International Journal on Advanced Science Engineering Information Technology

Vol.14 (2024) No. 1
ISSN: 2088-5334

Multi-Modal Deep Learning based Metadata Extensions for Video Clipping

Woo-Hyeon Kim [a], Geon-Woo Kim [a], Joo-Chang Kim [b,*]

[a] *Department of Computer Science, Kyonggi University, Suwon 16227, Republic of Korea*
[b] *Division of AI Computer Science and Engineering, Kyonggi University, Suwon 16227, Republic of Korea*
Corresponding author: [*]kjc2232@naver.com

***Abstract*— General video search and recommendation systems primarily rely on metadata and personal information. Metadata includes file names, keywords, tags, and genres, among others, and is used to describe the video's content. The video platform assesses the relevance of user search queries to the video metadata and presents search results in order of highest relevance. Recommendations are based on videos with metadata judged to be similar to the one the user is currently watching. Most platforms offer search and recommendation services by employing separate algorithms for metadata and personal information. Therefore, metadata plays a vital role in video search. Video service platforms develop various algorithms to provide users with more accurate search results and recommendations. Quantifying video similarity is essential to enhance the accuracy of search results and recommendations. Since content producers primarily provide basic metadata, it can be abused. Additionally, the resemblance between similar video segments may diminish depending on its duration. This paper proposes a metadata expansion model that utilizes object recognition and Speech-to-Text (STT) technology. The model selects key objects by analyzing the frequency of their appearance in the video, extracts audio separately, transcribes it into text, and extracts the script. Scripts are quantified by tokenizing them into words using text-mining techniques. By augmenting metadata with key objects and script tokens, various video content search and recommendation platforms are expected to deliver results closer to user search terms and recommend related content.**

***Keywords*— Video Analysis; recommendation; speech recognition; contextualized data; metadata.**

Manuscript received 15 Sep. 2023; revised 29 Dec. 2023; accepted 11 Jan. 2024. Date of publication 29 Feb. 2024.

I. INTRODUCTION

Currently, general video search and recommendation systems primarily rely on user cookies or weblogs. These methods utilize initial surveys or personal information recorded on the existing web to determine the preferences of new users. The multimedia industry originated with user-created content (UCC), and with the proliferation of personal Internet broadcasting and Over-the-top (OTT) platforms, a large volume of new video content is being generated. General video search and recommendation systems utilize metadata such as title, synopsis, keywords, tags, genre, and cast, which are added in advance by producers or distributors [1]–[3]. In cases where there is no index for the video beyond metadata, accessing video content can be time-consuming.

OTT platforms or video platforms with extensive content catalogs can analyze user search history, viewing history, subscription channels, etc., stored in users' web browsers to understand their preferences and provide customized video recommendations. This is achieved by comparing the user's history with metadata to recommend content judged as relevant or by comparing it with the records of other users with similar interests. Based on user analysis, recommended videos are selected and provided in a Top-N format according to priority. While most search and recommendation algorithms work effectively, they heavily rely on users' personal information, including viewing history, search history, and Internet browser cookies. Therefore, if a user deletes their information or sets their information protection to a high level, the information required for recommendations may be insufficient, resulting in inaccurate recommendations. This is commonly called the 'cold start problem'[4]–[6].

Since content creators or distributors primarily provide metadata, it may lack information about the actual content, potentially reducing the accuracy of searches and recommendations or being susceptible to misuse. Recommendation systems for similar content on OTT platforms often exclude significant content that could be

considered a spoiler, even if the plot is included in the recommendation process. Consequently, even if the content is recommended as similar due to similarities in title, genre, performers, etc., the perceived degree of similarity by the actual user may differ. Inaccurate searches or recommendations erode trust in the platform and lead to a decrease in the number of users. Furthermore, recommendation algorithms based on personal history can lead to the 'bubble effect,' wherein only content related to the user's history is presented, exposing users to biased content [7]–[10]. There have been social issues where crimes exploited this to expose children to harmful videos. As a result, concerns and interests regarding recommendation algorithms and video censorship are growing [11].

To address these problems, efforts to enhance censorship and the implementation of stricter regulations to block and remove harmful content are required, which consumes a significant amount of time and resources. Additionally, videos with long running times may contain various content, including parts that users may not desire to access [12].

