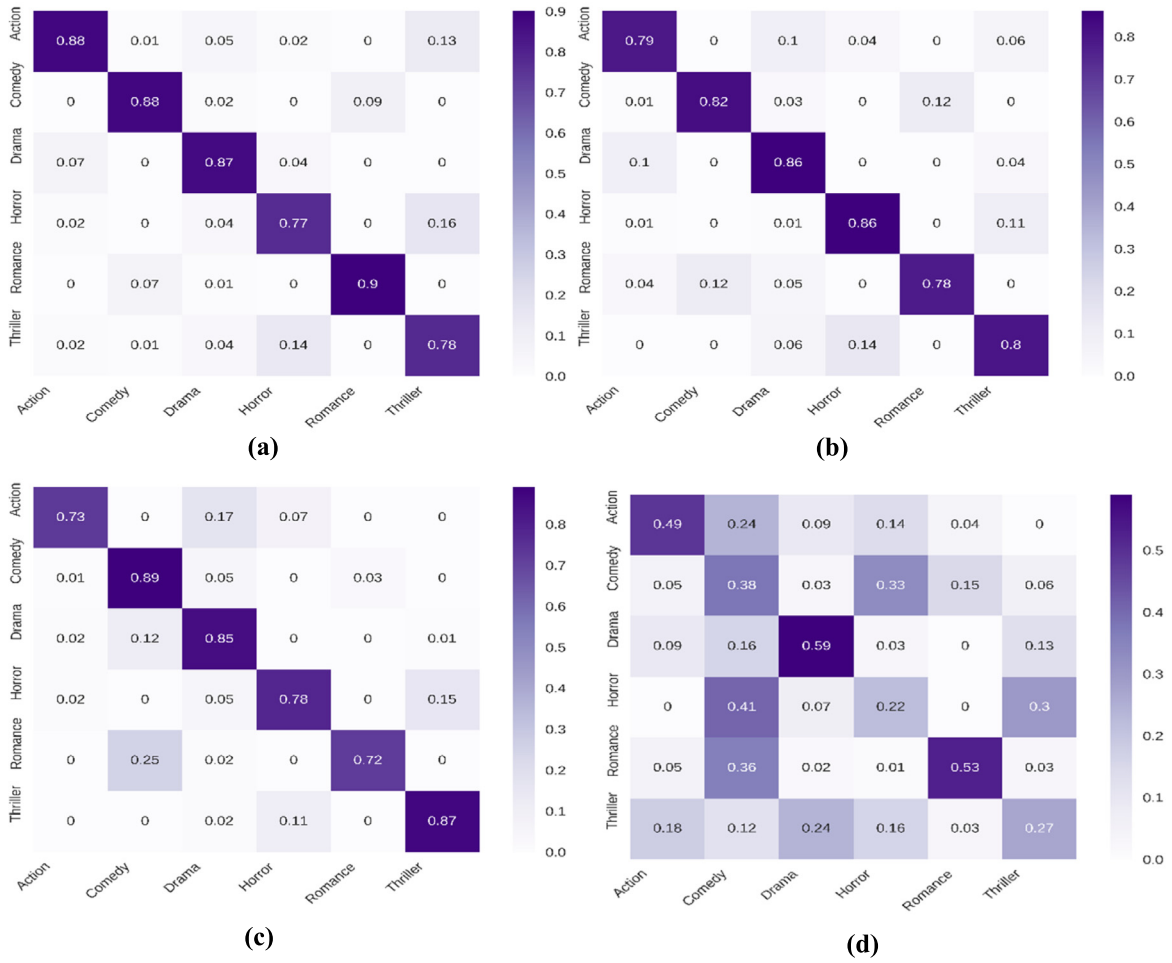

Fig. 9. Confusion matrix for (a) LMTD-9 (b) MMTF-14K (c) ML-25M dataset generated with model trained on five emotions (d) Confusion matrix for LMTD-9 dataset generated with model.

Table 5
Comparison of ILDNet architecture with previous works using micro-average AU(PRC) metric.

