International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-3, September 2019

Multimodal KDK Classifier for Automatic Classification of Movie Trailers

Prashant Giridhar Shambharkar, M N Doja, Dhruv Chandel, Kartik Bansal, Kunal Taneja

Abstract*: Movie trailer classification is a field of automation of analyzing the movie trailers and classify them into one of the various genres. In this paper, we proposed a classifier to identify the genre of a movie trailer by analyzing it's audio and visual features simultaneously. Our Approach decomposes a trailer video into frames and audio file and then analyze them based on certain specific features to categorize them into four genres. Our aim was to minimize the number of parameters involved in analyzing the trailer since other papers use many arguments which are impractical. The proposed classifier was trained on 4 audio, 2 broad visual features extracted from over 900 movie trailers distributed across 4 different genres, namely Drama, Horror, Romance, and Action. The Classifier Model has been trained using Neural Networks and Convolutional Neural Networks. Our Classifier Model can be used in Recommendation Systems and various websites like IMDB for automation of the genre classification process. As the common humanly approach is to generalize the results obtained from many inputs, the same way we use multiple models to obtain different outputs from multiple ANN models and then combine all the obtained results to get a final output. Also a Dataset containing 1000 movie trailers was introduced in this paper with trailers spanning to almost all Hollywood movies from 2010-2019 .After training and conducting Experiments on around 1000 movie trailers, the classifier model showed a maximum accuracy of 81 percent in determining top 1 genre and 91 percent in determining top 2 genres of a movie trailers in the test set.

Keywords*: movie trailer, movie genre classification, movie trailer genre classification, neural network, audio visual features, movie data-set.

I. INTRODUCTION

Video Classification is a well-researched topic; still, the results yet obtained are not satisfactory. Same is the issue with the topic we have researched upon, movie trailer genre classification..

Manuscript published on 30 September 2019
* Correspondence Author
Prashant Giridhar Shambharkar*. Assistant Professor in Department of Computer Science & Engineering, Delhi Technological University, Delhi
Dr. M.N. Doja, Professor in the Department of Computer Engineering, Faculty of Engineering and Technology, Jamia Millia Islamia, New Delhi-
Dhruv Chandel Information Technology from Delhi Technological University.
Kunal Taneja Technology in the field of Information Technology Delhi Technological University, Information
Kartik Bansal, (B.Tech) in Information Technology from Delhi pursuing Bachelors in Technological University

The difficulty in precisely approaching this field is that a single movie trailer is not always confined to a single genre. A single trailer may contain kiss scene stating it as a romantic trailer and few long talks stating it as a drama or a fight scene stating it as action trailer. So because of the absence of any such specific hard-and-fast rules of specifying a genre, it is not such an easy task to do

A.Movie Trailers

Movie Trailers are a short representation of a movie, released before the whole movie is released. Movies are a vast money-making profession if approached correctly, and the way to reach out audiences is through previews or trailers. Trailers are a short and crisp representation of the whole movie plot at once. Trailers depict and give an idea to the audiences about what they can expect from the movie at whole. Genre is such an integral choice for targeting huge audiences so that producers and directors of the movie can release a trailer according to the type of audiences they want to target by analyzing their genre using our model.

B. The Curse Of Incorrect Correctness

Interpreting Genre of a movie trailer isn't an easy task as there is no particular set of rules describing different genres. Deciding genre of a movie is relatively is an easy task instead then interpreting the actual genre of trailers because trailers are like an ultra-compressed form of a plot of the movie.
Also, multiple genres can be interpreted by different persons for the precisely same trailers because of the individuality and everyone's different mood, setting, and way of thinking. So what sources are trustable to find correct genre?
One of the most reliable sources for determining the genre of movies is IMDB, but here comes the curse of incorrect correctness. IMDB describes the genre of the movie as a whole and not of the trailer. Many a time the genre depicted by the IMDB rating is not the genre depicted by the trailer. To state some :

- Drishyam: A movie with rated Genre by IMDB as Crime, Drama, and Thriller was interpreted as Horror by our model due to overall vibes of Horror from the movie trailer.
- IRON MAN 3:- A movie that we all know as an Action movie had minimal features of action in its trailer and thus our model classified it into Drama Genre.
- Buried(2010):-This movie is rated as Drama by IMDB while its trailer has a very dark background and audio like that of Horror and action and thus it was classified as Horror by our model.

So for overcoming these error, we have manually watched all the trailers of the test, and training set; compared them to IMDB rated genre to correctly classify them to correct classes.

C. *The Curse of Multiple Genre Correctness*

The Old times are gone where there were only 3-4 genres over which all the movies were based on. With time presently, there exist more than 30 movie genres like Action, Drama, Horror, Romance, Mystery, Sci-fi, Adaptation, Thriller, Melodrama, Psychic Thriller, and many more.

These genres aren't much different, but the thing is that different compositions of core genres give rise to these secondary genres. For instance:

- Thriller: This Genre can be considered as a crudely a combination both of Action and Horror.
- Melodrama: Melodrama genre can be considered as a combination of both Drama and Romance.
- Romance: Romance although considered a core genre but is more or less a type of Drama only with added sensuality and generous talks.
- Psychic Thriller: It is new rising core genre.

These are only a few to state many more different genres have their alternative interpretations like the above stated only.