In this study, we propose a Multi-Modal Deep Learning based Metadata Extensions Model for Video Clipping. Extract contextual data from real content through artificial intelligence-based content analysis. We propose a model that indexes extracted contextual data into videos and clips to extend metadata and improve search and recommendation systems.

The key contributions of our approach include:

1) Object Frequency Analysis: We employ the YoloV8m deep learning model to perform frame-by-frame object detection and extraction, enabling us to select key objects based on their frequency within the video content.

2) Textual Analysis: We utilize STT technology to transcribe audio content into textual scripts, providing a text-based representation of the audio component of the video.

3) Contextual Data Integration: We seamlessly integrate object frequency analysis and textual scripts on a frame-by-frame basis, creating rich contextual data that captures both visual and auditory elements.

4) Video Clipping: We introduce an innovative video clipping algorithm based on contextual data, optimizing the categorization and segmentation of video content.

II. Material and Method

A. Video Platforms

Advances on the Internet, communication technology, and smartphone technology have created an environment where high-resolution video can be transmitted in real-time. In particular, the spread of smartphones and hardware advancements have enabled anyone to create content. People are content creators and consumers, and a large amount of video content is produced and consumed. This is the background for the emergence of various services such as video streaming platforms, live streaming platforms, education and expertise platforms, social media platforms, and video sharing platforms. Video service platforms have revolutionized media consumption patterns and play an essential role in providing opportunities for new video content creators. As more and more video content is being created, the demand for advancements in search and recommendation technologies targeting video is also increasing [13].

B. Video Service Platforms

A representative example of a video streaming platform is OTT. It is media content such as video, audio, and text provided over the Internet. OTT services deliver content directly to consumers without going through traditional media distribution channels such as broadcast or cable TV. Critical features of OTT services include Internet-based delivery, diverse content, a subscription model, personalized recommendations, content diversity, and creator opportunities. Internet-based delivery means that OTT services deliver content to users via an Internet connection, allowing users to watch or listen to content anytime, anywhere on various devices, including smartphones, tablets, and smart TVs. Subscribers can use these services to watch ad-free content of their choice. Additionally, many OTT platforms analyze users' viewing habits to provide personalized content recommendations. This makes it easier for users to find content that matches their tastes and interests [14]–[16].

C. Video Scene Detection

Shot boundary detection is fundamental for identifying scene transitions within video content. Videos are composed of a sequential series of frames, and the process involves identifying alterations in aspects such as color, lighting, and objects within a scene. This automated technique partitions the video into elementary segments termed "shots." A shot comprises consecutive video frames typically originating from a single camera. Various shot boundaries exist, including cuts, dissolves, wipes, and fades out/in. Cuts, signifying an instantaneous transition from one frame to the next, are the most commonly encountered shot boundaries. Dissolves involve a gradual fading of one scene while the next scene gradually emerges. Wipes denote transitions between scenes, often characterized by a wiping pattern across or vertically across the screen. Fades out/in signify scenes slowly fading out or brightening before disappearing. Over the years, the evolution of shot boundary detection algorithms has progressed from rudimentary feature comparisons to the utilization of intricate and probabilistic models. Furthermore, orthogonal polynomials extract features from the orthogonal transition region to enhance transition detection accuracy, facilitating the identification of complex transitions within video sequences [17]–[20].

D. Video Summarization

Video summarization is briefly extracting the main or lengthy content from a video. Significant portions are removed in the case of videos, and the core content is presented in a concise text format, providing users with a brief overview of the entire video. This involves detecting important video scenes, key moments, and events using various technologies, including object detection, face recognition, speech processing, and captioning. A summary sentence or paragraph is generated and made available to the user based on the identified content. Additionally, it may be presented in audio or video format. Video summarization finds applications in various fields, including information

retrieval, news reporting, educational content, online advertising, and movie trailers, allowing users to save time and quickly access the desired information. Recent research has been actively exploring ways to enhance the quality and accuracy of video summaries through artificial intelligence technology [21]–[23].

