International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 04 | April - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

Query Based Video Segment Retrieval Using Knowledge Graphs

Imaan Junaid[1], Abdul Raheem Khan[2], Sai Kiran[3], Mr. Shaik Kareem Basha[4]

Student, Department of Artificial Intelligence & Data Science, Methodist College of Engineering & Technology, Hyderabad, Telangana, India [1 2 3]

Assistant Professor, Department of Computer Science Engineering, Methodist College of Engineering & Technology, Hyderabad, Telangana, India

Abstract– The growth in the amount of videos causes problems in finding relevant data in videos by text search. The current state-of-the-art approach uses vector searches based on semantics and demonstrates poor multimodal data understanding capabilities. The presented paper describes the method for retrieving segments of videos that utilizes semantic vectors and knowledge graphs to make accurate video search. The proposed method includes multimodal analysis like speech recognition, automatic recognition of characters in the text, and then video contextual segmentation. Knowledge relations between objects are found by relation extraction models, which allow storing structured data in a graph representation. Relations in graphs can be used to query data and retrieve semantically related information. Efficient ranking with a combination of graph and vector scores is developed. The experiments demonstrate the effectiveness of the proposed technique in terms of mean reciprocal rank (MRR) and Recall@K metrics.

Keywords: Video Retrieval, Knowledge Graphs, Semantic Search, Multimodal Learning, Natural Language Processing, Hybrid Retrieval

1. INTRODUCTION

This rapid growth causes a challenge of looking up the relevant data within the video through textual queries. However, the methods used now are dominated by vector-based searches concerning semantics and provide poor results regarding understanding of multimodal data. This paper describes an approach of retrieving the video data relying on semantical vectors and knowledge graphs. Specifically, there is a multimodal analysis performed, which involves speech recognition, and the optical character recognition with subsequent video segmentation depending on the context. Relation extraction model allows to determine relations between the nodes and represents the data in the form of a graph. Relations between the graph nodes help to look up the data based on the same semantical information. A unified ranking approach is used.

2. LITERATURE SURVEY

Sun et al. (2024) proposed CLIP2TF, a multimodal approach for video-text retrieval was proposed that utilized CLIP architecture, which is efficient at capturing audiovisual information from multimedia content for educational purposes. But this system cannot capture the contextual relationship between the multimodal information presented in long form videos. This problem is overcome by our approach through the utilization of knowledge graph reasoning techniques [1].

Feng et al. (2023) Proposed MKVSE, which was a multimodal knowledge enhanced visual semantic embedding framework for retrieving images and texts. The technique enhances the alignment of semantics through the use of knowledge graphs. However, despite this advancement, it only addresses image and textual retrieval issues while ignoring the aspect of time in videos. Our research aims to bridge that gap by applying this concept to videos [2].

Kou et al. (2023) Created KnowER, which uses knowledge graphs and CLIP encoder for better text-video retrieval. The model increases semantic understanding using external knowledge but cannot effectively deal with noisy data and sparse representation. Our model increases efficiency by integrating the benefits of graph reasoning and semantic similarity scoring [3].

Li et al. (2024) proposed an approach to use a multi-modal hypergraph network for retrieving videos based on text. Although it captures the higher order relations between frames and text, it is complex in computation. To overcome this problem, our approach incorporates the concept of graph and embedding based hybrid retrieval [4].

Che and Guo (2024) A dual alignment method was used for cross-modal video retrieval to bridge the gap between text and video semantics. While this approach is effective in aligning different modalities, it could fail to capture any complicated relationship. The proposed approach enhances retrieval by capturing semantic alignment and relationship modeling explicitly using knowledge graphs [5].

Guo et al. (2024) Hypergraphs are also employed in extended hypergraph-based retrieval for multimodal video text tasks. This, however, has been found to have limitations with regard to scalability and is highly reliant on data quality. Our proposed system has managed to overcome this limitation by using vector retrieval combined with graph-based reasoning [6].

Feng et al. (2023) proposed a heterogeneous graph-based retrieval system named ORANGE was proposed. Though successful, this model depends on the user interaction information, which is not always available. In our proposed system, we have avoided such dependency by developing a content-based retrieval system [7].

Zhao et al. (2023) proposed an uncertainty-based retrieval framework for adaptive matching using probabilistic models. While it is more adaptable, it adds complexity to the model and increases training costs. Our system retains efficiency and achieves similar semantic comprehension using hybrid scoring [8].

Zhang et al. (2025) proposed a multi-stage retrieval scheme utilizing multi-modal tagging and pre-screening to optimize performance. Although this results in savings in terms of computation, it can result in a loss of pertinent information because of pre-filtering. However, our approach bypasses this problem entirely by incorporating the use of semantic and graph-based searches in one system [9].