Some examples for this are listed as :

- Khamoshiyan: It is classified in genres Drama, Horror and Romance by IMDB.
- Passengers: It's classified as Drama, Romance and Sci-fi by IMDB.
- KGF Chapter 1: It's classified as Action and Drama by IMDB.
- Crawl: It's classified as Action, Adventure and Drama by IMDB.

The contributions of this work are **(1)** Introducing a movie trailer video data-set named 'ALOO' (All Languages Organized Omnipotent) dataset publicly available for a research fraternity. **(2)** Introducing 3 new Audio and 1 new visual state of the art features useful in interpreting the genre of a movie trailer. **(3)** Source Code is made publicly available.

Further, in the paper In Section II, we discuss the related work done in this field, in section III we discuss The Public Dataset We Introduced for research community further in section IV we discuss the methodology that involves Data Pre-processing and Architecture Used in the classifier Models. Further, in section V, VI, and VII, we describe Implementation Details, Experimental Results, and Comparison with Other Models and Classifiers. In section VIII and IX, we discuss Future Scope and Conclusion in this Field of Research.

II. RELATED WORK

Video Classification field is already well researched by many people. Here we discuss about the work happening in the field of video classification using deep neural networks.

One of the foremost work was done in [8] , who basically used CNN as well as LSTM for video classification. They used the VGG convolution base and used techniques like Dense Optical Flow and Loose Labelling of frames to capture both spatial and temporal features of video along with CNN. Also, LSTM used by them worked on 9 frames at a time to give better accuracy than CNN. One of the drawbacks in their model is a large number of parameters involved along with VGG(about 13 crores), thereby leading to substantial computation time as compared to our model.

This paper [10] works on the concept of multiple label system, which enable the model to classify the trailer into more than 1 genre which is realistic in comparison to the single labeled which imparts single genre to all scenes. They used max-pooling and convolution to extract features. They made a deep neural network of 152 layers and trained on the ImageNet(1.2 million images and 21,000 classes) and Places365(1.8 million images and 365 classes). Residual net is pre-trained on ILSVRC 2012 subset of ImageNet.They also used audio features, but instead of extraction they transformed the spectrogram of the audio signal into 130x530 and passed it to CTT for the classification. Moreover, we got the result of 66% from audio. They used 3 different algorithms for calculating the results and thus got 64%, 74%, 72% accuracy. They claim that their all CTT models outperformed the LSTM models in MMC.

In [4] they Worked on audio and visual features for movie trailer genre classification. After using Bruteforce approach for extracting 277 Audio and Visual Features, they used SVM for classification of 7 movie genres and applied SAHS algorithm on it for feature selection and dimensionality reduction of their classifier model. However, the thing to think about is that their Dataset contains 223 movie trailer for 7 movie genres i.e., a tiny dataset and can easily overfit upon so many features to choose from. Also, similar Dataset not made publicly available for comparison make things even more thoughtful.

In [5] was a paper on which they worked specifically on movie clips and not a movie trailer, which is relatively less complicated field. They used 5 Audio and 4 Visual Features to differentiate between 5 genres, but the dataset again used was really small like only 200 movie clips for 5 moviegenres but the classifier model they presented was absurdly trained using neural network with 20000 epocs and 0.05 learning rate, which is very likely to show overfittingas suggested by some experts in the field of machine learning. Their classifier showed an accuracy of 87.5 percent, which is undoubtedly due to the high chances of over-fitting.

III. ALOO DATASET

D. *Training Set*

In this paper, we launch a state of the art data-set containing about 900 latest movie trailers, most of them ranging from movies released in the years 2010-2019.

Choosing the trailers for the movies during this period is a crucial step for any researcher because as times change the perspective of people, film directors and societies also change and so there is a surely a noteworthy change in the trailer of the movies over the period. Our movie trailer dataset mostly has HD trailers only so that, their quality can easily be moderated for any usage into lower resolutions also. Moreover, we have made the dataset available publicly for the research community so that no researcher has to devote their precious time in collecting and making an adequate dataset. Moreover, due to the availability of the dataset in raw video form, the data can be pre-processed in any accessible format. The bifurcation of the dataset is stated in TABLE I.

Table 1: Dataset Bifurcation

GENRE OF TRAILER	TRAILER COUNT
ACTION	250
DRAMA	250
HORROR	200
ROMANCE	200

E. *Test Set*

Testing a Model isn't an easy task as it seems to be. What most of the research papers do is that they train and test their models purely based on a common dataset on which they are working on. If we think mostly dataset on which researchers have worked, haven't been so vast, so in a way even if every time they reshuffle their data virtually their model's performance has been over-fitted to mostly every trailer of their dataset. So to counter this issue, in this paper, we have brutally tested the model and tried to make it work in all conditions. We have tested the model in 3 ways, as stated below.

1) First Approach was reshuffling the data every time and testing classifier's performance on the test set, which is 10 percent of the total dataset.
2) Second approach was using 10 percent of the training dataset to find validation accuracy of the model using k-fold cross-validation algorithm[9].
3) The third approach was that, we made a new dataset which consisting of new trailers which were not more than a month older from the date when the model was finalized and all these trailers in new test set were absent from the training set, therefore a total blind unseen test set for the classifier, and hence classifier was finally tested upon the latest movie trailers which were the need of the hour.

IV. PROPOSED METHODOLOGY

A. Data Preprocessing

Data Processing means converting a raw form of data into an analyzable and understandable form to perform various operations on it and extract various features from the raw data about the subject. Data Preprocessing was, again, an integral step to be performed.

