Dissertation on

"Video Trailer Generation using Multimodal Data Analysis"

Submitted in partial fulfilment of the requirements for the award of degree of

Bachelor of Technology
in
Computer Science & Engineering

UE21CS390A – Capstone Project Phase - 1

Submitted by:

Nikhil Giridhar	***PES1UG21CS384***
Prajna R	***PES1UG21CS417***
Pranathi Praveen	***PES1UG21CS428***
Shreeja Rajesh	***PES1UG21CS564***

Under the guidance of

Dr. Surabhi Narayan
Professor

January - May 2024

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
FACULTY OF ENGINEERING
PES UNIVERSITY
(Established under Karnataka Act No. 16 of 2013)
100ft Ring Road, Bengaluru – 560 085, Karnataka, India

PES UNIVERSITY

(Established under Karnataka Act No. 16 of 2013)
100 Feet Ring Road, Bengaluru – 560 085, Karnataka, India

FACULTY OF ENGINEERING

CERTIFICATE

This is to certify that the dissertation entitled

‘Video Trailer Generation using Multimodal Data Analysis’

is a bonafide work carried out by

Nikhil Giridhar	***PES1UG21CS384***
Prajna R	***PES1UG21CS417***
Pranathi Praveen	***PES1UG21CS428***
Shreeja Rajesh	***PES1UG21CS564***

in partial fulfilment for the completion of sixth semester Capstone Project Phase - 1 (UE21CS390A) in the Program of Study - **Bachelor of Technology in Computer Science and Engineering** under rules and regulations of PES University, Bengaluru during the period Jan. 2024 – May. 2024. It is certified that all corrections / suggestions indicated for internal assessment have been incorporated in the report. The dissertation has been approved as it satisfies the 6th semester academic requirements in respect of project work.

Signature	Signature	Signature
Dr. Surabhi Narayan	Dr. Mamatha H R	Dr. B K Keshavan
Designation	Chairperson	Dean of Faculty

External Viva

Name of the Examiners **Signature with Date**

1. ____________________________ ____________________________

2. ____________________________ ____________________________

DECLARATION

We hereby declare that the Capstone Project Phase - 1 entitled **“Video Trailer Generation using Multimodal Data Analysis”** has been carried out by us under the guidance of Dr. Surabhi Narayan, Professor and submitted in partial fulfillment of the completion of sixth semester of Bachelor of Technology in Computer Science and Engineering of PES University, Bengaluru during the academic semester January – May 2024. The matter embodied in this report has not been submitted to any other University or Institution for the award of any degree.

PES1UG21CS384	**Nikhil Giridhar**
PES1UG21CS417	**Prajna R**
PES1UG21CS428	**Pranathi Praveen**
PES1UG21CS564	**Shreeja Rajesh**

Acknowledgment

First of all, we would like to express our heartfelt gratitude for our project mentor, Dr. Surabhi Narayan for the persistent support and expertise she offered to us, especially in conducting the project where her insights really proved to be so crucial by enabling us to achieve some project milestones.

We wish to sincerely thank our Project Coordinator, Dr. Priyanka H, for her constant guidance and support. Her help has really been instrumental in the smooth operation of the first phase of our capstone project.

We are honestly grateful to our honourable Chairperson, Dr. Mamatha H R who has been our guiding force. Her insightful words have been a true source of inspiration for all of us.

We are really thankful to the Dean of Faculty, for his unwavering help and constant support.

We also extend our deepest thanks to the honourable Vice Chancellor for continuously giving us counsel and moral support.

We owe the honourable Pro Chancellor a debt of gratitude for his unflinching support and priceless advice.

We appreciate our honourable Chancellor a lot because he presented us with this golden chance to get involved in a project that will make an impact in the field of technology.

Finally, we would like to express our gratitude to our families and friends for their constant support and encouragement.

Abstract

Manually crafting emotion-aware and attractive trailers is often cumbersome and challenging. There is a need to automatically generate intelligent and attractive trailers. Our project aims to address this problem by using multi-modal analysis integrating audio, video and other metadata. The purpose of our project is to explore and create novel approaches for audio-video processing to capture the "trailer-worthy" key moments from short movies and tailor them into visually attractive and emotionally aware trailers.

This project aims to develop a unique audio-guided video feature extraction technique which helps automatically generate an appropriate trailer when a short movie is given as input. This strategy ensures equal prominence is given to both auditory and visual features when getting the timestamps. The project also introduces "feature-fusion" which takes in the timestamps obtained from audio and video processing modules and gives an output of plausible "trailer-worthy" scenes. These scenes are ordered to craft an attractive and intelligent trailer capturing the essence of the short movie. We also aim to show that there is no need for synthetic content to generate good quality trailers.

TABLE OF CONTENT

LIST OF FIGURES

CHAPTER-1

INTRODUCTION

Starring as the primary portal for audiences to explore the variegated world of modern digital content consumption film trailers are the primary access points to upcoming cinema experiences, distilling films' core idea into easily understandable miniatures, appealing vistas, and emotionally fraught sound patterns that invite viewers into worlds they have grown accustomed to. Crafting good trailers is a complex and laborious process, involving a fine balance of creativity, intuition, and technical know-how. Therefore, as the demand for interesting advertising materials keeps rising in a media world that is more competitive by the day, there is a pronounced need for inventiveness when it comes to the production of trailers.

Influencing the audience's perception, raising anticipatory anxiety, prompting engagement, shaping the audience's thoughts and driving suspense – all these things and more can only be achieved through the effective generation of trailers. A well-done video clip can stir up curiosity which in turn makes people feel something leading to their viewing a movie of their choice or interest through a particular channel. Thus, it is an indispensable tool in the hands of movie-makers, producers as well as vendors. In addition, trailers profoundly influence the narrative discourse around movies and in turn critical reception, the expectations of the audience and cultural dialogues. This in turn means that generating great trailers is not only important in influencing how financially successful a certain film may become but also its standing as a work of art while majoring towards which it will be part of the culture.

The point of our project is to basically change how trailers are made by creatively combining analysis techniques of information from many modes; usage of sophisticated artificial intelligence tools, and keeping user interfaces simple and user-friendly. It does this by using sound, speech and text mining which makes it possible to make movies before they're even shot within a short space of time. Our project focuses on a comprehensive exploration of trailer generation methodologies, encompassing a variety of functionalities such as feature extraction, content analysis, stylistic modeling, user customization etc.

Our video preview generator system is a sophisticated layer with a variety of features that include audio processing, visual analysis, text mining, and trailer synthesis. Use cases for the same can extend to a variety of domains ranging from film marketing to content promotions to academic research. We base our approach to trailer creation on the interdisciplinary merging of computer vision, audio processing, natural language processing, and machine learning methods. Our goal is to continuously modify and enhance our trailer generation framework using algorithmic development, data-informed experiments, and user feedback inclusion in order to improve its performance and user satisfaction.

It is the purpose of our project to expand the limits of creating trailers by employing advanced technologies and pioneer methods to provide an innovative tool for producing exciting movie trailers to artists, film directors and various authors of advertising content. In the course of our activities, we set forth towards a new creativity-oriented way of making trailers – one that would be much more accessible as far as users are concerned and would have more influence on promoting digital media than it has ever been before.

CHAPTER-2

PROBLEM STATEMENT

Movie trailers have been around for a century, yet they always pique viewers' curiosity and excitement. However, creating engaging short movie trailers with conventional techniques such as selecting key moments is difficult and time-consuming. Due to their reliance on manual labour, traditional methods are not very effective for creating engaging trailers.

The proposed solution involves using an automated trailer generation approach. The challenge is creating an automated approach for choosing keyframes that conveys story details without producing spoilers.The success of this solution will lead to increase in efficiency and time saving.

CHAPTER - 3

LITERATURE SURVEY

In the following sections, we present a detailed Literature Survey for the domains of Video Trailer Generation, Audio-Video Processing and Video Summarization. From our review of the research papers, we have come across substantial findings that help shape, inform and reform our study.

Papers such as [1] describe new techniques to automatically detect key moments using video analysis, focusing on the narrative structure of the movie. Related to it are papers [2]-[4] which show that highlight detection is important in obtaining the key frames which not only are visually appealing but also consider context and relevance to the movie. Paper [5] shows that multimodal analysis aids in obtaining scenes which are most trailer-worthy. Work on embedded learning for identifying common key components for trailers in [6] and automatic summarization using textual information and multimodal data in [8] provide more insights. Papers [9] and [10] deal with deciphering video, audio and text files and deriving meaningful insights to classify and synthesize frames using language-vision models. Experiments on the impact of editing biases on viewer impressions in [11] and efforts towards developing frameworks for genre classification based on textual features extracted from movie subtitles in [12] make up a complete platform for advancing video trailer generation techniques.

P. Papalampidi et al. [1]

Movie narratives involve intricate relationships between the character and the events; hence they require us to understand their long-term cause-and-effect. The sheer size (e.g., movie length) and multimodal nature (video, audio, text) of movies make it difficult for standard AI/ML models to process. Moreover, creating large datasets with the video and corresponding textual information (metadata) for movie narratives, is a very challenging task, thus limiting the training data available.
Existing research normally focuses on either analyzing full text narratives, ignoring the visual aspects.

This paper proposes the following Hypothesis - Summarizing a movie is possible using the information about its narrative structure. They introduce a new dataset called **TRIPOD**. Their proposed workflow is as follows:

- Identifying the crucial moments or turning points based on the plot using information from the video which is broken down into segments or shots
- Summarizing screenplays from their underlying narrative structure.
- Summarizing movies using both audio and visual information and building a graph structure to represent the relationships between key moments.
- Creating the trailers by relying on human input for certain prompts.

Based on the results of this paper, it can be concluded that:

- Screenplays are used during training the model, but not needed while using the model.
- The model can identify the turning points in the story and trailer-worthy moments.

