Video Trailer Generation using Multimodal Data Analysis

Nikhil Giridhar
Dept. of Computer Science and Engineering
PES University
Bangalore, India
nikhilgiridhar3@gmail.com

Prajna R
Dept. of Computer Science and Engineering
PES University
Bangalore, India
prajna.ramamurthy@gmail.com

Pranathi Praveen
Dept. of Computer Science and Engineering
PES University
Bangalore, India
pranathipraveen03@gmail.com

Shreeja Rajesh
Dept. of Computer Science and Engineering
PES University
Bangalore, India
shreejar197@gmail.com

Dr. Surabhi Narayan
Dept. of Computer Science and Engineering
PES University
Bangalore, India
surabhinarayan@pes.edu

***Abstract* — Trailers have been a vital part of the entertainment industry to pique audience interest. The process of trailer generation has been evolving to provide effective trailers to the public. In this paper we propose a novel approach by utilizing the audio and video modalities to extract key trailer segments that are compiled to create an effective trailer which highlights the impactful scenes of the film. In contrast to previous work in this domain , the novel approach gives greater emphasis to the auditory features, since auditory features have a significant impact on the film in the horror genre. Our approach introduces an audio guided visual model that compliments the acoustic features ensuring that the extracted segments are key moments both in terms of audio and video. The results obtained indicate that a significant proportion of predicted segments are deemed highly trailer.**

Keywords—multimodal, trailer, Siamese Network, Support Vector Machine, audio, visual.

I. Introduction

Within the multimedia context, trailers are among the most effective ways to capture the audience's interest in wanting to know more and promoting content. An efficiently created trailer is a wrap-up of a movie, game, or series in which only the best moments are aired to entice future viewers. The combination of visual and audio in the trailers is actually what communicates the essence of that content; thus, it becomes the main determinant of viewership and box-office outcomes.

A meaningful trailer, which really captures the attention of the audience, is not an easy process to create. In most instances, these editors need to carefully choose the key moments out of countless clips that can only visually impact and at the same time be able to preserve the storyline of the movie. This does not stop here. These must be woven into an interesting and understandable snippet. This can be quite a cumbersome and time-consuming task, leaving filmmakers with less time to devote to the creative aspects of their projects. Indeed, automated trailer creation based on this multimodal analysis will revolutionize filmmaking in such areas.

It is a technology that will process audio and video using the most advanced algorithms in AI and in machine learning. This simply means it will create trailers by itself with no human assistance whatsoever. Such systems bear the advantage of acceleration in production terms of trailer development against an editor who cannot catch the constancy and insights unique to such systems. This will also automate the generation of trailers and reduce the burden on creators of the movie, hence rapid iterations can be produced and scale in support of content promotion. Essentially, the technology could democratize the creation of high-quality trailers, making it feasible and smooth for a wide range of content producers. Various studies report that there is great influence from trailers on the behaviors and expectations of audiences.

For example, a study undertaken by Johnston et al. (2016) has evidenced that watching trailers can be framed as consumption practice in and of itself, independent of the consumption of watching feature films3. In doing so, they identified that trailers have both informational as well as prefigurative value for the audience, driven by factors such as emotional attachment, cultural value, and social expectation. According to other research by Kuppelwieser and Finsterwalder (2012)[2], a trailer forms an expectation of the consumer about the contents of any film and its quality, where actors and genre are in leading positions:2. In all, the automation of trailer generation will not only smoothen the process of production but can also improve engagement and satisfaction on behalf of audiences. Through advanced technologies, content creators may prepare efficient and effective trailers contributing to the very success of the content they promote.

II. Literature Survey

The review of research papers brings us to important findings that shape, inform, and reform our study.