Firstly we took the raw form of video and extracted frames from it at 1 FPS using OpenCV[3] and converted frames to the dimensions of 100*100*3. Then we extracted Audio from video using movies [1] into .mp3 format and Audio Features were extracted using Librosa[7].

This extracted data was stored and passed to respective models so to analyze them and study and extract various features from them. Later on, considering space efficiency, all the extracted data is deleted after performing adequate operations on them.

B. Features And Models Used

- Visual Features and Convolution Neural Network Model

This section shows us how Convolution Neural Network was able to extract features from the video and gave results.

Every video trailer is made up of many frames and frames are rendered at 30 fps (mostly). We extracted these frames but at a lower fps as at high fps frames are repeated, and that would not contribute any new information to the model. Frames are extracted from the video at 1 fps and then passed to the CNN model for the evaluation of scenes(8 classes). After obtaining these 8 features, we aimed to merge these 8 features to get the final results of 4 different classes. For this purpose, we implemented another Neural Network which classifies the trailer into 4 classes on the bases of these 8 features extracted from the Convolution Neural Network.

- Convolution Neural Network is one of the advanced forms of neural network. It was introduced by Yann LeCun in the year 1994, as stated in \cite{lecun1995convolutional}. Every layer tends to extract features from the image and pass it to the next layer. Each next layer created by convolution has degraded size by 2 in each dimension. A Convolving layer generates the next layer by convolving on the input layer. In our case, we used this algorithm to classify the images into different subcategories. Namely these category include :
  - Fight/Action
  - Drama/Comedy
  - Cars
  - Horror
  - Explosion
  - Party
  - Robot
  - Adult Content

We obtain percentages of each subclass through the Convolution Neural Network(CNN) model. This output layer of 8 values is used as the input layer for the next ANN model which uses a neural network to use these features and obtain the final result as *action, drama, horror, and romance.*

Convolution Neural Network(CNN) model was separately trained on the dataset of 5600+ images, which was self-made from the extracted images from the trailers. further ANN model was trained on the ALOO Dataset.

Unlike in [7] where only blue frames(middle frames) which were only from 5^{th} to 120^{th} second, which can be limited on the video dataset as most of the videos having length greater than 150 \textit{sec} could give wrong results, We extracted all images in the video at 1 fps, an to prevent the model on working not on trailer we set the limits to 500 *sec* since no trailer would be above 200 *sec*.

C. Audio Model And Audio Features

The Audio is extracted from the .mp4 format of the trailer in .mp3 format. The Audio is stored and passed to different classifier models to find the number of Significant Peaks, Average Time between peaks, MFCCs and RMS energy of the audio track.

Audio is then analyzed on the basis of features described using Neural Networks, and after passing any test audio, the classifier returns output in the form of 4 Genres.

- MFCC Features

  MFCC [11](Mel Frequency Cepstral Coefficients) is one of the most important single-frame audio features that describe the overall spectral shape of our spectrogram. MFCC is calculated by taking Discrete Cosine Transform of the logarithm of Mel frequencies obtained by mapping power of spectrum obtained by taking Fourier Transform of time series of an audio sample. MFCC helps in characterizing the spectral shape of sound on a non-linear mel scale(eg.0th coefficient tells us about overall loudness of audio,1st coefficient gives us an overview of the spectrum and other features dig into further detail of spectrum). We took first 13 MFCC of audio sample to classify trailers into 4 genres respectively by taking the average of 13 MFCC over all time frames since first 13 features give almost entire detail about the audio sample as used by other papers as well.

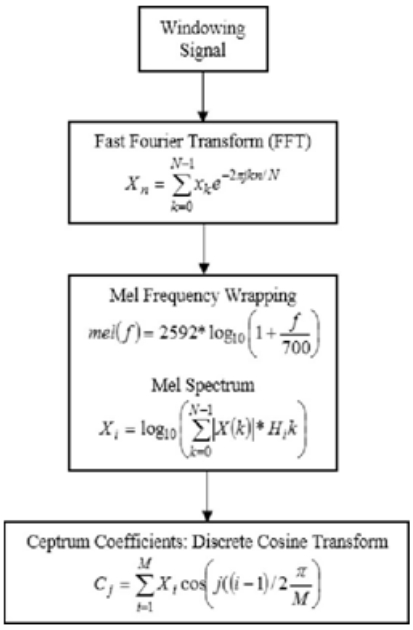

Figure 1: MFCC Evaluation [12]

- Number Of Significant Peaks

Number of Significant peaks is one of the new features we have introduced to this field of research. It was earlier used in onset detection of music signals [2]. Sound plays an important role in creating the mood of a person watching the video trailer. It modifies the subconscious of a human brain just to change the aura of one's mind. One of the factor human minds interpret about the sound while inspecting its genre is the sudden change in the sound effects.

These sudden changes are picked up by our model, and the total number of significant peaks per movie trailer are counted. Intuitively thinking more the movie trailer is active more will be the number of peaks and vice versa.

Therefore our model was trained on this feature to predict the genre of a movie trailer.