E. Speech recognition

Speech recognition is an innovative technology designed to identify and transcribe human speech into text, facilitating interaction between users and computers through voice commands or input. Commonly referred to as speech-to-text (STT), its core components encompass speech input, digital signal processing (DSP), feature extraction, acoustic modeling, and language modeling. The process begins when a user inputs information using a microphone, where the voice signal converts from analog to digital format. Subsequently digital signal processing is subsequently applied, encompassing preprocessing tasks such as noise elimination and echo cancellation. The crucial step involves extracting significant features from the digitized speech, typically achieved through algorithms like Mel Frequency Cepstral Coefficient (MFCC) [24]–[26]. These extracted features are then utilized with pre-trained models, such as Hidden Markov Models (HMM) or Deep Neural Networks (DNN), for acoustic modeling, enabling word recognition and pronunciation analysis.

Additionally, language modeling is performed by selecting the most contextually appropriate word among several candidates, ensuring natural sentence progression. STT finds applications in various domains, including smartphone personal assistant services like Siri, Google Assistant, and Bixby, automatic translation services, assistive devices, customer service centers, and educational fields. With advancements in artificial intelligence and machine learning, STT is continually evolving. Techniques such as Long Short-Term Memory (LSTM) and Transformers have significantly enhanced STT performance. Nonetheless, achieving 100% accuracy remains challenging due to unclear pronunciation, diverse accents, dialects, and background noise [27].

F. Method

We introduce a novel approach titled "Multi-Modal Deep Learning based Metadata Extensions for Video Clipping" aimed at significantly enhancing the functionality of search and recommendation systems within the vast landscape of video content. Our methodology revolves around a meticulous analysis of videos, dissecting them frame by frame. While each frame encapsulates visual information in images, the audio component spans multiple frames, giving rise to a distinct data format. This inherent duality serves as the cornerstone of our multi-modal video analysis approach.

The Contextualization-based Metadata Extension Tool is central to our method, which generates two key components: Script and Object Detection results. These components are seamlessly integrated frame-by-frame, effectively creating rich contextual data that encapsulates visual and auditory information. Leveraging this contextual data, we employ advanced object frequency analysis and text mining techniques to extract meaningful content segments, facilitating precise video clipping.

The result of our efforts is an enriched metadata structure that extends beyond the traditional bounds of basic video descriptors. This expanded metadata encompasses a broader spectrum of information closely related to the video content. We envision this enriched metadata as a valuable resource, one that holds the potential to empower search engines and recommendation systems by providing them with a more comprehensive and nuanced understanding of video content. Ultimately, our approach is poised to revolutionize the performance and capabilities of video service platforms, offering viewers an enhanced and tailored experience in the ever-expanding realm of video content. Fig. 1 illustrates the metadata extensions for video clipping architecture.

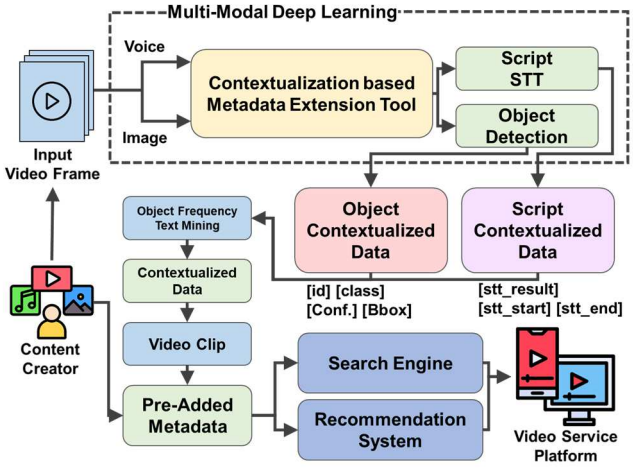

Fig. 1 Metadata Extensions for Video Clipping Architecture

III. Results and Discussion

A. Data Processing

This section outlines our proposed methodology to enhance the video recommendation system by selecting key objects based on their frequency and integrating textual analysis. The process involves three key steps:

Firstly, we identify essential objects required for the video recommendation system by conducting object frequency analysis. We employ the YoloV8m [28] deep learning model, capable of frame-by-frame object detection and extraction to achieve this. Trained on a labeled image dataset and initially on the Coco dataset [29], this model can proficiently detect and classify 91 objects, ranging from 1 to 91.