Novel approaches have recently been suggested that automatically and semi-automatically identify relevant trailer moments directly from video analysis by automatic scene segmentation into shots, turning point detection, and identification of trailer-suitable moments. This is further

enhanced through highlight detection based on visually appealing and contextually relevant frames. Multimodal analysis combines visual, audio, and textual cues to pinpoint scenes that best depict the movie. Embedded learning refines this process while factors such as editing biases, emphases on scenes, and focuses on characters add up to the dimensions of trailer component identifications in order to understand the evaluation factors of a trailer. The following literature review represents some of the different ways and advancements in the field of automatic trailer generation, putting a focus on multimodal analysis. It discusses crucial works that shaped not only the understanding but also developed the methods to find impactful video moments using audio, visual, and text data for optimal trailer creation. These have created foundations for making automation in trailer generation effective and efficient.

Papalampidi et al. [3] proposes new techniques for the automatic identification of key moments by video analysis, focusing on the movie's storytelling pattern. Wang et al. [4] presented the model CCANet, which performs the selection of trailer highlights using both highlight detection and multimodal approaches to capture relevant frames in movies both visually and contextually. Gan, Bei et al. propose learning movie highlights from noisy labels by training a model on trailers to segment movies into scenes. It shall use a new method called Collaborative Noisy Label Cleaner (CLC) which will handle the noisy labels, realizing not all scenes in the trailer contain highlight moments. Wei, Fanyue et al. [6] introduced a method of "Pixel-Level Distinction Video Highlight Detection" which can model movies at the pixel level, learning both temporal and spatial relations and subtle context to detect viewer-appealing segments of unedited videos. In [7], Bretti, Carlo, et al. introduce a multi-modal method based on both visual and subtitle data that predicts the "trailerness" of TV episodes and films to help editors choose the most optimal moments for inclusion in trailers. The Trailerness Transformer can improve the results over traditional methods by combining the predictions of various data streams to improve trailer-worthy scene selection, especially in narrative-based videos such as soap operas.

Work on embedded learning for the identification of common key components for trailers in Sheng, Jiachuan, et al. [8] proposes an automated trailer generation method by leveraging the VGG-F model in feature extraction and matching with SURF. This kind of approach gives high accuracy to those action-packed films by stitching clips around key frames but struggle when genre films rely on atmospheric elements. Sadoughi et al. propose MEGA, a multimodal alignment and fusion approach for effectively processing long-form videos with high efficiency, surpassing other models in scene and act segmentation. MEGA takes up a scalable method utilizing multiple media modalities and knowledge distillation. However, it would have more improvements that could add to actor identification to improve segmentation. More information is obtained about automatic summarization in P. Mishra et al., which will be supported by the textual information and multimodal data. Papers such as K. Porwal et al. [11] and Dong et al. [12] deal with deciphering video, audio, and text files with a view to deriving meaningful insights that classify and synthesize frames using language-vision models. Experiments on the influence of editing biases in H. Kakimoto et. al. [13] and efforts towards developing frameworks for genre classification based on textual features extracted from movie subtitles in M. Hesham et. al. [14] make up a complete platform for advancing video trailer generation techniques.

Although there have been significant advances in the area of automatic trailer generation, substantial methodological gaps still exist regarding integration and scalability for different media types. Although highly accurate in certain contexts, as revealed by many studies, there is still much room for improvement regarding segmentation accuracy and integration of visual, audio, and textual data. Further, other methods such as MEGA by Sadoughi et al. [9] offer very promising results but also bear further extensions; for example, identification of actors could be made for improving performance.

Furthermore, most works operate either on video or on audio components separately, with no approach on the holistic utilization of multimodal data to make the most of its benefit. While works such as automatic summarization by P. Mishra et al. [10], and the works of K. Porwal et al. [11] and Dong et al. [12] on language-vision models indicate that multiple data modalities must be integrated, a more integrated framework is warranted. This work will attempt to fill in all these lacunas by proposing an end-to-end methodology that will include advanced machine learning techniques with multimodal analysis so as to come up with a high-quality, engaging trailer.

III. Tools And Methods

In this section, the repository is discussed initially, followed by the tools and technologies and finally our novel method of multimodal trailer generation.