Wang, L. et al. [2]

In order to eliminate the exhaustive process of annotating the highlight moments in movies, the authors propose a novel method “Co-contrastive Attention network (CCANet)” to automatically detect trailer-worthy moments in movies by learning these key moments directly from the movie The authors have constructed a novel dataset “Trailer Moment Detection Dataset (TMDD)” to test their model. The contrastive attention network was also tested using YouTube Highlights, TV Sum and other on the VHD benchmark datasets

A two-step approach is has been proposed as a part of this research :

- Co-Attention Network : “Co-Attention” between movies and trailers help determine the training pairs where the moments are scored higher if they are more visually correlated to the scenes from the movie.
- Contrastive Attention Network : This module enhances the feature representation of the visual content, where the contrast between the moments in trailer and the movie are highlighted by maximising it.

The authors have used MAP, Rank@N and Rank@Global to test their findings; they have also compared CANet to other benchmark methods.

Co-Attention scores are maximum when trailer shots are from the movie itself
The augmenting makes the highlight features more distinguishable from the non-highlight features
The proposed Contrastive Attention module explicitly models the relations between the key and non-key moments so that “trailer” features stand out from the “non-trailer” moments helping build a more robust model.

The following strengths and weaknesses were observed from the findings of the paper
There is an area of opportunity to utilize multi-modal features to boost the trailer moment detection. This method shows performance gain on benchmark datasets and hence CCANet can be used for detecting key highlights in videos. It is also is targeted toward large movies, so scaling would be eased

Gan, Bei et al. [3]

The authors propose a method to learn highlights in movies through noisy labels. This model is trained on trailers which helps segment the movie into scenes. The shots in these scenes are noisy labels as not all trailer scenes incorporate highlight moments and most scenes are tailored to appeal to the audience. A novel approach known as CLC or collaborative noisy label cleaner is proposed to handle noisy labels in highlight detection.
The research uses a novel dataset “MovieLights” on which the approach has been tested. The research also uses “YouTube highlights” for validation purposes.

ACP - Augmented Cross Propagation exploits the related audio and visual signals and fuses them to learn single multi-modal representation.
MMC - Multi modal cleaning helps obtain cleaner labels by observing multi-modal losses

A three-step workflow has been proposed as a part of this research

- Feature extraction
- ACP : This module helps to capture multi-modal feature interactions. It also comprises of a “consistency loss” to show how much these modalities agree with each other in a feature space
- MMC : to tackle noisy labels

The research work shows that CLC outperforms the baseline Video Highlight Detection models by almost 20 - 23% . This model is also robust to varying label noise levels. CLC accounts for temporal and multi modalities helping in better understanding which scenes must be included in the highlights. ACP and MMC can be used in video highlight detection and further enhanced with other techniques to generate trailers. The authors, however have not implemented scene understanding

Wei, Fanyue et al. [4]

The authors propose a novel method "Pixel-Level Distinction Video Highlight Detection" (PLD-VHD) that models movies to obtain pixel level demarcations. This accounts for temporal and spatial relations between the content and also explores fine-level context to suggest what segments are appealing to the viewer from an unedited video. The authors have exploited the property of movie highlights being "context-dependent"
The following datasets have been used in this research : "YouTube Highlights", "TvSum", "CoSum" – existing benchmark datasets for testing

A two-step workflow has been followed

- Modelling Temporal Dependency : The video is segmented, each segment is associated with a label y, which determines if the segment is a highlight or not. Instead of adding individual frames as input to the estimation function, they add clips (L frames to I) and find the loss with ground truth
- Visual Saliency: In order to find out pixel level distinctions, visual attraction of the frame is considered. The saliency mask is used as pseudo-labels to obtain pixel level distinctions.

The following results have been observed

- PLD-VHD improves the benchmark methods on "TvSum" and "CoSum" by 3.1% and 9.9%
- TASED-Net with temporal and spatial works best on YouTube highlights dataset
- The PLD-VHD with TASEDNet accounts for spatial and temporal context. This may help in finding proper clips or shots to include in parts of the trailer

In addition to achieving best performance, the proposed approach has also shown expandability. However, their method mainly fails in “first-person” videos with a lot of cluttered background

Bretti, Carlo, et al. [5]

The authors proposed a multimodal approach based on utilizing pre-existing trailers to analyze which segments exhibit high trailerness, utilizing visual and subtitle data over various durations. This methodology aims to forecast the trailerness of specific sections within a TV episode or film, aiding editors in identifying optimal moments for trailer inclusion from extended video content.

This paper utilised the GTST dataset. It consists of sixty three episodes from the ”Goede Tijden, Slechte Tijden” which is a long-running Dutch soap opera

A novel approach , Trailerness Transformer based on three main stages which are, encoding, multi - modal and multi - scale transformers, and prediction aggregation.

- When input is given as a video representing a movie or television series, creating encodings at both the clip and shot levels for the visual and textual modalities of the video.
- The data is further encoded and processed by individual transformers, with the transformer output then undergoing a sigmoid function for normalization.
- Subsequently, to predict trailerness, the four prediction sequences are combined.
- The late fusion of predictions involves combining the frame-level predictions from the four multi - modal multi - scale streams using various possible combinations, such as averaging the likelihood of predictions is performed

Looking at the results discussed.

Quantitatively,

The proposed model outperforms all three baselines (random, MLP based ,Vasnet) by incorporating sequential order and temporal positioning in terms of clips being assigned for trailerness

Qualitatively,

Best Performing model - To yield higher trailerness, scenes with brighter visuals and emphatic dialogue delivery are required.

A few points can be noted regarding the methodology of the paper as given below,

Strengths

- The trailerness for each clip or shot aids editors in creating trailers. This inturn helps the editors in boosting their creativity due to selection and recommendation moments they would otherwise not have picked initially.
- This approach emphasizes the benefits from contextual information.
- This proposed method is targeted towards narrative-based videos (such as soap operas)

Weaknesses

- Although it does estimate the trailerness for the clips and shots it is noteworthy to mention that segmentation of these shots in a trailer is crucial
- The clips and shots selected are based on parameters such as high emotion visuals and high pitched dialogue delivery. Incase of incorporating this to our project , subjectiveness plays a crucial role , soap operas are region specific and hence can't be scaled to movie , where trailerness is found for all genres.

Sheng, Jiachuan, et al. [6]

An embedded learning algorithm has been suggested for the automatic generation of movie trailers, eliminating the need for human involvement. The development of a novel embedded classification algorithm for the purpose of identifying common key components within movie trailers.

The paper utilised ImageNet dataset for training. For testing dataset with 10 Movies namely The Dark Knight Rises , Inceptions, Transformers 3, Iron Man3 , Thor 2 , Prometheus and so on .

The workflow includes,

- Utilizing the VGG-F model for feature extraction from candidate frames.
- Applying SURF matching results to compare frames from a movie with frames from its trailer,
- Implementing semi-supervised learning based on "Friends being close and enemies being apart" principle, utilizing S4VM for the final classification of frames.

- Generating the trailer by stitching clips around the key frames identified during classification.

The results are as mentioned below ,

- Movies containing (a) a lot of attractive and exciting contents, such as actions and explosions (b) scenes with composition on protagonists' and dialogue delivery; have achieved high accuracy. An example for these are movies such as "Edge of Tomorrow" and "Resident Evil: Retribution". It can be seen here that "Prometheus" gets the least accuracy since the trailer does not have such characteristics.
- The proposed model outperforms the other three models based on similarity between the scenes picked by the model and trailer scenes

A few points can be noted regarding the methodology of the paper as given below,

Strengths

- Incorporates the creation of suspense through avoidance of story ending, thereby avoiding spoilers in the trailer generated.
- The features considered for the system aids in generating trailers for multigenre movies.

Weakness

- To enhance the presentation of movie content, the importance of the casts and sounds could be included, which may not satisfy conditions for creating a trailer.
- Trailers generated with just high-impact factors are not helpful.

Sadoughi, Najmeh, et al. [7]

A novel method is proposed for combining and aligning many modalities to interpret long-form videos (more than 60 minutes) in an effective and efficient manner. MEGA 1. MEGA addresses the issue of cinematic long-video segmentation by leveraging multiple media modalities .

The method is trained on Movienet-318 , IMDB , Places and testing on Movienet-318 , TRIPOD

The workflow includes,

The proposed method pipeline consists of the following steps .

- Initially, the video is preprocessed by dividing it into shots.
- Then, from each shot, features are retrieved and then pooling and normalisation are performed. The second step is the alignment and fusing of cross-modalities.
- Bottleneck fusion tokens and alignment positional encoding are used to accomplish this. Finally, scene and act segmentation is a part of the pipeline.
- Act segmentation is achieved by using knowledge transfer loss in this method, whereas scene segmentation uses CE loss.

The results are as mentioned below,

- In case of scene segmentation, MEGA Outperforms when trained on the three datasets M+P+I as compared to MEGA trained only on MovieNet 318, Mega outperforms all previous SoTA models in this aspect .
- In case of act segmentation, on the TRIPOD dataset, MEGA sets a new SoTA performance.
- Based on visual modality alone MEGA performs better than the prior SoTA, it also outperforms GRAPHTP, illustrates the alignment and fusion components and the suggested methodology. With +5.51% TA, +9.15% PA, and -%0.81 D, MEGA nearly doubles the performance of earlier studies.

A few points can be noted regarding the methodology of the paper as given below,

Strengths

- The suggested approach is versatile for use in real-world scenarios since it is scalable and generalizes to a variety of numbers of modalities at different scales.
- The method focuses on knowledge transfer between modalities using Knowledge Distillation

Weakness

- The appearance, location, activity, acoustic and textual features are the only areas explored in this work.
- Providing of actor name and identification can aid in scene/ act segmentation for long movie segmentation however it needs to be implemented to MEGA

P. Mishra et al. [8]

Author opposed the problem of manually developing trailers for online academic courses that consumes a lot of time and needs significant human labor and proficiency. Our proposed AI-based platform intends to computerize the creation of trailer content with regard to the text and visual components by employing machine learning technology plus natural language processing mechanisms.

Datasets for this included chapters from a textbook for ML and speech-to-text transcriptions of video lectures from an academic course on NLP..