Simultaneously, we extract audio content separately from the video and employ Google Cloud Speech [30] to transcribe it into textual scripts. These transcriptions are then synchronized with the corresponding video frames.

The text mining process involves several critical stages, such as tokenization to segment text, lemmatization to eliminate irrelevant words, root extraction to determine word basics structure, and meaning analysis. This comprehensive methodology integrates object-based video analysis with advanced textual quantification, providing a multifaceted approach to enhance the performance of video recommendation systems.

B. Extending Metadata with Contextual Data

In multi-modal deep learning, data exhibiting the characteristics of time series must be used as inputs at the

same time point at which they occur and input in the same order as the temporal flow. When analyzing images, analyzing every frame requires significant computing resources, and with audio, skipping frames can lead to reduced accuracy. Therefore, audio is often separated from the visual components in video processing and fed into different models. The results from each model are then integrated as contextual data, preventing the data's temporal characteristics from being distorted or lost.

Incorporating this approach makes the metadata extension with contextual data more effective. It maintains the temporal relationship between audio and visual elements, which is crucial for tasks such as video summarization, content retrieval, and automated monitoring systems. This method optimizes the use of computational resources by strategically selecting frames for image analysis and maintaining the integrity of audio information throughout the process. Integrating the outputs of each modality preserves the continuity and context of the data, thereby enhancing the overall quality and usability of the metadata.

C. *Data Architecture*

Metadata Extensions proposed in this study analyze videos using object recognition using the Yolo deep learning model and STT using the Google Cloud Speech model. The analysis results are stored in Json format for each frame. Data generated from the Yolo deep learning model uses ID, Class, Confidence, and BBox. In the case of STT, since it appears over several frames, [stt_start] tokens and [stt_end] tokens are created to indicate the first and last frames in which the voice is recognized. Fig. 2 illustrates Contextual Data Architecture.

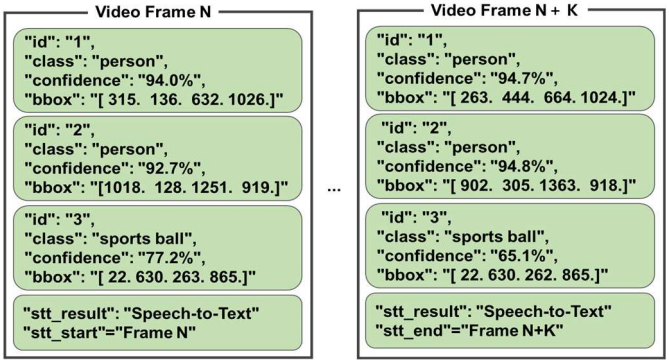

Fig. 2 Contextual Data Architecture

D. *Video Clipping*

Video content exhibits time series characteristics that evolve based on events and scene transitions. Traditional video clipping methods often rely on cumulative calculations of motion changes between successive frames or employ techniques like video dissolves and fades, which can demand significant computational resources. In contrast, our proposed approach introduces segmentation points using object tracking and script information, thus conserving computational resources by categorizing videos within metadata instead of physically trimming them.

The contextual data-driven video clipping algorithm, presented as Algorithm 1, is outlined as follows:

Algorithm 1 Video Clipping Algorithm using Object Recognition, Object Tracking, and Script Information

```
Read the video.
Initialize the list of clips.
Initialize the current clip.
while There is a next frame in the video do
  Read the current frame.
  if An object is recognized or the [stt_start] token is detected then
    if The current clip is not empty then
      Add the current clip to the list of clips.
      Initialize the current clip.
    end if
    Set the start frame of the current clip to the current frame.
  else if All recognized objects and [stt_end] have disappeared then
    if The current clip is not empty then
      Add the current clip to the list of clips.
      Initialize the current clip.
    end if
  end if
end while
```

Algorithm 1 Video Clipping Using Contextual Data

1) Object Recognition: This step involves recognizing objects within the video frames.

2) Clip Initialization: The video clipping process is initialized.

3) Clip Start Point: Identification of the start of a video clip occurs when an object is recognized or when the [stt_start] token is detected.

4) Clip End Point: Determination of the end of a video clip is made when both the recognized object and the [stt_end] token disappear.

5) Clip Recording: Each completed video clip's start and end frames are recorded.