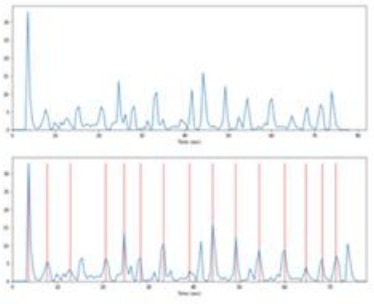

Figure 2:About Time(Romance)

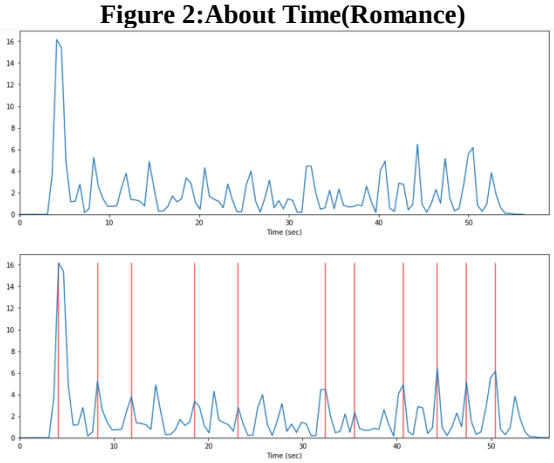

Figure 3: Blue Valentine(Drama)

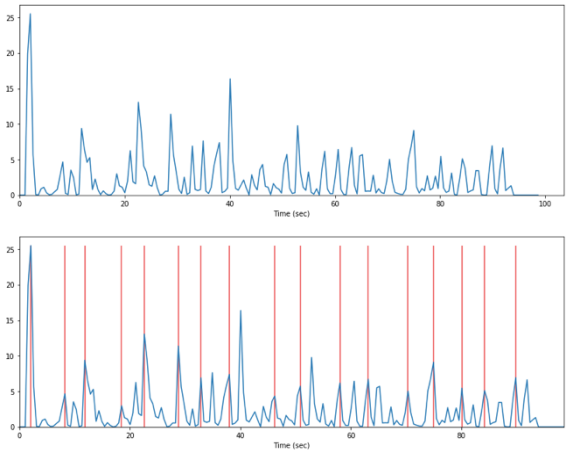

Figure 4: The Fate of Furious(Action)

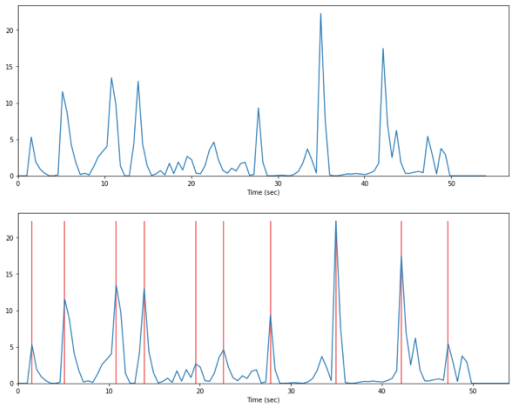

Figure 5: The Nun(Horror)

- Average Time Interval between Significant Peaks
  Average time interval between significant peaks is another new feature helpful in deciding the genre of a movie trailer. Either it is any genre, it has some significant peaks in it. Although the time span of movie trailers are almost the same still to eradicate any unwanted inaccuracy we take an average time interval in which the significant peaks occur throughout the movie trailer.
  As thought intuitively peaks occur at higher time intervals in horror trailers while in action trailer they occur more frequently. Drama and romance show similar kind of Average time intervals.
- RMS Energy
  The energy of an audio plot is approximately equal to the square of the amplitude of sound wave averaged over all time frames. RMS energy of audio is the root mean squared energy of an audio plot. Therefore average RMS energy of audio basically gives us an idea about average intensity of the sound of our movie trailer. This feature is useful since action trailers usually have quite high energy as compared to the drama and romance trailers in which intensity of sound remains quite low. Earlier energy plot has been used in order to detect significant peaks, but average RMS energy is used as a parameter value in our model itself.

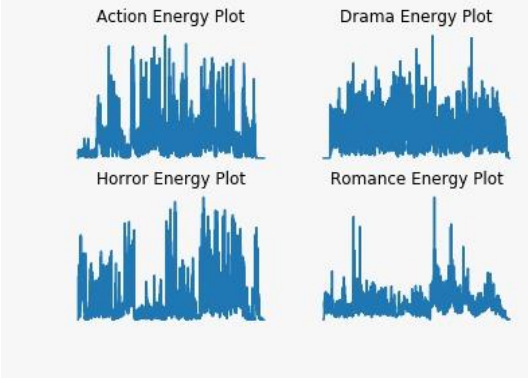

Figure 6: Energy Plots

D. Average RGB Factor And Net RGB Model

Net RGB model represents the whole movie trailer using a single color.

1. Firstly for all individual frames, an average RGB combination is found. For instance, if there are 200 frames, we will have a list of 200 RGB combinations for the whole trailer, with a single combination representing a single frame.
2. All the single combination are then summed up to find an average RGB combination representing the whole trailer.
3. This average RGB is passed through the Neural Network to analyze and give an output in the form of 4 genres.
   RGB factor is also a new factor what we introduced in this field of research. Earlier kind of similar approach which was used by many was using the lightning key, but it was a less accurate and more mathematical approach. We here took a simple approach to how humans interpret the visuals. Our subconscious mind acts to consider lightest and bright colors in romantic genre followed by drama, action, and horror movies. We used this approach to find a single mean color that could represent the whole trailer using one most dominant color throughout the movie trailer. This is a less mathematical and more accurate way to help classify movie trailers incorrect genres.