The proposed workflow is as follows,

- The system is an approach for producing video trailers by using templates that contain components like splash images, titles of trailers, information on authors, synopses, metadata, endorsements, calls-to-action.
- To create a trailer, the template constraints are used which specify how it should look – for example which fonts should be used, what should be the pacing, where should cut scenes be, audio editing etc.
- The system comprises a video combiner module stitching together all elements likes frame data, voice-over text, and text-to-speech to make the final trailer video.

The results obtained were as follows,

- A user evaluation with sixty-three human evaluators is part of the research and it offers positive feedback.
- Human evaluation results were positive suggesting that the developed approach can be effective in creating online educational courses trailers.
- The authors also plan to enhance the present system through incorporating user evaluation feedbacks and introducing more fascinating topics.

These were the key takeaways,

- Improving trailer content by refining the pacing, readability, and general effectiveness of the trailers according to user feedback suggestions it is possible to enhance the auto-generated content in the system.
- Interactive Dashboard: More customization options and flexibility in trailer creation could be incorporated into the proposed interactive dashboard for content creators wishing to make edits to the auto-generated content.
- Advanced Themes: The system's capabilities would be advanced by more advanced themes and features like automatically detecting learning outcomes given resources, thereby providing a more comprehensive overview of course content in the trailers.

K. Porwal et al. [9]

The paper addresses the challenges caused due to reduced time, tight schedules, and a growing number of online events and activities, using audio and video content online is hard. The research aims to provide a solution through using Natural Language Processing (NLP) methods for creating transcripts and summarizing video clips. With this approach, audio-visual content can be converted into textual format and summaries that are qualitative are produced while retaining the original essence of the content.

Datasets included generic video files (not explicitly mentioned about specifics).

The proposed workflow is as follows,

- The media file gets divided into audio chunks consisting of frames further cut into tokens; which on their own are fed into Hugging Face Model that extracts them into text.
- The main ones are Statistical Information Retrieval (SIR) based analyses and Natural Language Processing (NLP) based information extraction approaches.
- The document contains tables showing metrics such as duration of videos, memory use in gigabytes and processing time as an evaluation parameter for measuring the efficiency of the proposed algorithm for video summarization.
- An unsupervised graph-based text summarization technique, called the Text-Rank algorithm is utilized for generating the video summaries.

These were the key takeaways,

- The lack of specific datasets mentioned in the document are for video transcription and summarizer. The absence of comparative analysis in the paper does not compare the proposed video summarization algorithm with existing methods or tools within this area.
- If the author had included a comparative analysis with other video summarization techniques, it would have shown the advantages and disadvantages of their proposed method.
- The document does not clearly define potential future directions for the research or perceive possible improvements or applications for the project.

Dong et al. [10]

In this paper we consider the issue of text-to-speech conversion which can be quite problematic especially because it is not easy to obtain well-structured texts necessary for training audio recordings. The main objective of our work is to handle the lack of actual texts related to sound signals due to their unavailability through an exploration into raw instructional videos combined with visual linguistic pre-possessed systems

The paper utilizes the following datasets
1) VGGSound: Consisted of 171,899 YouTube videos which had around 10 sec duration, covering 310 classes of sounds.
2) MUSIC: Consisted of 1,055 full-length YouTube videos of people playing various musical instruments, with around 21 instrument types.

The authors propose the following workflow

- Teaching a condition diffusion model for producing video audio tracks from video frame images, by using pretrained contrastive language-image pretraining (CLIP) models. Researching zero-shot modality translation through conditioning the diffusion model with a CLIP encoded text query at test time.

- Utilization of a pretrained diffusion prior model has been applied to solve the modality gap problem between textual and visual zed queries which lead to an equal parity regarding text-to-audio transformation as well as image-to-audio synthesis.

A detailed analysis of the research showed the following results

- The proposed CLIPSonic model effectively learns text-to-audio synthesis without text-audio pairs, leveraging unlabeled videos and pretrained language-vision models.
- The model demonstrates competitive performance in both text-to-audio and image-to-audio synthesis, offering a promising research direction for leveraging images as rich conditioning signals for audio synthesis.
- The study identifies a noticeable performance drop in text-to-audio synthesis when using text queries in a zero-shot setting, indicating a modality gap between the CLIP's image and text embedding spaces.
- The CLIPSonic-ZS model shows a performance drop in fidelity and relevance when using text queries, suggesting a challenge in effectively synthesizing audio from text queries.

H. Kakimoto et. al [11]

Movie trailers are tailored to a specific target audience depending on the movie's genre. Very few scenes are taken from the movie and hence the duration of the trailer is very less. As a result, it is very hard to edit a trailer. If the trailer isn't captivating, viewers will lose interest in the movie.

When a movie is summarized and a trailer is edited, there could be some biases like background music, characters, dialogues, scenes, sentiments and so on. In this paper, seven of these biases are analyzed to check if they may be used in curating a trailer which will cater to the preferences of the majority of the viewers.

The movie plots and summaries are taken from Wikipedia and IMDb Websites. Users’ impressions are also evaluated for the movie trailers, through questionnaires, for analysis against the editing biases.

The proposed workflow is as follows:

- Defining two categories of video editing bias - Audio-Visual and Contents. 7 of these biases are extracted - scenes' length, reordering, background music, characters, lines, topics and sentiment.
- Their effects in the trailers are analysed through a preliminary experiment using existing movie trailers.

Based on the results, scene reordering, length of scenes, emphasis of characters and number of topics have the greatest influence on how the impressions of the viewers differ from the movie to the trailer.

M. Hesham et. al [12]

Video summarization is considered a promising approach for efficacious realization of video content through Identifying and picking out descriptive frames of the video.

In this paper, an adaptive framework called the Smart-Trailer is proposed to use only the subtitles of any English movie to automate the trailer creation process. The dataset used here is the Kaggle movie dataset which contains around 5000 movies belonging to different genres.
The following workflow is proposed in the paper:

- The framework analyzes the movie subtitles and classifies the movies into their corresponding genres.
- The system uses many deep learning methodologies to capture the opinions and the behaviors of users to recommend relevant scenes based on their preferences.

The results are as follows:

- Initial experimentation generated a corpus for genre based classification, which is tested on the dataset of real movies. The accuracy rate was 0.89.
- The system also gave automated trailers which have an average accuracy of 47% for selecting scenes which are there in the original trailer.

Video trailer generation has been in the research field for over 15 years, and there have been numerous attempts to solve the challenge of generating efficient, concise and contextual trailers.

This field has been marked with continuous advancements and refinements over the years. The aforementioned papers represent a few of them. From this we can conclude the following :

Incase of the method proposed by Bretti, Carlo, et al. [5] , the novel method proposed i.e the multimodal approach at shot level outperforms the unimodal baselines which are text and visual.

In movie highlight detection, Collaborative Noisy Label Cleaner [2] outperformed baselines by 20-23%, proving to be robust across label noise levels by utilizing temporal and multi-modal cues. “Smart Trailer” generates trailers only using movie subtitles, and automated trailers accurately selected scenes present in the original trailer with an average accuracy of 47%. The CLIPSonic model [10] exhibits good performance in text-to-audio and image-to-audio synthesis, but experiences a drop in relevance when using text queries in a zero-shot setting

The extensive literature survey performed gave us a good amount of insights into the existing approaches to tackle the challenge of trailer generation. We could conclude that processing modalities separately is lighter on computation and more effective in retrieving key scenes to be added in the trailer. We could also conclude that there are various baseline models and algorithms to infer from incase of possible challenges encountered in the future.

CHAPTER - 4

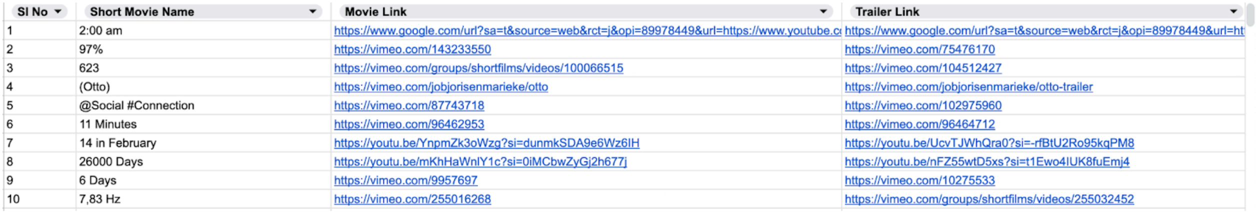
DATASET

The project focuses on generating trailers for short movies. Accordingly, our dataset comprises short movies and their trailers. A thorough analysis of the accompanying metadata including genre, duration and key elements of each movie has also been made.

Currently, 311 short movies and their trailers have been meticulously curated into a comprehensive dataset in an excel sheet. Around 30% of our dataset belongs to the ‘thriller’ genre, while another 30% comprises the ‘drama’ genre. The remaining movies are scattered among a diverse list of genres including romance, comedy, action, animation, and social awareness.

All the short movies and their corresponding trailers have been taken from the following online sources - YouTube, Vimeo, FilmShortage, ShortOfTheWeek, Reddit, FilmsShort, LetterBoxd.

Provided below is a snapshot of the curated dataset to provide a concise overview :



Sl No	Short Movie Name	Movie Link	Trailer Link
1	2:00 am	https://www.google.com/url?sa=t&source=web&rct=j&opi=89978449&url=https://www.youtube.co	https://www.google.com/url?sa=t&source=web&rct=j&opi=89978449&url=ht
2	97%	https://vimeo.com/143233550	https://vimeo.com/75476170
3	623	https://vimeo.com/groups/shortfilms/videos/100066515	https://vimeo.com/104512427
4	(Otto)	https://vimeo.com/jobjorisenmarieke/otto	https://vimeo.com/jobjorisenmarieke/otto-trailer
5	@Social #Connection	https://vimeo.com/87743718	https://vimeo.com/102975960
6	11 Minutes	https://vimeo.com/96462953	https://vimeo.com/96464712
7	14 in February	https://youtu.be/YnpmZk3oWzg?si=dunmkSDA9e6Wz6IH	https://youtu.be/UcvTJWhQra0?si=-rfBtU2Ro95kqPM8
8	26000 Days	https://youtu.be/mKhHaWnlY1c?si=0iMCbwZyGj2h677j	https://youtu.be/nFZ55wtD5xs?si=t1Ewo4IUK8fuEmj4
9	6 Days	https://vimeo.com/9957697	https://vimeo.com/10275533
10	7,83 Hz	https://vimeo.com/255016268	https://vimeo.com/groups/shortfilms/videos/255032452

Fig 4.1. Dataset Snapshot

CHAPTER - 5

SYSTEM REQUIREMENT SPECIFICATIONS

5.1 Introduction

5.1.1 Purpose

The intent of this document is to give a complete profile of the Project Requirements for “Video Trailer Generation using Multimodal Data Analysis” capstone project. Designing emotionally persuasive trailers for short films through ordinary means is a tiresome and scary job because it is hard to capture people’s interest by combining audio, video, and other metadata in an intelligent way through multimodal data analysis techniques. For this to be done there are some system requirements that you will need. They must have a strong ability when it comes to dealing with sound by enabling the removal of sounds from the radar as well as its analysis. Efficient video manipulation software should also be availed so as to facilitate proper scrutiny on the visual elements in any multimedia data. At the same time, there should be mechanisms of incorporating metadata into such systems so as to make sure these embedded files have necessary information for comprehension. In addition, there is also a need for computational capacities which can work many things on large multimedia files effectively.