6) Repeat: The process is repeated to create subsequent video clips.

This innovative approach allows for efficient video categorization and segmentation based on contextual information, leading to optimized resource utilization compared to traditional video clipping methods.

E. *Fine-tuning*

The proposed model facilitates fine-tuning by utilizing contextual data derived from videos. Initially, the extracted scripts obtained via STT are employed. To ensure the accuracy of these extracted scripts, they are cross-referenced with the embedded subtitle data within the video and subsequently calibrated.

Hyperparameters used to train the model are adjusted to enhance its performance through cross-validation. During cross-validation, a comparison is made between the extracted script and the contextual data comprising object frequency, which is acquired using the Yolo deep learning model.

Subsequently, an evaluation and testing phase is conducted on the developed model. This evaluation involves testing the model's accuracy by inputting select videos into the testing environment and verifying the accuracy of the extracted metadata.

F. *Contextualization-based Metadata Extension Tool*

Fig. 3 is a screenshot of the graphical user interface (GUI) for the "Contextualization-Based Metadata Extension Tool." This software seems to be designed for video analysis, where

it can identify and track objects within the video footage. There are two separate panels displayed in the image, each representing a different scene from the video being analyzed.

Fig. 3 Contextualization-based Metadata Extension Tool

The upper panel shows a news broadcast scene, where the faces of individuals have been pixelated for privacy. The software has drawn bounding boxes around each person and object of interest, with labels such as "Person", indicating the tool's ability to detect and categorize objects within the video frame. Below the video, there is a section labeled "STT Result", which likely stands for Speech-to-Text, showing the transcription of the audio from the video.

The lower panel shows a highway scene with vehicles on the road. Similarly to the upper panel, bounding boxes with labels have been applied to various vehicles, suggesting that the tool can also identify and track cars, trucks, and possibly their make and model. On the left-hand side of each panel, there are input parameters that the user can adjust:

1) "Confidence" can be set between 0 and 1. This parameter determines the threshold the software uses to decide whether the identification of an object is accurate enough to be considered correct. A higher confidence level means the software will only recognize detections it is more specific about, whereas a lower confidence level means it will accept more detections with less certainty

2) "IoU", or Intersection over Union, ranges from 0 to 1. It measures the overlap between the predicted and ground truth bounding boxes. A higher IoU represents a higher accuracy of object detection, as it indicates a more significant overlap between the predicted and actual positions of the objects.

3) "Frame Interval" is adjustable from 0 to 100. This setting controls the frequency of the video analysis by specifying which frames are analyzed. A lower frame interval means that more frames within the video will be investigated, which could lead to more detailed data but at the cost of requiring more computational power. A higher frame interval means fewer frames are analyzed, which conserves computational resources but may potentially miss some temporal details within the video content.

The "Frame Interval" parameter is applied to the YOLO (You Only Look Once) algorithm for object detection within the video frames. It is adjustable to determine the frequency at which the video is analyzed for objects. However, this parameter is not applied to Speech-to-Text (STT) processes, as audio data's sequential and temporal characteristics are essential. Therefore, for STT, every frame is analyzed to ensure that the temporal integrity of the audio data is preserved and to maintain the accuracy of the speech recognition.

G. Discussion

Utilizing the proposed method leads to extracting and transforming inherent information embedded within video content, effectively converting it into valuable contextual data. This innovative approach holds significant promise in enhancing the efficient management of the ever-expanding volume of video content that is continuously generated. It proves exceptionally advantageous when assessing the similarities among videos hosted on a video streaming platform. This newfound capability empowers video search engines and recommendation systems by providing them with a more comprehensive dataset for analysis.

Moreover, this method's utility extends to videos developed over extended durations, especially those featured on live-streaming platforms. Content creators frequently engage in the meticulous editing and processing of their video content. They often share curated summaries or highlights through various video-sharing platforms. In this context, the methodology introduced through this process offers a faster and more thorough means of managing and identifying editing points. This leads to enhanced video content curation, ensuring no crucial elements are overlooked or omitted.

The proposed approach thus promises to revolutionize how video content is analyzed, curated, and presented across a wide range of platforms, offering a more comprehensive and efficient solution for content creators and viewers. In future research, we plan to implement a search system using the proposed method and conduct research on using Confusion Matrix evaluation and video clipping for search and recommendation using contextual data.