Jeong et al. (2025) proposed an audio-guided video representation model based on gated attention architecture for better performance in text-based video retrieval tasks. Although the proposed model uses audio information efficiently, it adds to the complexity of the model and its sensitivity to the audio signal quality [10].

3. PROPOSED WORK

In this paper, we have developed a new framework for extracting segments from long videos based on a hybrid approach. While previous techniques either concentrate on similarities or keywords, the current method leverages the benefits of both techniques to deliver reliable context-aware video segment retrieval.

The first step involves pre-processing of the video data using a multimodal extraction technique. In the first case, the audio data of the video is transcribed with ASR while the visual information is processed with OCR to identify the text shown on the screen. Extraction of text data from both the audio and visual segments delivers enriched text data which considers both visual and auditory contexts.

For context-aware text segmentation, the first step involves segmentation of text into meaningful chunks based on similarity between different segments of text. Each text segment is then mapped to an embedding space using the Transformer architecture before applying similarity thresholding with respect to duration.

Once the segments are generated, structured knowledge extraction takes place based on named entity recognition and relation extraction. The named entities are extracted along with their relationships to form a graph where each node represents an entity while edges represent semantic relations. Moreover, dense vectors are generated for each text segment.

During the preprocessing phase, the query issued by the user is converted into embeddings. The query processing phase includes parallel processing, where vector-based retrieval with the cosine similarity score and graph-based retrieval with entity and relation matching take place. The results from both methods are combined according to the following equation using a weight-based score generation technique:

$$Score=\lambda1\cdot GraphScore+\lambda2\cdot SimilarityScoreSc$$

With this approach, the ranking system will be able to balance between semantic and relational search techniques according to the query type.

The output consists of video segments that are sorted and have timestamp information to enable the users to access specific parts of the video.

4. SYSTEM ARCHITECTURE

The suggested architecture is modular and hierarchical in its design and seeks to facilitate the process of extracting video segments from long-form videos through natural language queries.

It contains several interrelated modules including content extraction module, knowledge graph construction module, and query module among others. Through analysis of the input videos, the system structures the content and based on the users' queries, extracts the appropriate video segments through the hybrid retrieval model.

4.1 Content Extraction Module

In other words, the module will facilitate the process of video data processing and conversion into meaningful text data. The audio part of the video is separated first; then, it is automatically transcribed using the speech recognition model.

Moreover, the frames of the video are captured regularly; after that, they are processed using OCR technology to find any text that could be present in the video.

Text data retrieved from the audio and image are processed further to make it semantically meaningful using NLP technologies.

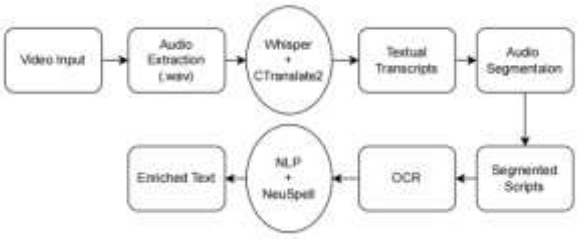

Fig 4.1: Contextual Extraction from Video Data to Enriched Text

4.2 Knowledge Graph Construction Module

In this module, the text is converted to structured knowledge graphs.

Tokenization, normalization, and noise removal processes are done in the preprocessing phase.

The NER algorithm is applied for entity detection, whereas relationship extraction methods are used to create triplets containing subject-relationship-object pairs.

Those triplet structures are saved to a graph database and finally a knowledge graph is formed where entities become nodes and relationships become edges.

Moreover, the text encoding process uses transformer based encoder to create dense vectors.

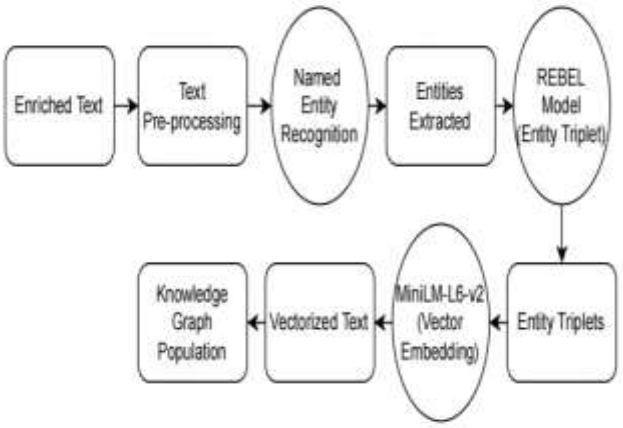

Figure 4.2: Text Processing to Knowledge Graph

4.3 Query Processing and Retrieval Module

In this module, interaction will happen between the user and the relevant video clip extraction process. In this module, the user will be able to formulate his queries in natural language, which is then preprocessed to get vector embeddings.