A. Dataset

The data repository consists of links compiled from various sources like Youtube, Director’s Note, Short of The Week, vimeo and other online resources. The data was mainly compiled from publicly available and reputable data sources. These platforms cater to short movie enthusiasts as well as serve as a stage for the directors to launch their upcoming movies and trailers to film festivals and garner new audiences’ attention. The total number of movies (with corresponding trailers) is 311. The average duration of the movies is around 12 minutes. The average duration of the trailers is around 50 seconds. The dataset has 5 columns. The movie name, movie video link, movie duration, trailer video link and trailer duration. A majority of these movies can be categorized as horror-thriller genre.

The data in the repository are readily available, high auditory and visual quality, and age appropriate feature content. 60% of the curated dataset is used for the training process, 20% for testing and the other 20% for validation purposes.

B. Tools and Technologies

Google Colaboratory which is a computational environment enables users to run code in a Jupyter Notebook environment. It has access to crucial computer resources like Google's GPU and TPU. This study made use of the

Python programming language (version 3.10.6) and its built-in libraries (Table 1).

TABLE I
TOOLS AND TECHNOLOGIES

Tool Used	***Purpose***
Librosa	Audio Feature Extraction
Sklearn	Data Analysis
Tensorflow - Keras	Building Machine Learning models
Numpy	Scientific Computing and feature storage
youtube_dl	Download videos directly from youtube
moviepy	Concatenate segments to form trailer video

C. Evaluation metrics

Evaluation metrics aid in the evaluation of performance of the machine learning models. The parameters such as Hamming Score(HS), Intersection Over Union(IOU) and Task Accuracy(TA) help assess the efficiency of the machine learning models.

Hamming Score (HS) : This score represents the exactness of the predictions made by comparing the predicted labels to the true labels.

Intersection Over Union (IOU) : This score represents the area of the intersection over the union of the predicted segments and the ground truth

Task Accuracy (TA) : measures the ratio between correctly predicted instances and the total number of instances, showing how well the model performs in a given task.

IV. Proposed Methodology

As mentioned earlier in Section II the previously proposed methods which focus rather singly on either the audio or the video aspects while generating the trailers. Hence we propose a novel approach below where the audio processing precedes the video processing ,making this approach a multimodal one. This approach greatly helps in extracting information from both the modalities, and also aids in reducing the compute resources required.

The data which is in the form of links extracted from various resources is preprocessed to extract the audio and video files in the required format. The data is subjected to initial data analysis where a range of audio and video features are examined, including MEL spectrograms, Mel-frequency cepstral coefficients, color spectrograms, Weibull distributions, and color histograms. On initial data analysis, we observed that results for the genre of horror were exceptional and since our novel proposed method is audio guided video processing , these results were found to be insightful. Hence the genre of horror was chosen for our novel approach.

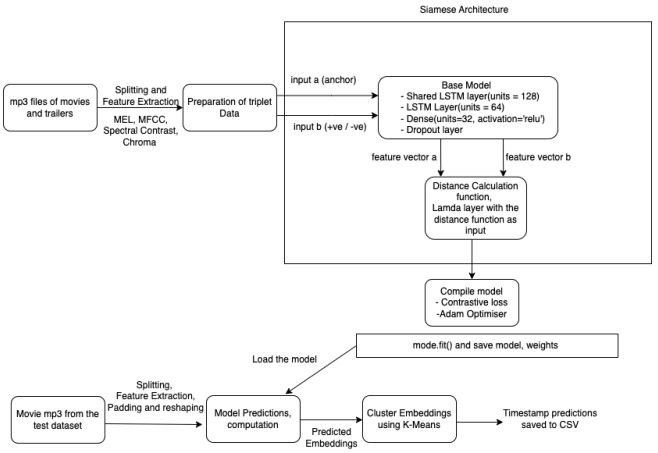


Fig. 1(a) Audio Model Architecture

As mentioned above regarding the audio processing, the audio files are first split into segments with a duration of 10 seconds for both movie and trailer. MEL, MFCC, chroma and spectral contrast features are extracted for all the segments. This is followed by the preparation of triplet data. For each movie segment (anchor segment) a corresponding trailer segment with similar features is paired. Positive pairs, drawn from trailers using the same feature types, are labeled "1" to signify a matching movie-trailer segment. Negative pairs are generated by reversing this setup: trailer segments act as anchors, paired with non-matching movie segments, and are labeled "0" to indicate dissimilarity.