Method	Dataset		
	(*EmoGDB* → *LMTD-9*)	(*EmoGDB* → *MMTF-14K*)	*EmoGDB* → *ML-25M)*
Low-level + SVM [8]	0.31	0.29	0.16
GIST + KNN [55]	0.46	0.48	0.31
CENTRIST + KNN [55]	0.49	0.55	0.39
w-CENTRIST + KNN [55]	0.48	0.59	0.53
CoNNeCT [33]	0.78	0.60	0.57
CNN-Motion [32]	0.41	0.46	0.44
CTT-MMC-C [52]	0.62	0.58	0.66
CTT-MMC-TN [34]	0.76	0.81	0.73
ILDNet	**0.81**	**0.94**	**0.89**

have imbalanced classes and vary in size, we validate the performance of our multilabel classifier by comparing the Area under the Precision–Recall curve *AU(PRC)* for each genre class. We combine the contribution of all the genre classes to calculate the micro-average scores for each dataset. This measure can adequately capture the noise resulting from the class imbalance problem in multilabel classification. The final performance across the datasets is evaluated by comparing the micro-average AU(PRC) metric, which is a stricter measure for validating a multilabel classification problem.

As seen in Table 5, we validate our work with state-of-the-art methods and several alternate approaches for movie genre classification. For extracting the low-level features, we compute the four video features as described in [8], namely, average shot length, color variance, motion content, and lighting key, followed by SVM classification. We extract the GIST, CENTRIST, and w-CENTRIST feature descriptors from keyframes of the trailers with the same parameters as discussed in [55] for each of the six movie genres. These are the state-of-the-art methods in low-level feature extraction for movie genre classification. For building CoNNeCT [33] architecture, we combined five different ConvNets models, each one designed to capture various features of the videos. The models are same as in [33], except that we apply GoogleNet architecture, which we pretrain on the LMTD-9 dataset. The CNN-Motion-S [32], extracts the video features using CNN architectures based on [56] and MFCC audio features. The CTT-MMC-C [52] extracts the video features using [39] and applies 2D convolution on them. Similarly, CTT-MMC-TN [21]

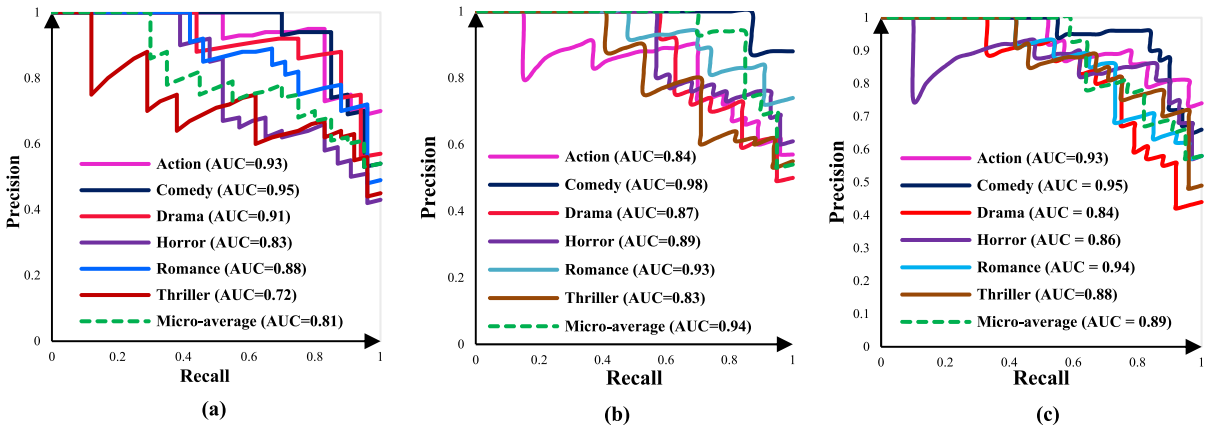

Fig. 10. Precision–Recall curves for (a) LMTD-9 (b) MMTF-14K, and (c) ML-25M datasets.