E. Result Combiner ANN

The Result Combiner is 4 layered neural network which takes in Outputs from all the classifier models in the form of percentages and after passing them through an ANN gives a final predicted genre of the movie trailer. This Final Result combiner acts like a feature selector for our classifier model, i.e., which feature to be considered for deciding one of the 4 genres accurately.

V. IMPLEMENTATION DETAILS

Implementation is to make sure what we planned works perfectly. Our implementation was based on the features extracted from 3 different models and ultimately achieving the genre from which most part of the genre belongs. Sometimes it's difficult to conclude the genre of the movie just on the basis of the trailer since the trailer constitutes the cut scenes to attract the mob. Our model aims not to give a positive result but what it sees from the trailer.

F. Convolution Neural Network Model

This sub-model works on the visual feature of the trailer and finds the 8 features, namely *fight, cars, drama/comedy, explosion, horror scenes, robots, romantic scenes, party* in the trailer and finally classify the video into one of the 4 classes.
When a video is passed to the model, firstly frames are extracted from the video at 1 fps. Then these frames having shape (SxSx3) are collected together to form a linear array using numpy and deducing the resolution of the image to 100 making shape (100x100x3), where S is the resolution of the image extracted, and 3 is for the RGB value layers thus formed is of the Red, Green, and Blue respectively. This whole array of shape (nx100x100x3) is passed as the input. The output is in the form of a linear array of shape (8,), which is interpreted as the percentage of frames belonging to each 8 sub-class. This array is further passed into another Artificial Neural Network(ANN) model, which ultimately classify into 4 class which we to make meaningful named to *action, drama, horror, romance*.

1. Architecture : We adopted very simple architecture to obtain good result instead of the huge and complex networks which take time to load. CNN model is has an architecture of 5 layers and further 4 more layers in the ANN model.

- First layer was the Convolution2D layer having activation as *relu* with input dimensions equal to 100x100x3 and output dimensions equal to 98x98x128, with convolving matrix of size 3x3.
- Second layer was of MaxPooling2D layer which picks the maximum from each set of 4 features. Output layer thus achieved is 49x49x128. Matrix used is of dimension 2x2.

- Third layer was again of the Convolution2D having activation *relu* with output dimension 47x47x64 and convolving matrix of dimension equal to 3x3.
- Fourth layer is the flatten layer which flatten all the values into a single array of size 141376.
- Fifth layer is the output layer and of the type dense with activation \textit{softmax} giving output as linear array of size 8.
- After that ANN model has 4 layers in which first layer is of dense type with input dimension as (8,) and output dimension equal to (128,) with activation *relu.*
- Second layer of ANN is again of dense type having output dimension equal to (64,) and activation *relu.*
- Third layer of the ANN model is dense layer with activation *relu* and output dimension as (32,).
- Fourth and the last layer of the ANN model is also dense layer with activation *softmax* and output dimension equal to (4,)

All these layers combine to form one of the 3 sub-models(CNN model).

Figure 7: CNN Model

2. Evaluation : We trained the CNN model on the dataset of 5600+ images classified into 8 classes, with 25 epochs. The optimizer used was Adam, and the learning rate was set to 10^{-4}. The evaluation was on the basis of accuracy. The loss function used was the Categorical Cross-entropy. We shuffled the dataset after every epoch. The batch size was set to 100, and validation split was 0.10. This experimental combination gave the accuracy equal to 73.2 percent and validation accuracy equal to 71.7 percent. The ANN model was trained on the dataset of 900 trailers whose features were first extracted from the CNN model in 8 classes. Then using these 8 features as the input ANN model was trained to have optimizer equal to Adam, learning rate as 10^{-4} loss function as categorical cross-entropy, and evaluation on the basis of accuracy. The batch size was set to 40, and data was shuffled after every epoch. From this, we got the validation accuracy of 73 percent.

3. Audio ModelAudio Model uses the ReLu activation function throughout the model, which outperforms all considerable activation functions like tanh, sigmoid, hard-sigmoid, relu.

The Audio model takes in input as all 16 audio features, namely the 13 MFCC, number of significant peaks, the average time difference in between the peaks and RMS energy.

Model is then applied to give out the percentages of each genre detected within a movie trailer using the Softmax activation function in the final prediction layer.

1. Architecture

   Audio ANN model basically consists of 4 Dense layers-3 hidden layers and 1 output layer.

- First Layer consists of 64 neurons taking in input of dimensions (16,) representing 16 audio features of a single movie trailer. ReLu activation is provided.
- Second Layer consists of 128 neurons and ReLu activation.
- Third Layer consists of 64 neurons and ReLu activation. Output from this layer is supplied to final output layer.
- Fourth Layer is final output layer that finally gives us outp't of dimension (4,) and it is Softmax activated.
- Total trainable parameters: 17,924 parameters.

2. Evaluation

We trained the ANN model on the dataset of about 672 movie trailers. For each trailer, audio features were extracted, and finally, a CSV file is stored containing 16 audio features of 672 movie trailers. The optimizer used was Adam(AMS grad variant), and the learning rate was set to10^{-4}. Model performance was evaluated on the basis of accuracy. Batch Size used was 64, and the number of epochs was set to 300. Validation set constitutes 10 percent of training data. The validation set was shuffled after each epoch. Loss Function used was Categorical Cross-Entropy. We achieved training accuracy close to 60 percent, whereas validation accuracy close to 55 percent.