IV. CONCLUSION

In this paper, we proposed a multi-modal deep learning-based metadata extension for video clipping. The reliance on user cookies and weblogs in general video search and recommendation systems is challenged by issues such as the 'cold start problem,' overreliance on personal information, and the potential for biased content recommendations. To tackle these challenges, our proposed approach leverages advanced technologies, including deep learning, object recognition, STT, and text mining. We seamlessly integrate these technologies to extract valuable contextual data from video content. This contextual data encapsulates visual and auditory

information, offering a multi-modal perspective for content analysis.

We employ the YoloV8m deep learning model to perform object detection and extraction frame by frame, allowing us to select crucial objects based on their frequency within the video content. Additionally, we utilize STT to transcribe audio content into textual scripts, providing a text-based representation of the video's audio component. These two elements, object frequency analysis and textual scripts, are seamlessly integrated on a frame-by-frame basis, creating rich contextual data that captures both visual and auditory elements. This contextual data forms the foundation for our innovative video clipping algorithm, optimizing the categorization and segmentation of video content. Furthermore, our model supports fine-tuning through cross-validation, ensuring script accuracy and optimizing model hyperparameters. By expanding the metadata associated with video content, our approach significantly enhances video search and recommendation systems. The enriched metadata provides a comprehensive dataset for analysis, leading to improved content recommendations.

Moreover, our method streamlines video content management, particularly for lengthy videos, live streams, and edited content, offering a more efficient solution for content creators and viewers alike. Overall, our proposed approach has the potential to revolutionize the way video content is managed, curated, and presented across various platforms. It empowers content creators and viewers with a more efficient, nuanced, and personalized video experience, ultimately advancing the capabilities of video service platforms in the rapidly evolving multimedia landscape.

ACKNOWLEDGMENT

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No. 2020R1A6A1A03040583). Additionally, this work was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(No: RS-2023-00248899).

REFERENCES

[1] Kim J. C. and Chung K. Y., "Knowledge expansion of metadata using script mining analysis in multimedia recommendation", *Multimedia Tools and Applications*, vol. 80, pp. 34679-34695, 2020.

[2] Lee J. H., "The Growth and Impact of OTT on Video Viewing Behavior", *Asian-pacific Journal of Convergent Research Interchange*, vol. 6, no. 1, pp. 41-50, 2020.

[3] Shah S. and Mehta N., "Over-the-top (OTT) streaming services: studying users' behaviour through the UTAUT model", *Management and Labour Studies*, vol. 48, no. 4, pp. 531-547, 2023.

[4] Luo, M., Chen, F., Cheng, P., Dong, Z., He, X., Feng, J. and Li, Z., "Metaselector: Meta-learning for recommendation with user-level adaptive model selection", *In Proceedings of The Web Conference 2020*, pp. 2507-2513, 2020.

[5] N. Silva, T. Silva, H. Werneck, L. Rocha and A. Pereira, "User cold-start problem in multi-armed bandits: When the first recommendations guide the user's experience", *ACM Transactions on Recommender Systems*, vol. 1, no. 1, pp. 1-24, 2023.

[6] Hao, B., Yin, H., Zhang, J., Li, C. and Chen, H., "A Multi-strategy-based Pre-training Method for Cold-start Recommendation", *ACM Transactions on Information Systems*, vol. 41, no. 2, pp. 1-24, 2023.

[7] A. Ishikawa, E. Bollis and S. Avila, "Combating the elsagate phenomenon: Deep learning architectures for disturbing cartoons," *in Proc.IWBF'19*, pp. 1-6. 2019.

[8] Matakupan, N. I., "The Study of'Don't Hug Me I'm Scared'Web Series Storytelling For IP Design Regarding Safe Viewing Content For Children", *IJVCDC (Indonesian Journal of Visual Culture, Design, and Cinema)*, vol. 2, no. 2, pp. 172-177, 2023.

[9] Yousaf, K. and Nawaz, T., "An attention mechanism-based CNN-BiLSTM classification model for detection of inappropriate content in cartoon videos", *Multimedia Tools and Applications*, pp. 1-24, 2023.