5.1.2 Intended Audience and Reading Suggestions:

Intended Audience includes Production houses, Professional Video Editors and Film Enthusiasts who want to use this software to explore automatically created trailers for short movies; and Researchers who intend to work with Audio/Video Processing to improve the existing approaches for trailer generation.

5.1.3 Project Scope:

This software is intended to create trailers for short movies. The intent is to build on or improvise existing research methods for automatic trailer generation. The scope of the project also entails curating a novel dataset for short movies and their trailers along with the associated metadata including genre, language, subtitles etc.

5.2 Overall Description

5.2.1 Product Perspective:

The software is developed for crafting intelligent trailers using Multi-Modal Analysis and Deep Learning Methodologies. It can be used by Professionals to generate trailers for advertising their short movies. This can also be used by Enthusiasts for experimentation and entertainment, for movies that don’t have trailers.

5.2.1.1 System Interface

The system would contain multiple modules for Audio Processing, Video Processing, Frame Selection, Shot Selection, Trailer Scene Rearrangement, Metadata Analysis and Trailer Generation.

5.2.1.2 Software Interface

The software includes Python Libraries, Deep Learning Models and other Audio-Video Processing tools. The Software will be compatible on Linux, Mac and Windows Operating Systems.

5.2.2 Product Functions

- Given the video of a short movie, the software separates the audio and video, processes the audio and video files separately to find certain key moments.
- Based on the anomalies detected in the audio files, or attractiveness of the video frames, shot selection is done.
- The Scenes are combined in such a way that the end trailer is visually attractive and also has key moments based on the audio (music or dialogues).

5.2.3 User Classes and Characteristics

The user characteristics in this research project revolve around educational or experimental roles. Those who contribute are usually academic researchers, short-film producers or social media content creators who supply video datasets and set guidelines for creating trailers, which drives the project forward.

5.2.4 Design and Implementation Constraints:

Climax or Story Reveal: The system must be designed to correctly identify key moments like surprise reveals or tension-building scenes based on the genre of the movie. Also, ensuring that these trailer moments do not spoil the movie by revealing the story or climax is very essential.

Length of the movie and trailer: Creating a captivating trailer with the limited duration of 30-90 seconds from a 15-30 minutes short film is a huge task requiring careful selection and editing. The algorithms used have to be optimised to work effectively with this constraint.

Multimodal Analysis: There is a lack of publicly available benchmarks for processing multimodal data, especially for movie trailer generation. To evaluate the proposed software, a novel evaluation metric must be developed.

Data Acquisition: Finding a publicly available dataset of short films with clear audio and video, obtained legally and with proper permissions, is extremely challenging. Alternate data collection methods have to be explored, such as collaborating with short film festivals or communities.

5.2.5 Assumptions and Dependencies:

The main assumptions of our project would be:

1. Genre Consistency: It is assumed that the short films provided mostly belong to the horror, thriller, or mystery genres. This assumption makes it possible to create analytical algorithms that are especially suited to these narrative modalities. By concentrating on these genres, the study can go more deeply into the particular components and methods employed in these genres, improving the analysis's accuracy.

2. Content Adherence: The process of creating trailers will closely follow the short films' original content.The story won't be changed, and no outside components (like generated content) will be included.

In addition to these assumptions, there are dependencies that must be considered:

1. Deep Learning Frameworks: To create and train models for audio and video analysis, the project mostly uses deep learning frameworks like TensorFlow or PyTorch. These frameworks enable the project to extract significant information from the audio and visual components of the short films by offering strong tools and algorithms for processing and evaluating complex data.

2. Audio/Video Processing Libraries: The extraction and manipulation of features from audio and video data will be greatly aided by libraries such as Librosa or OpenCV. With the help of these libraries, the processing and manipulation of audio and video data in a multitude of ways can occur , extracting essential elements that aid in analysis and trailer creation.

3. Computational Resources: Training of Deep learning models can incur significant computational costs. Hence, in order to effectively train and improve the models, access to computational resources like GPUs or cloud computing services may be required. These resources ensure the fast and accurate analysis of the short films required for trailer curation.

A thorough study plan that takes into account these presumptions and dependencies can be created to meet the obstacles and take advantage of the opportunities this project presents. The implementation and accomplishment of the research objectives will be facilitated by the careful consideration of these elements.

5.3 Functional Requirements

The core functionalities that our Automatic Trailer Generation software will offer are as follows:

5.3.1 Separation of Audio and Video

The user must give a short movie as input to the software. The system will separate the short movie into its audio and video components.

5.3.2 Audio Processing:

The audio is processed first, to retrieve specific segments of audio that highlight the key moments of the short movie. For example, “screaming” in the case of a horror movie. The audio segment determines to what extent the dialogues, music etc., matter while a trailer is generated.

5.3.3 Video Processing:

The video processing component does the following

1. Picks the video segments corresponding to the audio segment and analyses the attractiveness
2. Selects any other frames / shots that may be visually and in terms of content, a highlight and therefore a necessity to include in the trailer

5.3.4 Meta-Data Processing:

This component is concerned with introducing some important metadata like cast and other important information which may be useful for audio or video segment selection

5.3.5 Shot Arrangement:

The video and audio segments selected will be combined and an algorithm will arrange them in a visually attractive trailer with any additional effects

5.4 Analysis Model

5.4.1 Gantt Chart

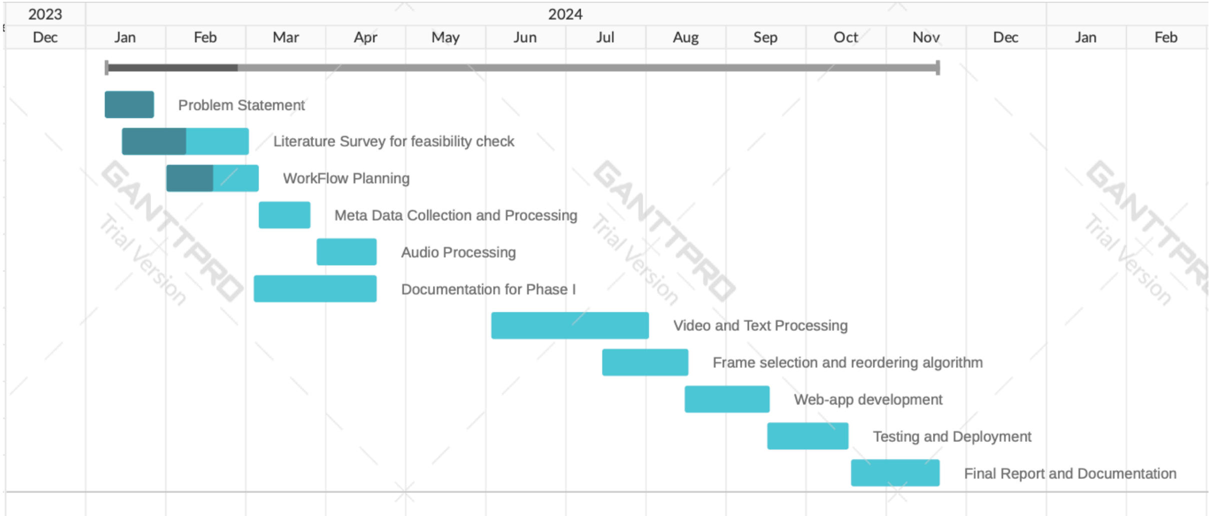


Fig 5.1. Gantt Chart

CHAPTER - 6

HIGH LEVEL DESIGN

6.1 Introduction

This document explains the High Level Design of the Capstone Project titled - "Video Trailer Generation using Multimodal Data Analysis".

This project addresses the need for effective and automatic trailer generation by using multimodal data analysis, integrating audio, video, and other metadata to intelligently craft meaningful and concise trailers that capture the attention of the viewers/audience. Our project aims to explore various Audio and Video Processing techniques for analyzing and capturing the key moments of a short movie.

6.2 Current System

Crafting emotionally resonant short movie trailers using traditional methods is laborious and challenging. Current softwares and approaches for automatic trailer generation include having a professional content creator or experienced director to provide inputs and suggestions for choosing key moments for the trailer.

6.3 Design Details

6.3.1 Novelty

The Novelty of this project is that a multi-modal analysis is performed for short movies. Audio is processed first to extract emotional moments and then the video processing is performed to extract attractive frames. Then a novel algorithm is applied for rearranging the scenes to create a captivating trailer. The metadata, like genre, is also taken into consideration.

6.3.2 Innovativeness

The project showcases a lot of innovation in the field of audio-video processing. The work expresses innovation by creating a comprehensive trailer using a multimodal data analysis approach which has been under-explored in prior work. The project is also aimed at creating a novel audio guided video processing technique and a sophisticated algorithm for ordering the obtained scenes.

This creative method helps produce a trailer with visual and auditory attractive features that captures the essence of the given short movie. The project guarantees that by using independent audio and video analysis, equal weightage will be given to both visual and audio features thereby ensuring a thorough comprehension of the short film's content.