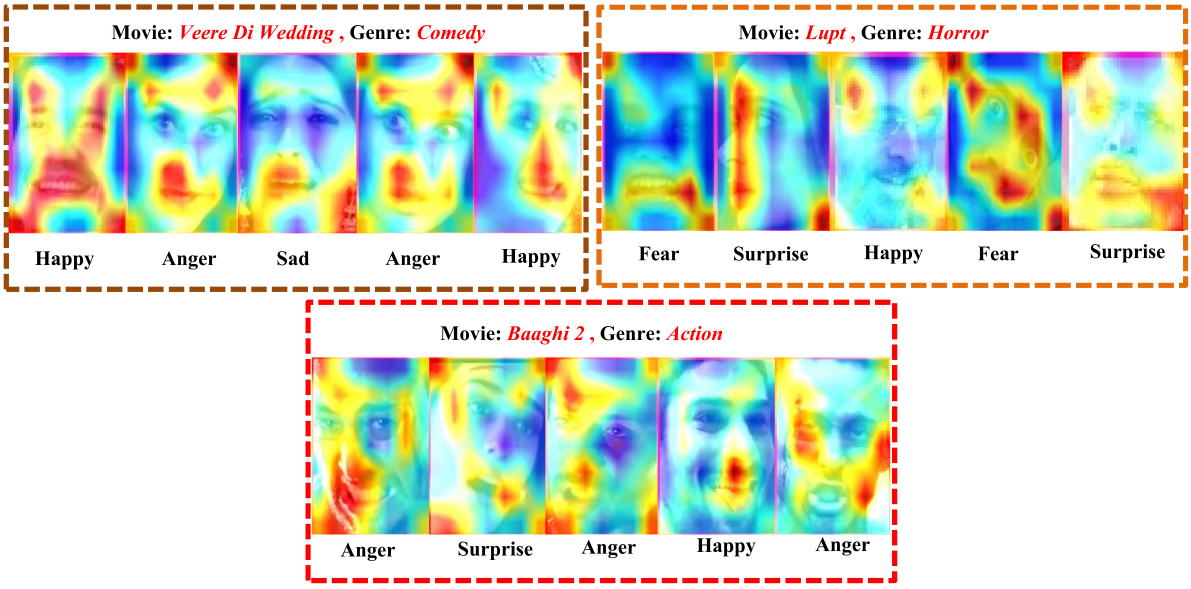

Fig. 11. Visualizing the discriminative image regions captured by the model for identifying different emotions in the movie trailers.

extends the work of CTT-MMC-C [52] by fusing the video features with audio features extracted by spectrograms. From Table 5, we can conclude that our proposed ILDNet architecture surpasses the low-level state-of-the-art methods by 30% and high-level approaches by 5%–16% on different datasets.

As discussed above, we report the performance of our architecture by visualizing the precision–recall curves for each output genre class in Fig. 10. We also compute AU(PRC) to compare the performance across the datasets. The best results of 0.81, 0.94, and 0.89 are obtained for LMTD-9, MMTF-14K, and ML-25M datasets, which reaffirms the adequate learning of the ILDNet architecture. The architecture can perform well across the movie trailers of different cinemas, despite being trained on the dataset of Indian cinema.

5.2.4. Computational efficiency

To evaluate the efficiency of our framework, we report the running time to classify a youtube video trailer on Windows 10, 64-bit machine with 128 GB RAM using NVIDIA Titan RTX GPUs, after training our network. Our testing phase involves the following steps:

- Selection of the facial frames from the videos, which takes 1.8 s.
- Abstraction of spatiotemporal features, which takes 12.3 s.
- Computing the emotion labels and referring the emotion-genre mapping for predicting the final genre labels of the videos. This process is finished in 2 s.

Thus, the proposed framework requires 16.1 s for a 3 min long video. We compare our results with other popular methods in movie genre classification. The results are summarized in Table 6.

As seen in Table 6, our framework runs efficiently than other competitors. Although CoNNeCT [32] shows less running time of 10.6 s, but its accuracy is lesser than the proposed works, which can be seen in Table 5. Hence, it can be summarized that the proposed framework is better than earlier works, which makes more acceptable.

5.2.5. Visualization

We also provide qualitative results in Fig. 11. We visualize the discriminative image regions captured by the ILDNet architecture for identifying different emotions in the movie trailers, by showing the gradient-based class activation maps [57] for comedy, horror, and action genre. We can easily see that ILDNet can detect the relevant regions in the images which evoke certain kinds of emotions. Thus, each of the genres elicits different types of emotions, which validates the fact that there exists a strong

Table 6
Comparison of Computation Time of movie genre classification methods.