Figure 8: Audio ANN Model

G. Net RGB Model

The RGB model is kind of straightforward model with only 3 dimensions in input as the average red, green, and blue color combination of the whole movie trailer. It serves similar architecture to that of the audio model but with a different combination of neurons to serve the best accuracy. In the RGB model, the mean value of RGB is found for each single movie trailer to get a (3,) vector, which serves as input to RGB ANN model.

1. Architecture : RGB ANN model comprises of total of 3 Dense Layers-2 Hidden Layers and 1 Output Layer.
   - First Layer consists of 128 neurons and ReLu activation and takes input of dimension (3,) for each movie trailer.
   - Second layer consists of 64 neurons and Relu activation and it's output is provided to output layer.
   - Third Layer is final output layer that finally gives us output of dimension (4,) and it is Softmax activated.
   - Total trainable parameters:9,028 parameters.

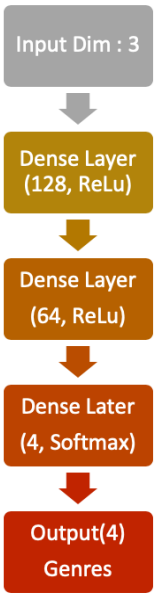

Figure 9: RGB ANN Model

2. Evaluation : We trained the RGB model on the dataset of about 520 movie trailers. For each trailer, mean RGB value is found, and finally, a CSV file is stored containing 3 RGB values of 520 movie trailers. The optimizer used was Adam(AMS grad variant was used), and the learning rate was set to 10^{-4}.Model performance was evaluated on the basis of accuracy. Batch Size used was 16, and the number of epochs was set to 300. Validation set constitutes 10 percent of training data. The validation set was shuffled after each epoch. Loss Function used was Categorical Cross-Entropy. We achieved both training and validation accuracy close to 50 percent.

H. Result Combiner ANN

Result Combiner ANN takes in input from all 3 models of dimension (12,) where each of the previous models was supplying (4,) output. Then Result Combiner ANN finally produces (4,) output representing percentages of all 4 genres. Its function is to combine the result of all 3 models.

1. Architecture : Result Combiner ANN consists of 4 Dense Layers- 3 Hidden Layers and 1 Output Layer.
   - First Layer consists of 32 neurons and ReLu activation and takes input of dimension (12,).
   - Second Layer consists of 64 neurons and ReLu activation.
   - Third Layer consists of 32 neurons and ReLu activation and supplies input to final output layer.
   - Fourth Layer is final output layer comprising of 4 neurons and Softmax activation thus giving final percentages of 4 genres predicted.
   - Total trainable parameters:4,740 parameters.

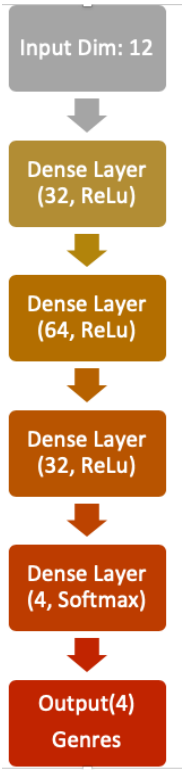

Figure 10: Result Combiner ANN Model

2. Evaluation : Result Combiner ANN model is trained on our training dataset "ALOO" comprising of 900 movie trailers, where 12 values are obtained for each trailer and stored in CSV file. The optimizer used was Adam(AMS grad variant was used), and the learning rate was set to 10^{-5}.Model performance was evaluated on the basis of accuracy. Batch Size used was 32, and the number of epochs was set to 500. Validation set constitutes 10 percent of training data. The validation set was shuffled after each epoch. Loss Function used was Categorical Cross-Entropy. The genre with the highest percentage obtained is our final predicted genre. We achieved both training and validation accuracy close to 80 percent.

VI. EXPERIMENTAL RESULTS

I. Test Dataset Description

We made a Test Dataset of 99 movie trailers on which finally our model performance can be seen.
Test Set comprises of :-

Genre	Number Of Trailer.
Action	24 Trailers
Drama	25 Trailers
Horror	25 Trailers
Romance	25 Trailers

Table 2: Test Set

J. Results

- TEST ACCURACY:- 81 percent
- CLASSIFICATION REPORT:-It is shown in Table III.

Table 3: Classification Report

GENRE	PRECISION	RECALL	F1-SCORE	SUPPORT
ACTION (0)	0.78	0.75	0.77	24
DRAMA (1)	0.79	0.92	0.85	25
HORROR (2)	0.78	0.84	0.81	25
ROMANCE (3)	0.9	0.72	0.8	25
MICRO AVG	0.81	0.81	0.81	99
MACRO AVG	0.81	0.81	0.81	99
WEIGHTED AVG	0.81	0.81	0.81	99