The trailer generation process gives emphasis on not giving out the plot or disclosing the climax. Key “trailer-specific” moments are chosen through the audio driven video analysis which ensures that the highlights are selected from the movie based on a certain threshold. Through this the project aims to manipulate the trailer scenes to not completely give out the plot.

This project represents a significant advancement in the trailer creation industry due to its innovative use of multimodal data analysis techniques.

6.3.3 Interoperability

The project follows a modular approach. The significance of this approach is that it becomes open to extension. Any further features like adding additional template content can easily be integrated into the existing code modules. The existing plan of action can be split into the following modules

1. Audio Analysis Module : This feature is specifically to return key timestamps corresponding to emotionally significant audio segments which have some auditory impact (ex. A Scream)
2. Video Analysis Module : This module is to utilize the segments returned by the audio module and select highlights from the video and return the corresponding timestamps

3. Feature Fusion : This module aims to combine the features obtained from the previous modules
4. Scene Formation: The video and audio segments selected will be combined and an algorithm will arrange them to form a visually attractive trailer.

6.3.4 Performance

Performance of an application is a measure of effectiveness and user satisfaction. The users of the trailer generation software must be happy with the trailer contents they get as output

Certain aspects like memory utilization, processing speed, correctness of output etc. show the performance of an application

Utilizing cloud services like AWS to store large short movie data and generated trailers helps reduce dependence on local storage which inturn prevents system failure.

The processing speed can be increased by the use of good quality GPUs. The trailer generation software must give an appropriate trailer, else the user will not be satisfied.

6.3.5 Reliability

The goal in terms of reliability is to establish a robust system which establishes smooth, uninterrupted operations. Comprehensive error handling and fault tolerance mechanisms to deal with potential failures and overall stability of the system shall be implemented.

We will also strive to implement data integrity and backup procedure checks with the goal to prohibit any form of data loss or corruption. Rigorous types of testing would be conducted to address as many potential vulnerabilities as possible.

6.3.6 Maintainability

A modular design approach is applied to maintain the project: Audio and visual analysis components are kept separated from each other to update them independently. Documentations describing data models, methods, and system architecture are updated on a regular basis. It helps keep the code name and relevant comments to maintain the code. The code is tracked with tools such as Git and other version control systems that encourage teamwork and perform system backups. High-level testing such as unit and integration tests helps ensure that the system is still reliable and accurate even after modifications. The system should be made as easy to maintain and flexible regarding changing needs as possible.

6.3.7 Portability

To ensure a smooth operation across various systems and platforms, platform-independent Python libraries and portable deep learning frameworks for audio and visual analysis are used. Cross-platform development tools and libraries will assist in the code being compiled and run on different operating systems. The use of Docker allows for containerization in which an application and its dependencies are bundled in standardized components. A flexible and scalable way to deploy is done with the use of cloud platforms such as AWS, Azure, or Google Cloud. A steady compatibility testing method will ensure a proper working scope of the application in various settings, screen resolutions, aspect ratios, and hardware combinations. This will make the application more accessible and portable in environments.

6.3.8 Reusability

The modular architecture makes it easy to reuse it, hence efficiency in pre-processing and frame extraction. The pre-processing module splits short movies into audio and video parts. The audio module simplifies audio processing across various movies while extracting significant anomalies or sentiments. It enhances processing coherence. The video module selects visually appealing frames and can adapt to various video inputs. The scene arrangement module uses inputs from both audio and video modules to arrange segments in the correct order. It can be reused for all trailer generation use cases.

6.3.9 Application compatibility

The system can run seamlessly on any operating system like Windows, Linux or Mac OS, Also, it supports all the commonly used audio formats like WAV and MP3, and video formats like MP4, MOV, and MVA. The software also offers a variety of formats for exporting or downloading the trailers, including MP4 and MOV.

6.3.10 Resource utilization

The resource utilization depends on many factors like the length and complexity of the movie. Short movies with basic cuts and transitions will require comparatively less processing power than those of longer duration having elaborate effects and color grading.

1. CPU: Primarily for basic audio and video loading and processing
2. GPU: A hardware accelerator to boost the speed of training, processing and feature extraction time
3. RAM: The software needs RAM to store project data, including video clips, audio files, and cached information.
4. Storage: The final trailer file size will depend on the chosen format and resolution.

CHAPTER - 7

SYSTEM DESIGN

The design approach follows the order of processing Audio followed by processing Video. Audio processing is computationally lighter. Selecting distinct timestamps from processing the audio and video optimises the model's ability to select key highlights.

Proposed Methodology:

1) **Pre-Processing** - The short film is taken as input and is pre-processed to convert the file into the standard format like MP4 or MOV.
2) **Splitting Audio and Video** - It is essential to separately process the audio and video segments, to extract all the relevant features that could be important in extracting trailer-worthy moments or scenes. So, the pre-processed file is split into audio file and video file and sent to different modules for further analysis..
3) **Audio Analysis** - This module deals with analyzing the audio components, that is, the vocals or dialogues and the background music or noise. It includes the following steps:
    a) Splitting the audio into vocals and background
    b) Analysing the vocals using sentiment analysis for key moment timestamps
    c) Analysing the background music or noise using audio evaluation for key highlight timestamps
4) **Video analysis** - After getting the key timestamps from the Audio Analysis step, further analysis of the video gives the timestamps of attractive segments of the video
5) **Shot/scene selection and audio-video feature fusion** - Feature fusion entails combining the results of analysis for audio and video to select proper frames or shots and generate coherent scenes.
6) **Scene Rearrangement & Trailer Creation** - After feature fusion, the scenes are rearranged using a novel algorithm for creating the visually appealing as well as emotionally stirring trailer.

The trailer created using this approach will only have original content taken from the short movie and will not use any other external source.

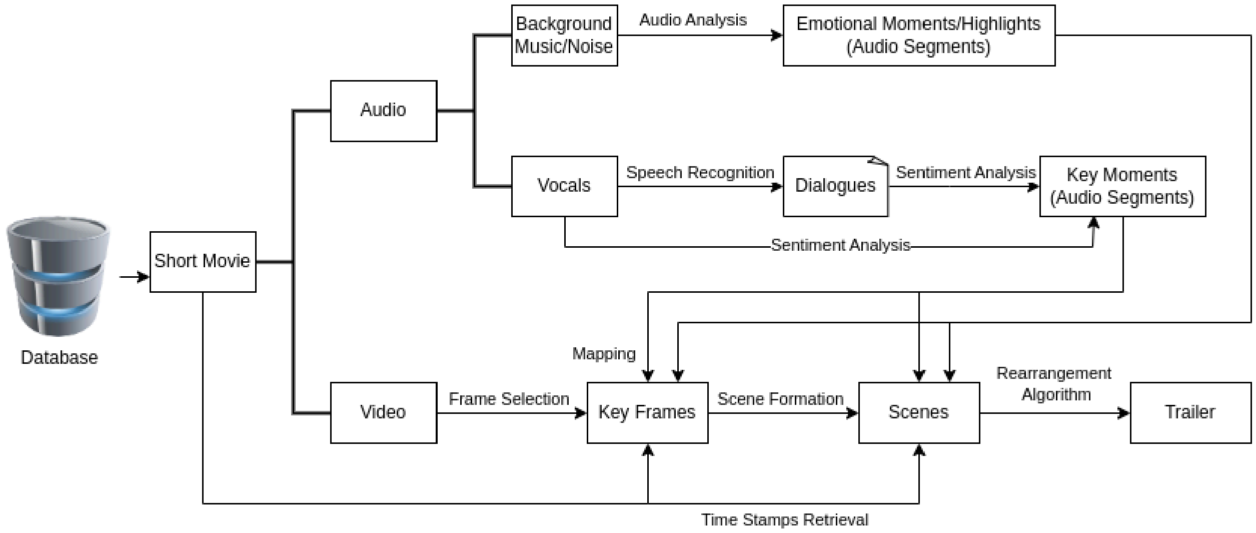


Fig 7.1 System Design

CHAPTER - 8

IMPLEMENTATION AND PSEUDO-CODE

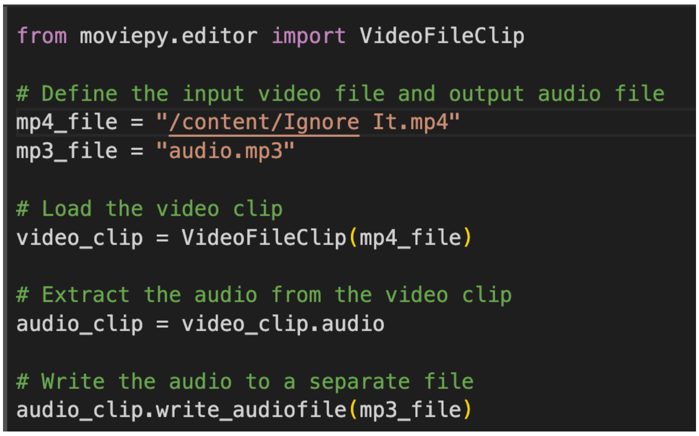
The following section details the implementation that has been performed for phase 1 and the pseudo code of the entire project as mentioned in the System Design(section 7). The implementation includes an analysis of how the audio is structured and its pattern using MEL Spectrogram and MFCC scores. These features give an understanding of what timestamps can be possibly extracted in the audio analysis phase that later guides the selection of key timestamps from the video.

The following are the details of the implementation:

A short movie named “Ignore It” [13] has been used to demonstrate the exploratory data analysis for the audio processing phase. This short film is of the “horror” genre. The movie is around 7 minutes

As seen in the system design (section 7), the video is supposed to be in mp4 or mov format. The short film is extracted to required format using “pytube” [16], a python library. Further , the audio from this short film is extracted using “moviepy” [17] for the first step of the pseudo code i.e. audio processing

Code for audio extraction - the extracted audio is saved as audio.mp3

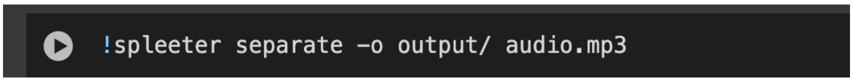


```python
from moviepy.editor import VideoFileClip

# Define the input video file and output audio file
mp4_file = "/content/Ignore It.mp4"
mp3_file = "audio.mp3"

# Load the video clip
video_clip = VideoFileClip(mp4_file)

# Extract the audio from the video clip
audio_clip = video_clip.audio

# Write the audio to a separate file
audio_clip.write_audiofile(mp3_file)
```

Fig 8.1 Code - audio extraction

Following this the extracted audio is subsequently split into vocal and accompaniment (background) audio files using "spleeter" , a python library[14].