Method	Time (s)
CoNNeCT [32]	10.6
CNN-Motion [33]	24.3
CTT-MMC-C [52]	18.6
CTT-MMC-TN [34]	25.2
Ours	**16.1**

correlation between the emotions and movie genre, justifying the motivation of our proposed work. This proves that induced emotions contribute significantly to identify and classify the genres of movie trailers.

6. Conclusion

In this paper, we present a novel deep affect-based movie genre classification framework, which is based on extracting the affective content from the movie trailers. We also develop an affect-based EmoGDB dataset, which contains 100 Bollywood movie trailers annotated with six different types of induced emotions for each of the six popular Bollywood genres. Firstly, we extract the video frames which have faces in them as facial expressions can distinguish emotions. Secondly, the spatial and temporal features are learned with ILDNet architecture. Finally, we develop an emotion-based multiple genre detection theory for effectively classifying the movie trailers into different genres based on emotions. We validate the performance of the proposed architecture by conducting several experiments on LMTD-9, MMTF-14K, and ML-25M datasets, which gives 0.81, 0.94, and 0.89 AU(PRC) values, respectively. Further, we conclude that the proposed affect-based framework successfully establishes the relationship between evoked emotions and movie genres, which helps to classify the movie trailers effectively. The prime advantage of our work is that without watching an entire film, the framework can classify the genre of the film by its trailer only. Moreover, it can be used for classifying the movie-trailers of different cinemas like Hollywood, Cinema of Japan, Denmark, South Korea, etc. efficiently.

The limitation of our work is that the facial expressions are not enough to depict the emotion categories of a trailer. Hence, in the future, we may develop a holistic approach for movie genre classification, which will consider the motion of whole body parts. We will also extend our proposed dataset to include the movement of the character and background of the scene to capture the motion in the videos for genre classification. We also plan to extend our work by incorporating multiple modalities like audio and text for multimodal genre-based classification.