Two distinct retrieval engines will work simultaneously – semantic retrieval engine through cosine similarity calculation and graph-based retrieval engine through entity and relation detection.

The results obtained from these two engines will then be combined using a hybrid scoring mechanism that will decide the relevance of video clips using a relevance measure.

The result of both these operations will contain metadata information such as timestamps and video IDs to allow users to navigate through the relevant parts of the video.

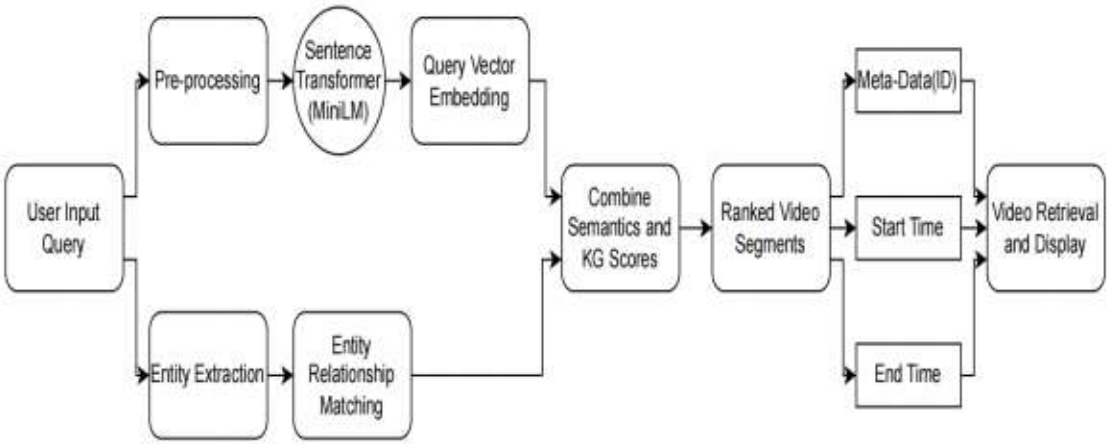

Fig 4.3: Query Processing to Video Retrieval

5. RESULTS

The evaluation process of the suggested system was done in order to study the efficiency of the retrieval of video clips based on natural language queries.

It can be seen from the result that the proposed system accepts the user's query as input and returns the video clips that are sorted and along with the time stamp context of the information in text format.

Fig 5.1: Application Interface

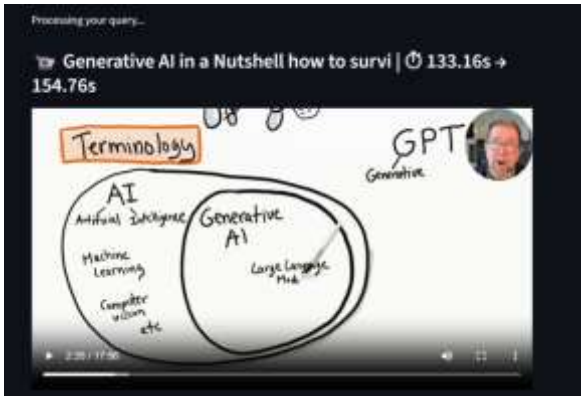

Figure 5.2: Query Output

5.1 Evaluation Metrics

The following two metrics have been employed in the assessment process of the proposed algorithm:

Firstly, we should mention Similarity Weighted Mean Reciprocal Rank (sMRR). The metric assesses the quality of ranking similar objects in the list. This metric takes into account both semantic similarity and ranking position and returns the result within [0; 1] range. The higher the number is, the better the result is

Secondly, there is a metric called Recall@K, which is the proportion of similar objects among the first K objects on the list. The metric uses a threshold value, based on which it determines whether the object is relevant or not.

5.2 Experimental Results

Several situations have been analyzed in which testing was performed on the system through a hybrid approach of semantic search and knowledge graph queries depending upon various weights.

In case of knowledge graph queries alone ($\lambda1=1.0$; $\lambda2=0.0$), MRR is 0.52, whereas in case of semantic search queries alone ($\lambda1=0.0$; $\lambda2=1.0$), MRR is 0.80.

But in a hybrid approach, it has been observed that performance optimization can be done by providing higher importance to semantic search. It can be shown by the following situations:

- $\lambda2=0.8$; $\lambda1=0.2 \rightarrow$ MRR = 0.84
- $\lambda2=0.7$; $\lambda1=0.3 \rightarrow$ MRR = 0.78

5.3 Recall Analysis

In addition, the performance of the proposed system was evaluated using the Recall@K measures, where the value of $\alpha=0.8$ and $\tau=0.8$.