Code for splitting audio - the "vocals and background" files are stored in /output directory



```
!spleeter separate -o output/ audio.mp3
```

Fig 8.2 Code - audio splitting

A Fourier analysis on the vocals and background audio is performed, however fourier analysis only gives a compressed visual of the audio in a fixed interval of time, and there was a need to dynamically analyse the audio to get highlight information

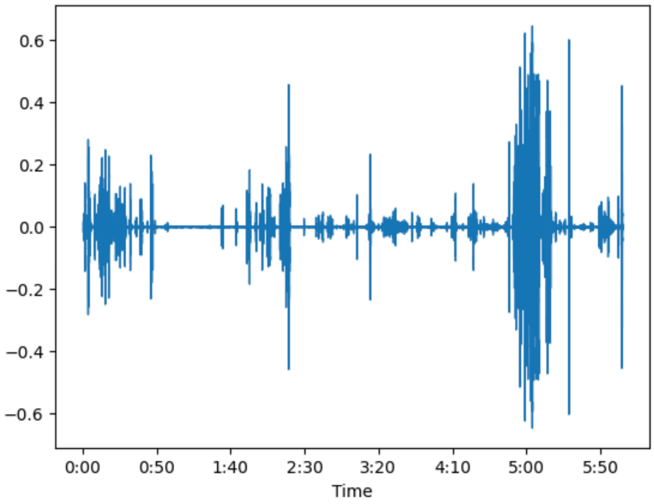
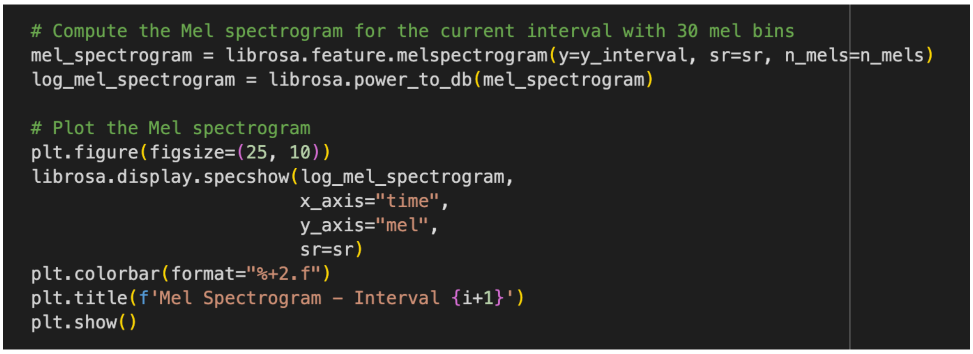


Fig 8.3 Fourier Analysis of the vocals audio

The MEL spectrogram is a spectrogram in which the y axis is in the scale of "MELs". The MEL scale normalises the pitch difference. The intensity of audio is measured in terms of Decibels instead of amplitude. Decibels make the audio easier for humans to analyse. MEL spectrograms are plotted using the "librosa" library in python [15]. The MEL spectrograms are plotted in a 10 seconds interval.

Figure 8.4 shows the code for plotting MEL spectrograms for the given audio file of the short film.



```
# Compute the Mel spectrogram for the current interval with 30 mel bins
mel_spectrogram = librosa.feature.melspectrogram(y=y_interval, sr=sr, n_mels=n_mels)
log_mel_spectrogram = librosa.power_to_db(mel_spectrogram)

# Plot the Mel spectrogram
plt.figure(figsize=(25, 10))
librosa.display.specshow(log_mel_spectrogram,
                         x_axis="time",
                         y_axis="mel",
                         sr=sr)
plt.colorbar(format="%+2.f")
plt.title(f'Mel Spectrogram - Interval {i+1}')
plt.show()
```

Fig 8.4 Code - MEL Spectrogram

Figure 8.5 shows the MEL spectrogram for the 32nd Interval (5 min 33 secs - 5 mins 43 secs). This shows intervals where there is loud periodic audio i,e. 1.5 s - 3 s depicted by close bands and darker orange indicating that the person is speaking with more intensity. The intervals 4.5s - 6s and 8.5s - 9s correspond to the man screaming “No Leave her alone !” and “Jessica, Pass”, both of which are told with tremendous intensity and fear (in the context of the movie, Jessica is about to be possessed by an evil spirit and the man, her husband is screaming). There are also continuous bands which indicate there is continuous vocals like talking.

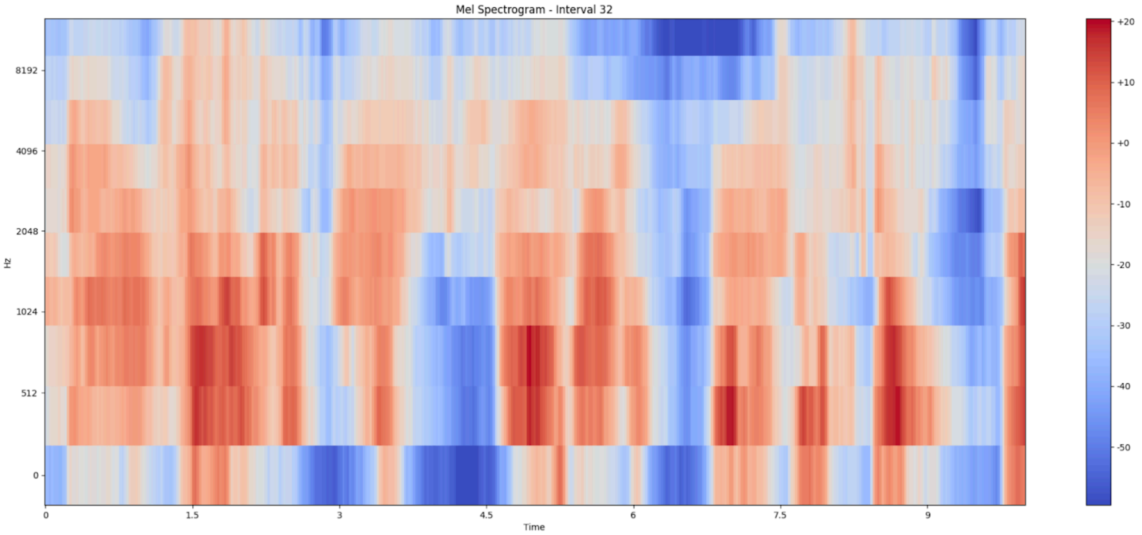


Fig 8.5 MEL Spectrogram for the vocals file

Figure 8.6 shows a comparison between MEL spectrograms for the vocals and background file. The MEL spectrogram for vocals between 4.5s and 9s shows high intensity which indicates a shout or very loud, intense talking. However the interval 1.5s - 4s is blank indicating there was no talking.

On analysing the background we can infer that there was a significant audio event, like heavy breathing when the vocals were static.

These details are important in selecting key timestamps during audio analysis. A horror short film trailer usually has one scene in which there is a loud sound like a scream, a significant tagline dialogue and either eerie silence or breathing in the background.