CRediT authorship contribution statement

Ashima Yadav: Software, Validation, Investigation, Data curation, Writing - original draft, Visualization, Formal analysis, Resources. **Dinesh Kumar Vishwakarma:** Conceptualization, Methodology, Writing - review & editing, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] M.U. S., B.C. Kovoor, Towards genre-specific frameworks for video summarisation: A survey, J. Vis. Commun. Image Represent. 62 (2019) 340–358.
[2] G. Irie, T. Satou, A. Kojima, T. Yamasaki, Affective audio-visual words and latent topic driving model for realizing movie affective scene classification, IEEE Trans. Multimed. 12 (6) (2010) 523–535.
[3] G.M. Smith, Film Structure and the Emotion System, Cambridge University Press, 2003.
[4] H.L. Wang, L.-F. Cheong, Affective understanding in film, IEEE Trans. Circuits Syst. Video Technol. 16 (6) (2006) 689–704.
[5] Y. Hou, T. Xiao, S. Zhang, .X. Jiang, X. Li, X. Hu, J. Han, L. Guo, S. Miller, R. Neupert, T. Liu, Predicting movie trailer viewer's like/dislike via learned shot editing patterns, IEEE Trans. Affect. Comput. 7 (1) (2016) 1949–3045.
[6] A. Hanjalic, L.-Q. Xu, Affective video content representation and modeling, IEEE Trans. Multimed. 7 (1) (2005) 143–154.
[7] A. Lang, J. Newhagen, B. Reeves, Negative video as structure: Emotion, attention, capacity, and memory, J. Broadcast. Electron. Media 40 (4) (1996) 460–477.
[8] Z. Rasheed, Y. Sheikh, M. Shah, On the use of computable features for film classification, IEEE Trans. Circuits Syst. Video Technol. 15 (1) (2005) 52–64.
[9] L.-H. Chen, Y.-C. Lai, H.-Y.M. Liao, Movie scene segmentation using background information, Pattern Recognit. 41 (3) (2008) 1056–1065.
[10] A. Yadav, D.K. Vishwakarma, Sentiment analysis using deep learning architectures: a review, Artif. Intell. Rev. (2019) 1–51.
[11] A. Yadav, D.K. Vishwakarma, A comparative study on bio-inspired algorithms for sentiment analysis, Cluster Comput. (2020) 1–21.
[12] J.G. Ellis, W.S. Lin, C.-Y. Lin, S.-F. Chang, Predicting evoked emotions in video, in: IEEE International Symposium on Multimedia, 2014.
[13] S. Wang, Q. Ji, Video affective content analysis: a survey of state-of-the-art methods, IEEE Trans. Affect. Comput. 6 (4) (2015) 410–430.
[14] S. Mo, J. Niu, Y. Su, S.K. Das, A novel feature set for video emotion recognition, Neurocomputing 291 (2018) 11–20.
[15] J. Russell, A circumplex model of affect, J. Pers. Soc. Psychol. 39 (6) (1980) 1161.
[16] E. Cambria, A. Livingstone, A. Hussain, The Hourglass of Emotions, Springer, Heidelberg, 2012, pp. 144–157.
[17] P. Ekman, An argument for basic emotions, Cogn. Emot. 3 (4) (1992) 169–200.
[18] A. Ortony, T.J. Turner, What's basic about basic emotions, Psychol. Rev. 97 (3) (1990) 315.
[19] B.H. Detenber, R.F. Simons, G.G.B. Jr, Roll em!: The effects of picture motion on emotional responses, J. Broadcast. Electron. Media 42 (1) (1998) 113–127.
[20] M. Xu, C. Xu, X. He, J.S. Jin, S. Luo, Y. Rui, Hierarchical affective content analysis in arousal and valence dimensions, Signal Process. 93 (8) (2013) 2140–2150.
[21] L. Canini, S. Benini, R. Leonardi, Affective recommendation of movies based on selected connotative features, IEEE Trans. Circuits Syst. Video Technol. 23 (4) (2013) 636–647.
[22] M. Xu, J. Wang, X. He, J.S. Jin, S. Luo, H. Lu, A three-level framework for affective content analysis and its case studies, Multimedia Tools Appl. 70 (2) (2012) 1–23.
[23] J. Fleureau, P. Guillotel, Q. Huynh-Thu, Physiological-based affect event detector for entertainment video applications, IEEE Trans. Affect. Comput. 3 (3) (2012) 379–385.
[24] D. McDuff, R.E. Kaliouby, J.F. Cohn, R. Picard, Predicting ad liking and purchase intent: Large-scale analysis of facial responses to ads, IEEE Trans. Affect. Comput. 6 (3) (2015) 223–235.
[25] Y. Yi, H. Wang, Q. Li, Affective video content analysis with adaptive fusion recurrent network, IEEE Trans. Multimed. (2019).
[26] F. Noroozi, M. Marjanovic, A. Njegus, S. Escalera, G. Anbarjafari, Audio-visual emotion recognition in video clips, IEEE Trans. Affect. Comput. 10 (1) (2017) 60–75.
[27] S. Bansal, C. Gupta, A. Arora, User tweets based genre prediction and movie recommendation using LSI and SVD, in: Ninth International Conference on Contemporary Computing (IC3), India, 2016.
[28] P. Ghaemmaghami, M.K. Abadi, S.M. Kia, P. Avesani, N. Sebe, Movie genre classification by exploiting MEG brain signals, in: International Conference on Image Analysis and Processing, Cham, 2015.
[29] Y.-F. Huang, S.-H. Wang, Movie genre classification using svm with audio and video features, in: International Conference on Active Media Technology, Berlin, Heidelberg, 2012.
[30] G. Paıs, P. Lambert, D. Beauchêne, F. Deloule, B. Ionescu, Animated movie genre detection using symbolic fusion of text and image descriptors, in: 10th International Workshop on Content-Based Multimedia Indexing (CBMI), Annecy, 2012.
[31] K. Choroś, Video genre classification based on length analysis of temporally aggregated video shots, in: International Conference on Computational Collective Intelligence, Vietnam, 2018.