A Siamese network consisting of two identical LSTM subnetworks is trained on this triplet data. Each LSTM processes the input sequence and generates a fixed size embedding. This model is compiled with Adam optimizer and contrastive loss function. Euclidean distance is computed to get the similarity between input pairs.

Once the model is compiled and trained , in regards to the prediction, initially the features such as MFCC, Mel-spectrogram, Chroma, and Spectral Contrast are extracted followed by averaging to form a single feature vector. Audio segments are padded to a uniform length with zero-padding to ensure the consistent input shapes for machine learning models.Further K Means clustering is employed to find unique clustering for the features. Hence the key segments are identified based on their proximity to the cluster centroids.

Further audio features are standardized using Standard Scaler , which is given as an input to the Siamese LSTM model. Finally the predictions are made, and significant segments are identified using the K Means clustering. The identified key moments are then saved to a CSV file which is given as the output.

The audio timestamps extracted during audio processing are used to guide the video processing. During video processing

the video file is processed using the State of the Art PySceneDetect [19] algorithm, that segments the video into scenes and individual frames are extracted from and stored.

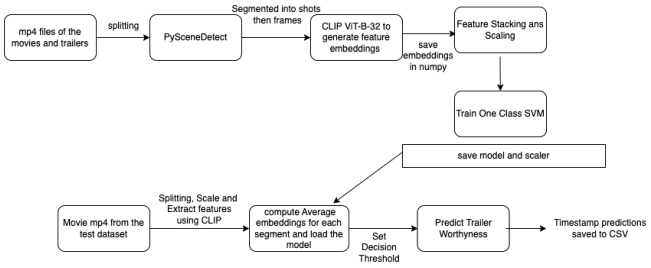


Fig. 1(b) Video Model Architecture

CLIP [20] is a neural network trained on image-text pairs, allowing it to predict relevant text for any image through natural language prompts. The image and its associated text are mapped to the same vector space. The CLIP model (ViT-B/32) is loaded and features from the movie and trailer frames are extracted, averaging them to obtain a single feature vector per scene.

Movie features (normal data) and trailer features (anomalies) are stacked and scaled using Standard Scaler. A one class Support vector machine (OCSVM) is trained to distinguish the anomalies from normal movie features. The model operates on the principle of novelty detection, where it learns a decision boundary around the normal data points (movie features) and identifies any new data points that fall outside this boundary as anomalies.

The trained model and scaler is saved. When timestamps from audio analysis for a new movie are fed to the model, the video is split into five second segments and frames are extracted from them. For each frame, the frames are extracted using the CLIP model, preprocessed, and averaged to represent each segment. Extracted features were scaled and fed into the One-Class SVM model to predict trailer-worthiness, with a decision threshold score determining the distance from the boundary. The timestamps closer to the boundary are considered as the key moments for the prediction and hence returned as the output in the form of a CSV file. Hence the selected timestamps are then matched to the corresponding movie and the returned segments are pieced together to give a trailer.

In summary , our novel methodology i.e, audio guided video processing describes a methodology for automatic trailer generation using advanced machine learning techniques coupled with multimodal analysis.It hence promises to reduce, by a huge margin, production time and cost traditionally used in the creation of trailers while the output is assured to be engaging the audiences.

V. RESULT

The split tasks of audio video processing proved not only to be computationally effective but also showed results that gave importance to the most impactful auditory and visual features that can be expected in a horror film. The final segments extracted highlighted key acoustic elements such as screams and unsettling atmospheres and key visual elements like darkness and fear filled imagery.