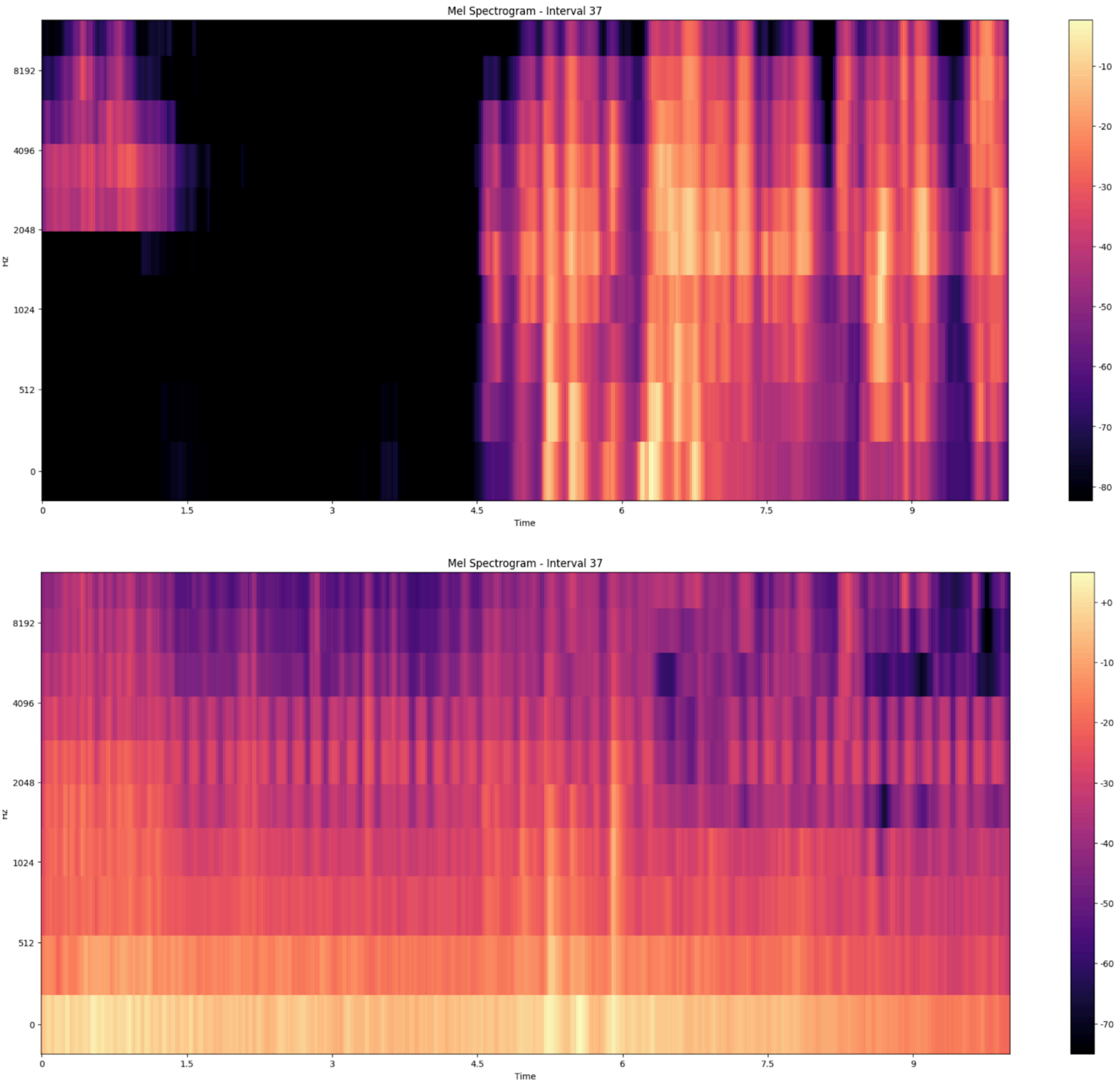


Fig 8.6 Comparison of MEL Spectrograms for Vocals and Background

Following the analysis of MEL Spectrograms , an analysis of MFCC coefficients has been made. MFCCs are a concise representation of audio features. Hence, MFCC coefficients contain information about the rate changes in the different spectrum bands. So they basically describe the shape of the spectral envelope.

MFCC are formed by applying DCT (Direct Cosine Transform) on a list of mel log powers. The MFCC are less sensitive to noise and the variation in pitch of the audio , hence making them most suitable for audio signal processing. As seen here MFCC are constricted to about 13 parameters to capture complex audio signals efficiently.

Given below is the MFCC plot of the entire short film spanning for about six minutes and 33 seconds.

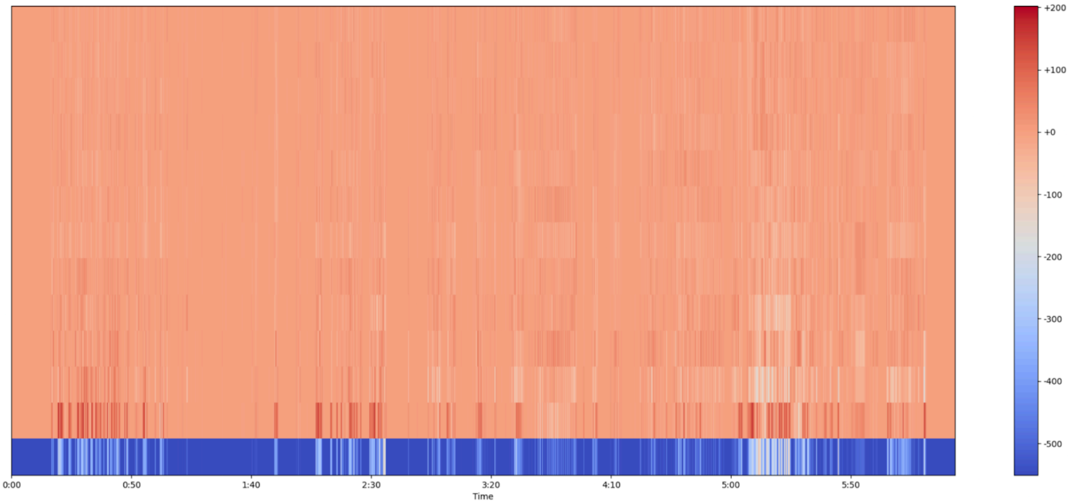
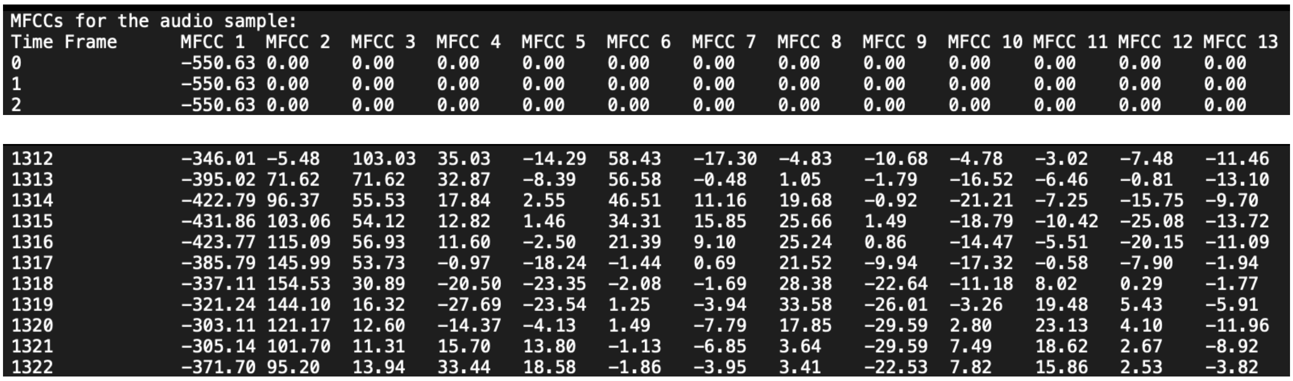


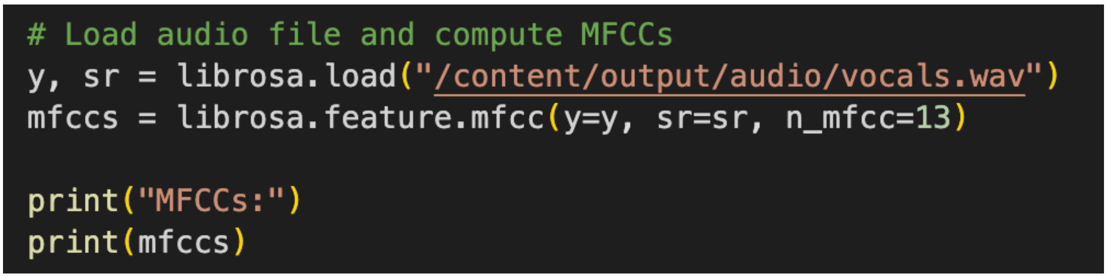
Fig 8.7 MFCC Power Spectrogram



```
MFCCs for the audio sample:
Time Frame       MFCC 1  MFCC 2  MFCC 3  MFCC 4  MFCC 5  MFCC 6  MFCC 7  MFCC 8  MFCC 9  MFCC 10 MFCC 11 MFCC 12 MFCC 13
0                -550.63 0.00    0.00    0.00    0.00    0.00    0.00    0.00    0.00    0.00    0.00    0.00    0.00
1                -550.63 0.00    0.00    0.00    0.00    0.00    0.00    0.00    0.00    0.00    0.00    0.00    0.00
2                -550.63 0.00    0.00    0.00    0.00    0.00    0.00    0.00    0.00    0.00    0.00    0.00    0.00

1312             -346.01 -5.48   103.03  35.03   -14.29  58.43   -17.30  -4.83   -10.68  -4.78   -3.02   -7.48   -11.46
1313             -395.02 71.62   71.62   32.87   -8.39   56.58   -0.48   1.05    -1.79   -16.52  -6.46   -0.81   -13.10
1314             -422.79 96.37   55.53   17.84   2.55    46.51   11.16   19.68   -0.92   -21.21  -7.25   -15.75  -9.70
1315             -431.86 103.06  54.12   12.82   1.46    34.31   15.85   25.66   1.49    -18.79  -10.42  -25.08  -13.72
1316             -423.77 115.09  56.93   11.60   -2.50   21.39   9.10    25.24   0.86    -14.47  -5.51   -20.15  -11.09
1317             -385.79 145.99  53.73   -0.97   -18.24  -1.44   0.69    21.52   -9.94   -17.32  -0.58   -7.90   -1.94
1318             -337.11 154.53  30.89   -20.50  -23.35  -2.08   -1.69   28.38   -22.64  -11.18  8.02    0.29    -1.77
1319             -321.24 144.10  16.32   -27.69  -23.54  1.25    -3.94   33.58   -26.01  -3.26   19.48   5.43    -5.91
1320             -303.11 121.17  12.60   -14.37  -4.13   1.49    -7.79   17.85   -29.59  2.80    23.13   4.10    -11.96
1321             -305.14 101.70  11.31   15.70   13.80   -1.13   -6.85   3.64    -29.59  7.49    18.62   2.67    -8.92
1322             -371.70 95.20   13.94   33.44   18.58   -1.86   -3.95   3.41    -22.53  7.82    15.86   2.53    -3.82
```

Fig 8.8 MFCC Coefficient Values

The values above represent the MFCC features against time frames . It can be noted that the 1st feature describes the overall signal energy while the next twelve features represent spectral features such as shape, slope, roughness, flatness etc . Here the negative values under MFCC1 show that in the time frames as seen above , indicates a high signal energy that seems to be consistent in these frames. There also occurs frames where the 1st coefficient contains positive values which indicate lower signal energy. MFCC plot and coefficient values can be obtained using “librosa”.



```
# Load audio file and compute MFCCs
y, sr = librosa.load("/content/output/audio/vocals.wav")
mfccs = librosa.feature.mfcc(y=y, sr=sr, n_mfcc=13)

print("MFCCs:")
print(mfccs)
```

Fig 8.9 Code - MFCC Coefficients

Pseudo Code

```
Step 1 : ProcessShortFilm( shortFilmFile )
    // read the short film
    // getters and setters for metadata
    // return the film in mp4 or mov format

Step 2 : SplitFilm()
    //ProcessShortFIlm( shortFilmFIle )
    // Separate audio and video files using libraries
    // Output: audioFile, videoFile

Step 3 : AudioAnalysis(audioFile)
            SplitAudioFile(audioFile)
                //returns a split of audio file into vocals and background sound
                // Output : vocals.mp3/wav , background.mp3/wav
            vocalAudioAnalysis(vocalsFile)
```

```
                        // perform vocal analysis to select key timestamps
                  bgAudioAnalysis (bgFile)
                        // perform background analysis to select key timestamps
      // return key timestamps

Step 4 : VideoAnalysis (videoFile)
                  AudioAnalysis (audioFile)
                        // returns a list of key timestamps
                  //  perform visual analysis based on attractiveness and emotion


            return [ all timestamps ]
Step 5 : FeatureFusion
                  // VideoAnalysis (videoFile)
              // selection of frames / shots from the retrieved timestamps
              // Fusion of features
             // scene formation
            return [ scenes ]
Step 6 : Scene Rearrangement
                        FeatureFusion ()
                        // reorders the retrieved scenes based on an arrangement algo
                        // compile the trailer
                  returns the Trailer
```

CHAPTER - 9

CONCLUSION OF CAPSTONE PROJECT PHASE - 1

The first phase of the capstone project included a thorough analysis to identify viable topics within the realm of audio and video processing. The techniques analysed will be a benchmark in the audio video processing domain. Exploratory analysis and extensive literature survey has shown that the audio processing technique being proposed is feasible and reliable to obtain attractive audio segments. As shown in the Implementation, the audio analysis done using MEL Spectrogram and MFCC, proves that the approach of separately processing audio and video is helpful in obtaining key moments in the movie. The subsequent workflow being proposed in the System Design is also inspired from prior research work. These techniques ensure both novelty and feasibility.