[10] Huang, S., Liu, G., Chen, Y., Zhou, H. and Wang, Y., "Video Recommendation Method Based on Deep Learning of Group Evaluation Behavior Sequences", *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 37, no. 02, pp. 2352002, 2023.

[11] A. Yousaf., A. Mishra., B. Taheri and M. Kesgin, "A cross-country analysis of the determinants of customer recommendation intentions for over-the-top (OTT) platforms," *Information & Management*, vol. 58, no. 8, pp. 103543, 2021.

[12] Hashemi, M., "Web page classification: a survey of perspectives, gaps, and future directions", *Multimedia Tools and Applications*, vol. 79, no. 17-18, pp. 11921-11945, 2020.

[13] Hesmondhalgh, D. and Lotz, A., "Video screen interfaces as new sites of media circulation power", *International Journal of Communication,* vol. 14, pp. 386-409, 2020.

[14] Patnaik, R., Shah, R. and More, U., "Rise of OTT platforms: effect of the C-19 pandemic", *PalArch's Journal of Archaeology of Egypt/Egyptology*, vol. 18, no. 7, pp. 2277-2287, 2021.

[15] Singh, N., Arora, S. and Kapur, B., "Trends in over the top (OTT) research: a bibliometric analysis", *VINE Journal of Information and Knowledge Management Systems*, vol. 52, no. 3, 411-425, 2022.

[16] Sontakke, K. S., "Trends in OTT Platforms Usage During COVID-19 Lockdown in India", *Journal of Scientific Research*, vol. 65. no. 8, pp. 23, 2021.

[17] Sun, C., Jia, Y., Hu, Y. and Wu, Y., "Scene-aware context reasoning for unsupervised abnormal event detection in videos", *In Proceedings of the 28th ACM International Conference on Multimedia*, pp. 184-192, 2020.

[18] Ramachandra, B., Jones, M. J. and Vatsavai, R. R., "A survey of single-scene video anomaly detection", *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no.5, pp. 2293-2312, 2020.

[19] Raja, R., Sharma, P. C., Mahmood, M. R. and Saini, D. K., "Analysis of anomaly detection in surveillance video: recent trends and future vision", *Multimedia Tools and Applications*, vol. 82, no. 8, pp. 12635-12651, 2023.

[20] Luo, H., Ji, L., Zhong, M., Chen, Y., Lei, W., Duan, N. And Li, T., "Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning", *Neurocomputing*, vol. 508, pp. 293-304, 2022.

[21] Haq, H. B. U., Asif, M. and Ahmad, M. B., "Video summarization techniques: a review", *Int. J. Sci. Technol. Res*, vol. 9, no. 11, pp. 146-153, 2020.

[22] Workie, A., Sharma, R. and Chung, Y. K., "Digital video summarization techniques: A survey", *Int. J. Eng. Technol*, vol. 9. no. 1, pp. 81-85, 2020.

[23] Tiwari, V. and Bhatnagar, C., "A survey of recent work on video summarization: approaches and techniques", *Multimedia Tools and Applications*, vol. 80, no. 18, pp. 27187-27221, 2021.

[24] Malik, M., Malik, M. K., Mehmood, K. and Makhdoom, I., "Automatic speech recognition: a survey", *Multimedia Tools and Applications*, vol. 80, pp. 9411-9457, 2021.

[25] Li, J., "Recent advances in end-to-end automatic speech recognition", *APSIPA Transactions on Signal and Information Processing*, vol. 11, no. 1, 2022.

[26] Guo, Z., Leng, Y., Wu, Y., Zhao, S. and Tan, X, "PromptTTS: Controllable text-to-speech with text descriptions," *In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1-5. IEEE, 2023.

[27] Saranya, V., Devi, T. and Deepa, N, "Text Normalization by Bi-LSTM Model with Enhanced Features to Improve Tribal English Knowledge", *In 2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS),* pp. 1674-1679, 2023.

[28] Ultralytics YOLOv8, 2024. 01. 01. https://docs.ultralytics.com/.

[29] CoCo Dataset, 2024. 01. 01. https://cocodataset.org/.

[30] Google Cloud Speech, 2024. 01. 01. https://cloud.google.com/.