As mentioned in section III , the evaluated metrics aids in the effectiveness of the Machine learning models. As previously stated , three parameters Hamming Score(HS), Intersection Over Union(IOU) and Task Accuracy(TA) have been considered.

Hamming Score as a metric here represents the ratio of the number of segments that are highly trailer worthy(closer to the decision threshold) to the total number of extracted segments. When tested over 20 percent of the dataset the average Hamming Score was 0.6930. This showed that over 60 percent of predicted segments were found to be highly trailer worthy. Fig 2. illustrates the top 10 Hamming Scores for the movies analyzed.

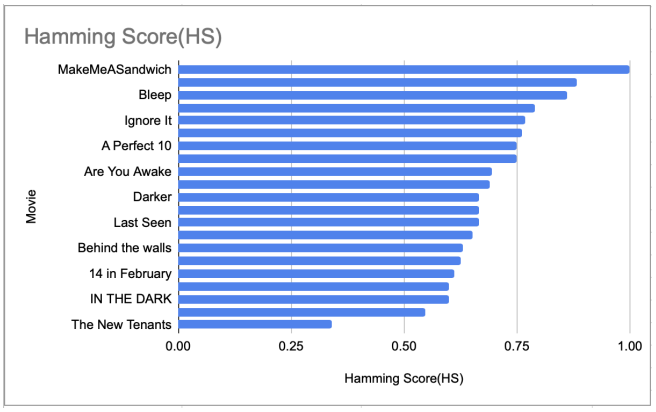


Fig. 2. Top ten Hamming Scores

Intersection Over Union is the ratio of the number of segments that lie in both extracted trailer segments and the actual trailer segments(ground truth) to the union of trailer worthy segments between the original trailer segments and the extracted trailer segments. The average IOU was computed as 0.3455. This indicates that over 30 percent of the extracted segments are present in the ground truth while there is a scope for other trailer worthy segments that could be used for the trailer. The top 10 IOU Scores are displayed in Fig 3.

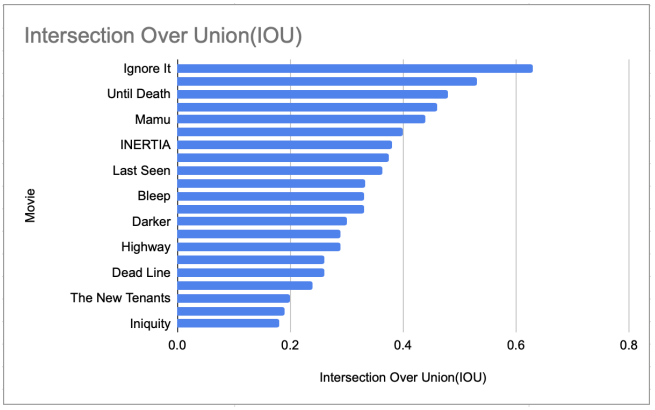
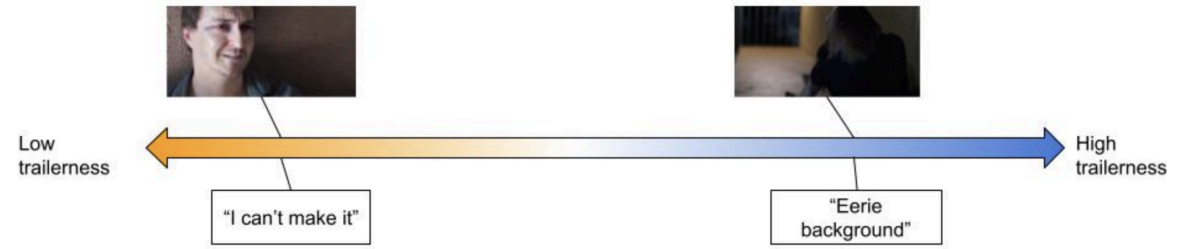
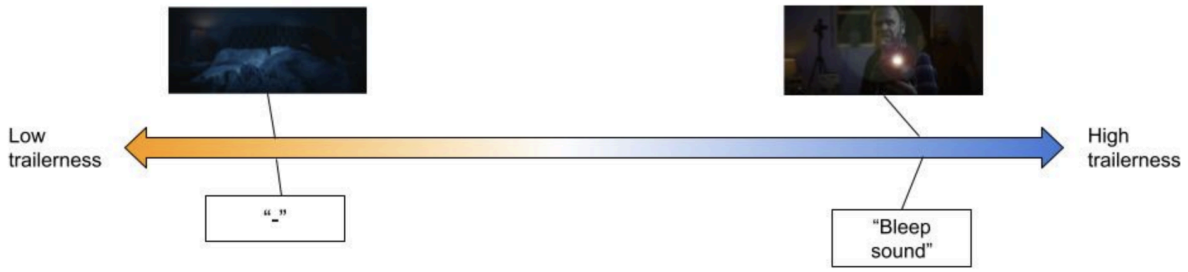