- CONFUSION MATRIX:- Refer to Fig. 11

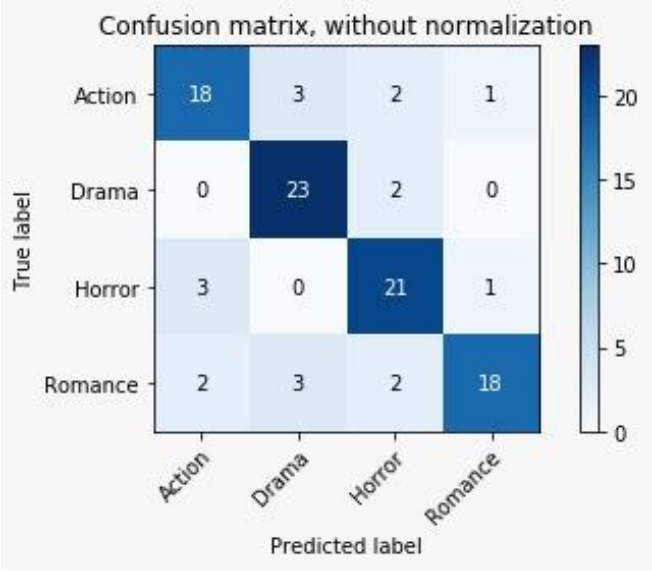

Figure 11: Confusion Matrix

- We tried different models to classify movie trailers into different genres before reaching to our final model. Results of all our models are shown in TABLE 4.

Table 4: Different Classifier Models

MODELS USED	FEATURES USED	ACCURACY
SVM (SUPPORT VECTOR MACHINES)	VISUAL	47%
CNN (CONVOLUTIONAL NEURAL NETWORKS)	VISUAL	62%
ANN (ARTIFICIAL NEURAL NETWORKS) AND CNN (CONVOLUTIONAL NEURAL NETWORKS)	VISUAL AND AUDIO	81%

VII. COMPARISON WITH OTHER MODELS/CLASSIFIERS

The metrics we here take in consideration while making the movie genre classifier are classifier are 'accuracy' and 'time-efficiency.'
While some classifiers work only over accuracy as a metric and do not consider the practicality of such applications, i.e., what will be the use of such movie trailer genre classifiers of they take up more than what a human will take to categorize the same by watching the trailer.
When a movie trailer of average duration (150 sec) is tested on machine with configurations of processor- intel i7 8th gen 8700Q , DDR4 8gb ram, it took about 45-60 sec to classify the trailer.
The 45-60 seconds included the time required to extract images, convert images, extract audio, analyze them, delete cache and give out results in .csv format, for a movie trailer of an average length of 2 minutes and 30 seconds which is a big thing in itself. Also, the parameters to consider, the introduced model uses overall very less number of trainable and non-trainable parameters if considered relative to the heftiness of the task of video processing.

VIII. FUTURE SCOPE

We thought about another feature which hasn't yet been introduced to this field of research. If we closely watch a movie trailer, we always notice that in Action and Horror Movies, the human voice content concerning the length of a video is less to what if Drama and Romantic movies are considered. This factor can be used to be incorporated in the audio model of the introduced classifier.
Further improvement in the time domain can be made by extracting frames in a lower resolution than the currently extracted HD resolution, so less disk read and write time is needed for the task.
Also Currently, we first extract the audio and convert it to .mp3 format for proper analysis, direct extraction of audio in .mp3 format from .mp4 format will help in improvement in time required for movie trailer analysis.

IX. CONCLUSION

Although the classifier we built is one of the best in class in terms of time and accuracy taken together giving an accuracy of nearly 81 percent.

Sometimes the classifier gets confused in between the Drama-Romance and Horror-Action genre of movie trailers due to similar cinematic attributes in the respective genres.
As a measure of accuracy, we tried to take the top 2 genres of movie trailers, which resulted in increased accuracy of 91 percent.
The parameters we considered while making the classifier model were of Time and accuracy hand-in-hand. The Time thus required to analyze a trailer of 150 at once was about 45-60 seconds.
Also, we considered, when we used the model to distinguish between Action and Non-Action movies, considering Action and Horror as Action and Drama and Horror as non-action movies the classifier gave us results with an overall accuracy of 80.795 percent.

ACKNOWLEDGEMENT

We want to express our sincere thanks towards Delhi Technological University, Jamia Millia Islamia, New Delhi for providing infrastructure and platform. Our heartfelt thanks to Siddhant Bhambri and Tenzin for their valuable and constructive suggestions during the planning and development of this research work. Their willingness to give their time so generously has been very much appreciated.

REFERENCES

1. Audio Extraction using moviepy Python library. https:// pypi.org/project/moviepy/.
2. Juan Pablo Bello et al. “A tutorial on onset detection in music signals”. In: IEEE Transactions on speech and audio processing 13.5 (2005), pp. 1035–1047.
3. Gary Bradski and Adrian Kaehler. Learning OpenCV: Computer vision with the OpenCV library. ” O’Reilly Media, Inc.”, 2008.
4. Yin-Fu Huang and Shih-Hao Wang. “Movie genre clas- sification using svm with audio and video features”. In: International Conference on Active Media Technology. Springer. 2012, pp. 1–10.
5. Sanjay K Jain and RS Jadon. “Movies genres classifier using neural network”. In: 2009 24th International Sympo- sium on Computer and Information Sciences. IEEE. 2009, pp. 575–580.
6. Yann LeCun, Yoshua Bengio, et al. “Convolutional net- works for images, speech, and time series”. In: The handbook of brain theory and neural networks 3361.10 (1995), p. 1995.
7. Brian McFee et al. “librosa: Audio and music signal analysis in python”. In: Proceedings of the 14th python in science conference. Vol. 8. 2015.
8. K Sivaraman and Gautam Somappa. “Moviescope: Movie trailer classification using deep neural networks”. In: Uni- versity of Virginia (2016).
9. Mervyn Stone. “Cross-validatory choice and assessment of statistical predictions”. In: Journal of the Royal Statistical Society: Series B (Methodological) 36.2 (1974), pp. 111– 133.
10. Joˆnatas Wehrmann and Rodrigo C Barros. “Movie genre classification: A multi-label approach based on convolutions through time”. In: Applied Soft Computing 61 (2017), pp. 973–982.
11. Wikipedia: Mel-frequency cepstrum. https://en.wikipedia. org/wiki/Mel-frequency cepstrum.
12. Figure 1 : https://www.semanticscholar.org/paper/Infant's-cry-sound-classification-using-Cepstrum-Rosita-Junaedi/76999fe56627320afc4ea1e5d626da48e4c36634/figure/6