CHAPTER - 10

PLAN OF WORK FOR CAPSTONE PROJECT PHASE - 2

Exploratory analysis on audio content of the movies and trailers have given us an abundance of information. The subsequent workflow is as follows :

1. Building a novel architecture to obtain timestamps of significant audio segments
2. Extending the model to utilize these timestamps to guide the process of analysing video to obtain timestamps of attractive segments
3. Building a feature fusion model to accept timestamps from both audio and video content and giving the most important “trailer-appropriate” scenes
4. Creating a novel algorithm for scene arrangement and finally generating an emotion aware, intelligent trailer.
5. Performing Unit Testing, Integration Testing and a custom test for comparing the generated trailer with the existing trailer in the dataset

REFERENCES

[1] P. Papalampidi, “Structure-aware narrative summarization from multiple views,” era.ed.ac.uk, Jan. 2023, doi: https://doi.org/10.7488/era/2946.

[2] Wang, L., Liu, D., Puri, R., Metaxas, D.N. (2020). Learning Trailer Moments in Full-Length Movies with Co-Contrastive Attention. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, JM. (eds) Computer Vision – ECCV 2020. ECCV 2020. Lecture Notes in Computer Science(), vol 12363. Springer, Cham. https://doi.org/10.1007/978-3-030-58523-5_18

[3] Gan, Bei & Shu, Xiujun & Qiao, Ruizhi & Wu, Haoqian & Chen, Keyu & Li, Hanjun & Ren, Bo. (2023). Collaborative Noisy Label Cleaner: Learning Scene-aware Trailers for Multi-modal Highlight Detection in Movies.

[4] Wei, Fanyue et al. “Learning Pixel-Level Distinctions for Video Highlight Detection.” 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022): 3063-3072.

[5] C. Bretti, P. Mettes, Hendrik Vincent Koops, Daan Odijk, and Nanne van Noord, “Find the Cliffhanger: Multi-modal Trailerness in Soap Operas,” Lecture Notes in Computer Science, pp. 199–212, Jan. 2024, doi: https://doi.org/10.1007/978-3-031-53308-2_15.

[6] J. Sheng, Y. Chen, Y. Li, and L. Li, “Embedded learning for computerized production of movie trailers,” Multimedia Tools and Applications, vol. 77, no. 22, pp. 29347–29365, Apr. 2018, doi: https://doi.org/10.1007/s11042-018-5943-3.

[7] N. Sadoughi et al., “MEGA: Multimodal Alignment Aggregation and Distillation For Cinematic Video Segmentation Supplementary Material.” Accessed:Mar.14,2024.[Online]. Available: https://openaccess.thecvf.com/content/ICCV2023/supplemental/Sadoughi_MEGA_Multimodal_Alignment_ICCV_2023_supplemental.pdf

[8] P. Mishra, C. Diwan, S. Srinivasa, and G. Srinivasaraghavan, “AI based approach to trailer generation for online educational courses,” CSI Transactions on ICT, vol. 11, no. 4, pp. 193–201, Nov. 2023, doi: https://doi.org/10.1007/s40012-023-00390-1.

[9] K. Porwal, H. Srivastava, R. Gupta, S. Pratap Mall, and N. Gupta, “Video Transcription and Summarization using NLP,” SSRN Electronic Journal, 2022, doi: https://doi.org/10.2139/ssrn.4157647.

[10] Dong, Hao-Wen et al. “CLIPSonic: Text-to-Audio Synthesis with Unlabeled Videos and Pretrained Language-Vision Models.” 2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) (2023): 1-5.

[11] H. Kakimoto, Y. Wang, Y. Kawai and K. Sumiya, "Extraction of Movie Trailer Biases Based on Editing Features for Trailer Generation," 2018 IEEE International Symposium on Multimedia (ISM), Taichung, Taiwan, 2018, pp. 204-208, doi: 10.1109/ISM.2018.000-6.

[12] M. Hesham, B. Hani, N. Fouad and E. Amer, "Smart trailer: Automatic generation of movie trailer using only subtitles," 2018 First International Workshop on Deep and Representation Learning (IWDRL), Cairo, Egypt, 2018, pp. 26-30, doi: 10.1109/IWDRL.2018.8358211

[13] Ignore It: https://www.youtube.com/watch?v=hs3paMLb9Qg

[14] Spleeter library: https://pypi.org/project/spleeter/

[15] Librosa MEL Spectrogram :
https://librosa.org/doc/latest/generated/librosa.feature.melspectrogram.html

[16] MoviePy : https://pypi.org/project/moviepy/

[17] Pytube : https://pytube.io/en/latest/

Prajna R PW24_SBN_08 Report - Google Docs

ORIGINALITY REPORT

7%	6%	6%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	link.springer.com Internet Source	1%
2	arxiv.org Internet Source	1%
3	Jiachuan Sheng, Yaqi Chen, Yuzhi Li, Liang Li. "Embedded learning for computerized production of movie trailers", Multimedia Tools and Applications, 2018 Publication	1%
4	journal.linguaculture.ro Internet Source	<1%
5	pt.scribd.com Internet Source	<1%
6	Yibo Sun, Zhe Sun, Weitong Chen. "The evolution of object detection methods", Engineering Applications of Artificial Intelligence, 2024 Publication	<1%
7	Submitted to Edge Hill University Student Paper	<1%

8	www.coursehero.com Internet Source	<1%
9	Honoka Kakimoto, Yuanyuan Wang, Yukiko Kawai, Kazutoshi Sumiya. "Extraction of Movie Trailer Biases Based on Editing Features for Trailer Generation", 2018 IEEE International Symposium on Multimedia (ISM), 2018 Publication	<1%
10	dblp.dagstuhl.de Internet Source	<1%
11	deepai.org Internet Source	<1%
12	"MultiMedia Modeling", Springer Science and Business Media LLC, 2024 Publication	<1%
13	wsl.iiitb.ac.in Internet Source	<1%
14	ojs.aaai.org Internet Source	<1%
15	salu133445.github.io Internet Source	<1%
16	"Computer Vision – ECCV 2020", Springer Science and Business Media LLC, 2020 Publication	<1%

17	research.aalto.fi Internet Source	<1 %
18	medium.com Internet Source	<1 %
19	"Information and Communication Technologies", Springer Nature, 2010 Publication	<1 %
20	Nishant Rai, Haofeng Chen, Jingwei Ji, Rishi Desai, Kazuki Kozuka, Shun Ishizaka, Ehsan Adeli, Juan Carlos Niebles. "Home Action Genome: Cooperative Compositional Action Understanding", 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021 Publication	<1 %
21	Submitted to University of Tennessee Knoxville Student Paper	<1 %
22	lestehjournal.ru Internet Source	<1 %
23	Eslam Amer, Ayman Nabil. "A Framework to Automate the generation of movies' trailers using only subtitles", Proceedings of the 7th International Conference on Software and Information Engineering - ICSIE '18, 2018 Publication	<1 %

24	Submitted to University of Canberra Student Paper	<1 %
25	assets.amazon.science Internet Source	<1 %
26	openaccess.thecvf.com Internet Source	<1 %
27	export.arxiv.org Internet Source	<1 %
28	www.research.ed.ac.uk Internet Source	<1 %
29	Hao-Wen Dong, Xiaoyu Liu, Jordi Pons, Gautam Bhattacharya et al. "CLIPSonic: Text-to-Audio Synthesis with Unlabeled Videos and Pretrained Language-Vision Models", 2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 2023 Publication	<1 %
30	Najmeh Sadoughi, Xinyu Li, Avijit Vajpayee, David Fan, Bing Shuai, Hector Santos-Villalobos, Vimal Bhat, Rohith Mv. "MEGA: Multimodal Alignment Aggregation and Distillation For Cinematic Video Segmentation", 2023 IEEE/CVF International Conference on Computer Vision (ICCV), 2023 Publication	<1 %

31	Prakhar Mishra, Chaitali Diwan, Srinath Srinivasa, G. Srinivasaraghavan. "AI based approach to trailer generation for online educational courses", CSI Transactions on ICT, 2023 Publication	<1 %
32	"Advances in Computing and Information Technology", Springer Nature, 2013 Publication	<1 %

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		

Gmail

Prajna R <prajna.ramamurthy@gmail.com>

Request for Plagiarism and AI Check for Capstone Report

plagiarism@_Library PESU <original@pes.edu> 6 May 2024 at 14:30
To: Prajna R <prajna.ramamurthy@gmail.com>
Cc: Surabhi Narayan PESU RR CSE Staff <surabhinarayan@pes.edu>

Dear Student

Please find the attached reports

Plagiarism -7%
AI Content - 0%

Best regards
Savitha

[Quoted text hidden]

Prajna R PW24_SBN_08 Report - Google Docs-1.pdf
5662K