Fig. 3. Top ten IOU Scores



(a) Beware of Zombies



(b) Bleep

Fig 5. Visual and Audio Element Impact on Trailer-Worthy Scores for Beware of Zombies(a) and Bleep(b)

Task Accuracy here is the ratio of the number of predicted segments that are present in ground truth to the total number of segments. When computed over the testing data ,the average Task Accuracy was 0.5625.This likely indicates that over 50 percent of the extracted segments were predicted correctly. Fig 4. represents the top 10 Task Accuracy scores among the movies tested.

.

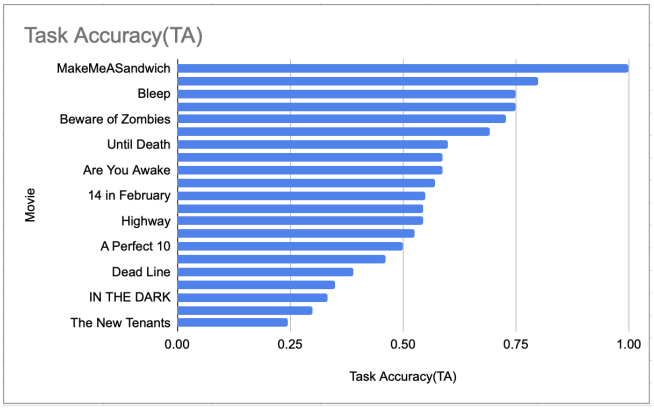


Fig. 4. Top ten Task Accuracy Scores

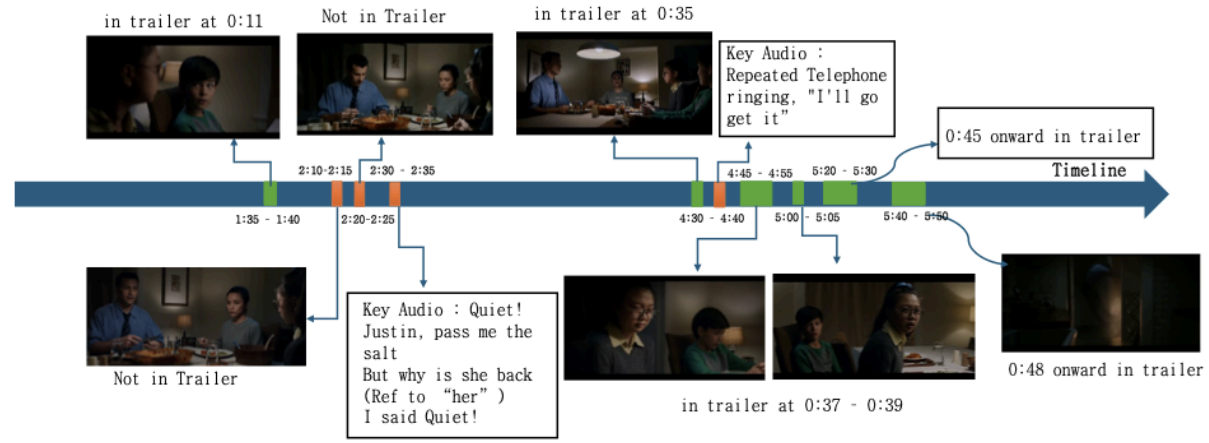
A. Quantitative Results

All of the above analyses on trailer-worthy scores in various scenes are a very important insight into the inter-relation between visual and audio, illustrated in Fig. xx which represents trailer worthiness of two short films, Beware of Zombies, and Bleep. Primary attention is given to scene characteristics, like visual brightness, presence of audio, impacting computed trailer-worthy scores.

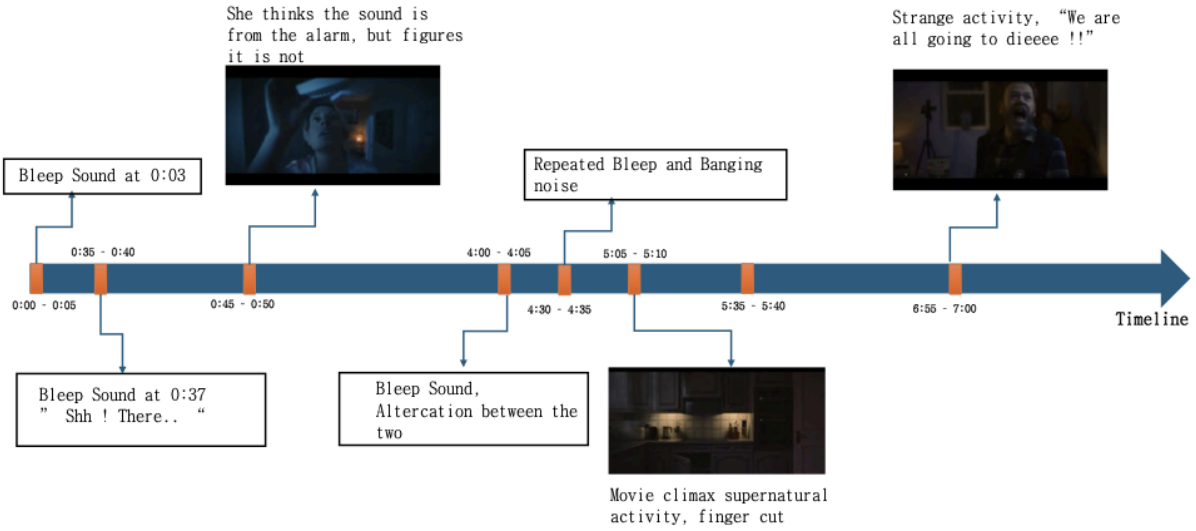
The scores taken for the trailer-worthy present an interesting trend for Beware of Zombies, a short film dominated by visually bright scenes. As seen by Fig. 5(a), most of the darker hue scenes, even with low counts for the number of audio elements in those scenes, score higher on the trailer-worthy scores. This perhaps tells us that the visual composition, particularly those with darker tones, are more trailer-worthy than when there are prominent audio segments present in a visually bright scene. That's possibly due to the emotional or psychological impact of the darker visuals engaging the audience better even with minimal audio cues. These findings suggest that, in the case of Beware of Zombies, changes in visual mood and/or tone, rather than interaction between visual brightness and key auditory elements, can be crucial for the efficiency of a film trailer.

On the other hand, from Fig. 5(b), it can be observed that Bleep is more filled with darker visual settings and has a rather very specific pattern in trailer-worthiness scores, tending to be much heavier in the audio component of most of the scenes that use more distinct horror-themed sound associations. Salient audio cues, like the use of dissonant tones, suspenseful soundscapes, and other audio cues typical for the genre, go a long way in making those scenes worth the trailer. Therefore, results would indicate that audio, especially those that are characteristic and specific to horror movies, go a long way in enhancing the perceived intensity and engagement of a film trailer where dark visuals have been dominant. Again, this shows how important audio is in creating atmosphere and emotion in scenes that are low-light dominated visually.