[32] S.G. Simoes, R.C. Barros, J. Wehrmann, D.D. Ruiz, Movie genre classification with convolutional neural networks, in: International Joint Conference on Neural Networks (IJCNN), Vancouver, 2016.
[33] J. Wehrmann, R.C. Barros, G.S. Simoes, T.S. Paula, D.D. Ruiz, (Deep) learning from frames, in: IEEE 5th Brazilian Conference on Intelligent Systems (BRACIS), Brazil, 2016.
[34] J. Wehrmann, R.C. Barros, Movie genre classification: A multi-label approach based on convolutions through time, Appl. Soft Comput. 61 (2017) 973–982.
[35] A.M. Ertugrul, P. Karagoz, Movie genre classification from plot summaries using bidirectional LSTM, in: 12th IEEE International Conference on Semantic Computing, California, 2018.
[36] J. Wehrmann, M.A. Lopes, R.C. Barros, Self-attention for synopsis-based multi-label movie genre classification, in: The Thirty-First International Florida Artificial Intelligence Research Society Conference (FLAIRS-31), Florida, 2018.
[37] J.A. Wi, S. Jang, Y. Kim, Poster-based multiple movie genre classification using inter-channel features, IEEE Access 8 (2020) 66615–66624.
[38] G. Tu, Y. Fu, B. Li, J. Gao, Y.-G. Jiang, X. Xue, A multi-task neural approach for emotion attribution, classification, and summarization, IEEE Trans. Multimed. 22 (1) (2019) 148–159.
[39] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, 2015.
[40] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A.C. Berg, SSD: Single shot multibox detector, 2016, arXiv:1512.02325v5.
[41] C. Szegedy, S. Ioffe, V. Vanhoucke, A. Alemi, Inception-v4, Inception-ResNet and the impact of residual connections on learning, in: Thirty-first AAAI conference on artificial intelligence, California, 2017.
[42] Y. Bengio, P. Simard, P. Frasconi, Learning long-term dependencies with gradient descent is difficult, IEEE Trans. Neural Netw. 5 (2) (1994) 157–166.
[43] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.
[44] K. Greff, R.K. Srivastava, J. Koutnık, B.R. Steunebrink, J. Schmidhuber, LSTM: A search space odyssey, IEEE Trans. Neural Netw. Learn. Syst. 28 (10) (2016) 2222–2232.
[45] F.A. Gers, J. Schmidhuber, F. Cummins, Learning to forget: Continual prediction with LSTM, in: Ninth International Conference on Artificial Neural Networks, 1999.
[46] F.A. Gers, J. Schmidhuber, Recurrent nets that time and count, in: Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks, Italy, 2000.
[47] K. Cho, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, B.v. Merrienboer, C. Gulcehre, Learning phrase representations using RNN encoder–decoder for statistical machine translation, 2014, arXiv preprint arXiv:1406.1078.
[48] M. Schuster, K.K. Paliwal, Bidirectional recurrent neural networks, IEEE Trans. Signal Process. 45 (11) (1997) 2673–2681.
[49] K. Yao, T. Cohn, K. Vylomova, K. Duh, C. Dyer, Depth-gated LSTM, 2015, arXiv preprint arXiv:1508.03790.
[50] J. Koutník, K. Greff, F. Gomez, J. Schmidhuber, A clockwork RNN, 2014, arXiv preprint arXiv:1402.3511.
[51] A. Graves, J. Schmidhuber, Framewise phoneme classification with bidirectional LSTM and other neural network architectures, Neural Netw. 18 (5–6) (2005) 602–610.
[52] J. Wehrmann, R.C. Barros, Convolutions through time for multi-label movie genre classification, in: Proceedings of the Symposium on Applied Computing, Morocco, 2017.
[53] Y. Deldjoo, M.G. Constantin, B. Ionescu, M. Schedl, P. Cremonesi, MMTF-14K: A multifaceted movie trailer feature dataset for recommendation and retrieval, in: Proceedings of the 9th ACM Multimedia Systems Conference. Netherlands, 2018.
[54] F.M. Harper, J.A. Konstan, The movielens datasets: History and context, ACM Trans. Interact. Intell. Syst. 5 (4) (2015) 1–19.
[55] H. Zhou, T. Hermans, A.V. Karandikar, J.M. Rehg, Movie genre classification via scene categorization, in: Proceedings of the 18th ACM International Conference on Multimedia, Italy, 2010.
[56] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2015, arXiv preprint arXiv:1409.1556.
[57] R.R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, D. Batra, Grad-CAM: Why did you say that? 2016, arXiv preprint arXiv:1611.07450.