AUTHORS PROFILE

Prashant Giridhar Shambharkar
B.E. From Amravati University, Done M.Tech From RGPV, Bhopal, Working as an Assistant Professor in Department of Computer Science & Engineering, Delhi Technological University, Delhi having 14+ years teaching experience of various Computer Engineering and IT related subjects, worked in various committee at responsible position, member of ISO 9001-2008 member at institution level, worked as exam coordinator, counsellor, mentor, time table in-charge, assistant centre superintendent in university exam, centre controller in UPSEE exam, Alumni Incharge, Organized Literary Activities, etc.
His area of interest includes Data mining, Real Time Systems. Guided many Graduate level projects and supervised M.Tech thesis. Member of ACM, Life member of CSI.

Dr. M.N. Doja is Professor in the Department of Computer Engineering, Faculty of Engineering and Technology, Jamia Millia Islamia, New Delhi-110025. He was the founder head of the Department of Computer Engineering for six years from its inception and has established the innovative and upcoming labs in the department. He started B.Tech. (Computer Engineering) and B.E. (Computer Engineering) as day and evening course respectively. He started Ph.D programme in the Department of Computer Engineering.
He has more than two decades of academic, research, training and administrative experiences in the field of Computer Engineering and Information Technology.
He received his B.Sc. (Engg), M.Tech. and Ph.D. degrees from B.I.T, I.I.T. Delhi and Jamia Millia Islamia, New Delhi respectively. His areas of research are Software Engineering, Networks, Security, Simulation, Operating System and Soft Computing. Prof. Doja is an active researcher. He has more than 100 publications in referred journals and conferences of international and national repute. He has been referee for a number of journals in the area of Computer Engineering and Information Technology. Several students have been awarded Ph.D under his able guidance and several students are working under his supervision. He is also author of a number of books in the area of Information Technology. He has successfully completed several projects including AICTE sponsored projects. He has chaired various sessions of various reputed conferences and delivered lectures on recent and burning topics in Computer Engineering and Information Technology as an invited speaker at a number of places across India and world. Prof. Doja is a member of Academic Council, Board of Research Studies and Board of Studies of a number of universities including Ambedkar University Lucknow, NSIT New Delhi, A.M.U. Aligarh, Guru Gobind Singh Indraprastha University Delhi, NIT Jalandhar, Hamdard University New Delhi etc. He has been a member of a number of committee for various universities in various capacities. He has been expert member/member of various committee constituted by UGC and AICTE. He has also been associated with the affiliation work of various universities like Guru Gobind Singh Indraprastha University Delhi, M.D. University Rohtak, U.P. Technical University Lucknow and Uttrakhand Technical University.

Dhruv Chandel Currently pursuing B.Tech in Information Technology from Delhi Technological University (formerly DCE). Proficient in languages like C, C++, Python. Skill Set consists of Web Developments using Django in backend and experience in fields of Data Mining, fine tuning Neural Networks and can handle various Machine Learning Techniques and algos. I am good at adapting and learning new skills when required and always interested in taking risks. Teamwork ,solving issue of compatibility with different OS's and confidence to present our work were the things are my speciality.
Have Good leadership and management skills and can adapt to new challenges and constraints.

Kunal Taneja 2nd year Student of Delhi Technological University, pursuing Bachelor of Technology in the field of Information Technology.
He has passed class X from Delhi public School ,Haridwar with 9.2 cgpa, completed class XII from Abhinav Public School with aggregate 94.2%.
He has always urge to constantly develop his skills and grow professionally. He is confident in his ability to come up with interesting ideas in the field of computer vision and artificial intelligence.
He have language proficiency in C, C++, Python, Dart. Also have dexterity in Data Structure and Algorithm, pursuing machine learning, artificial intelligence and have experience in openCV, keras, librosa. He have experience of learning app development using flutter.

Kartik Bansal Currently pursuing Bachelors in Technology(B.Tech) in Information Technology from Delhi Technological University (Previously known as DCE). Proficiency in languages like C, C++, Python, having good command over Data structures and Algorithms and currently learning and exploring the fields of Machine Learning, Deep Learning and Reinforcement Learning. Always ready to take up challenges and problems. Have curiosity in implementing different ideas and the approach for a particular task. Have excellent leadership qualities, persistent, focused, and confident about the work and always aspiring to learn new things. In this project he contributed in the model architecture design, Audio features extraction and optimization of neural networks. Wanted to grow both intellectually and professionally in coming years and explore the industry of Information Technology as far as possible.