These findings emphasize the integral relationship between the visual and audio components in movie trailers, in which both balance and interaction may hold a great bearing on viewers' engagement. The findings suggest that in films featuring bright visual scenes, perhaps the visual tone-that is, darkness or brightness-would have more influence in making the trailer worth watching, while in movies that are basically dark visually, the audio elements themselves, especially those hinting at genre-specific tones, may be critical in bringing effectiveness to trailers.



(a)



(b)

Fig 3: Timeline of Ignore It (a) and Bleep (b), highlighting key audio moments, matched scenes, and trailer-worthy elements from the generated trailer

B. Qualitative Analysis of Results

The predicted scenes are compiled into a trailer - the generated trailer - and these segments are analysed for trailer worthiness.

Fig. 3 represents the timeline of the short films - Ignore It (a) and Bleep (b). In "Ignore It", a family avoids acknowledging a deadly spirit ("her") to survive, while in "Bleep", A couple's relationship is pushed to the brink as they investigate a strange noise that woke them in the night. (a) represents some of the scenes from the generated trailer that are 1. in the original trailer i.e, the ground truth, 2. a "key audio" moment in the short film (represented by text) and 3. not in trailer and not a key moment i.e, failed predictions. The matched scenes are presented with the timestamps at which they actually occur in the original trailer. While (a) emphasises on scene matching, (b) gives more importance to the trailer worthiness of the scenes. This timeline represents all the "Key audio moments" (represented by text), often characterised by repeated sounds, intensity, eeriness and "Key segments" (represented by images) which have important audio and video features that highlight the theme of the movie and other significant aspects of the movie. It is necessary to note that certain official trailers rely heavily on video effects to showcase various details like the directorial team, movie title, release date, hence some movies in the test set have fewer detected "in trailer" scenes but more key moments (audio and video).

VI. Conclusion And Future Work

Trailers play a vital role to give a glimpse into an upcoming movie and pique audience attention. However manually curating a trailer for movies is labor intensive and time consuming. Hence prompting to automating the trailer generation process presents a valuable opportunity for the

production team to work on other creative aspects of filmmaking.

The repository utilized consists of 311 short films and their corresponding trailers which mostly are of the horror genre. This aforementioned data is split into separate modalities and processed. As mentioned in section IV, our novel methodology proved to be computationally effective.

The results obtained as seen in section V shows that a significant proportion of the predicted segments are highly trailer worthy and are present in the ground truth. The evaluation metrics demonstrate the effectiveness of our novel method in predicting trailer worthy segments. The hamming score indicates that more than 60% of the predicted segments are highly trailer worthy. The Intersection over Union score suggests that more 30% of the predicted segments overlap with the ground truth and the Task Accuracy signifies highlights the model's capacity to predict segments accurately

The Hamming Score and Task Accuracy Score of 1 highlights the model's effectiveness in prediction of the most high trailer worthy moments that match the ground truth. The IoU scores could be relatively lower because there are a significant number of key trailer segments in this film, and not all of them may make it into the final trailer.

While the results are promising, the future research could explore the inclusion of more sophisticated audio and visual features that capture the nuances of a horror movie. To enhance the model's robustness, a wider variety of horror films across different subgenres can be included in the dataset.

References

[1] K. M. Johnston, E. Vollans, and F. L. Greene, "Watching the trailer: Researching the film trailer audience," Participations: Journal of Audience and Reception Studies, vol. 13, no. 2, pp. 56-85, 2016.

[2] V. G. Kuppelwieser and J. Finsterwalder, "The effects of film trailers on shaping consumer expectations in the entertainment industry—A qualitative analysis," Journal of Retailing and Consumer Services, vol. 19, no. 6, pp. 589-595, 2012.

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

[4] 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

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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

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

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

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

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

[19] https://www.scenedetect.com/

[20] https://github.com/openai/CLIP