- When K=1, the Recall@K value = 0
- When K=3, the Recall@K value = 0.34
- When K=5, the Recall@K value = 0.80

From the results, it is clear that the higher the value of K, the better the recall. However, an increase in the value of K leads to a decrease in the MRR measure.

5.4 Result Analysis

The effectiveness of the suggested approach can clearly be observed through the experiment results.

In other words, the suggested model can effectively integrate semantic understanding and structural reasoning to improve accuracy.

However, one should bear in mind that the efficiency of this approach is contingent upon factors such as query design and segmentation accuracy. Overall, the suggested system can work efficiently in extracting video segments in relation to their context.

6. DISCUSSIONS

From the experiment results, it can be concluded that the hybrid video search system is quite effective in the process of extracting long form videos using natural language queries. Thanks to the combination of semantic embedding along with knowledge graph reasoning, we were able to enhance the efficiency of our system.

First of all, the analysis proved that the semantic method is preferable in case of searching accurate information, which is evident from the MRR improvement by increasing the importance of semantic search. Secondly, the application of graph retrieval was more efficient when dealing with entity and relation queries, as they increased precision of results.

Hence, the proposed hybrid approach is able to eliminate the drawbacks that were identified in the individual approaches, while combining their benefits. Therefore, semantic gap related to vector retrieval, as

well as flexibility of graphs, are successfully overcome by the proposed approach.

Notwithstanding all the above, some problems arose during the process of evaluating the proposed system. First of all, its effectiveness is strongly connected to the quality of the segmenting procedure, as well as to how users construct queries. At times, required information was found by the system, however, it was not located first.

The next thing that can be worked on concerns the fact that at present the system employs mostly the textual information extracted from the analysis of the audio-visual information. In the future, more attention can be paid to the use of advanced visual perception and processing.

Summing up, the presented points demonstrate that the system is highly efficient and scalable; however, some improvements still can be made in terms of ranking, multimodal perception, and query analysis.

7. CONCLUSION

This paper explores the application of hybrid techniques in the process of video segment retrieval from long video sequences through natural language search queries. In this paper, this approach has been demonstrated to be highly efficient in terms of video segment retrieval through the exploitation of the strengths of semantic embeddings and knowledge graph reasoning.

In this system, multimodal content analysis is performed using speech and vision analysis to generate textual embeddings of video sequences. These texts are then used to generate a knowledge graph as well as perform semantic embeddings for similarity and relation-based searches.

It can be observed from the experimental results that the proposed hybrid model outperforms independent methods for retrieval, and its performance increases even further when semantics search is used along with reasoning based on the knowledge graph.

However, there are certain limitations, such as depending on the text-based encoding and sensitivity to query construction. There are various ways in which the future work could improve on the current state-of-the-art approaches, including better visual reasoning and ranking.

To conclude, the proposed architecture provides an innovative approach to video retrieval.

8. REFERENCES

[1] X. Sun, T. Fan, H. Li, G. Wang, P. Ge, and X. Shang, "CLIP2TF: Multimodal video–text retrieval for adolescent education," *Displays*, vol. 84, p. 102801, 2024.

[2] D. Feng, X. He, and Y. Peng, "MKVSE: Multimodal knowledge enhanced visual-semantic embedding for image-text retrieval," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 19, no. 5, pp. 1–21, 2023.

[3] H. Kou, Y. Yang, and Y. Hua, "KnowER: Knowledge enhancement for efficient text-video retrieval," *Intelligent and Converged Networks*, vol. 4, no. 2, pp. 93–105, 2023.

[4] Q. Li *et al.*, "Text-video retrieval via multi-modal hypergraph networks," in *Proc. 17th ACM Int. Conf. Web Search and Data Mining (WSDM)*, 2024.

[5] Z. Che and H. Guo, "Cross-modal video retrieval model based on video-text dual alignment," *International Journal of Advanced Computer Science & Applications*, vol. 15, no. 2, 2024.

[6] Z. Guo *et al.*, "Text-video retrieval via multi-modal hypergraph networks," in *ACM Web Search and Data Mining (WSDM)*, 2024.

[7] J. Feng *et al.*, "ORANGE: Text-video retrieval via watch-time-aware heterogeneous graph contrastive learning," in *Proc. EMNLP Industry Track*, 2023.

[8] Y. Zhao *et al.*, "Uncertainty-adaptive text-video retrieval," *arXiv preprint arXiv:2301.06309*, 2023.

[9] L. Zhang *et al.*, "Efficient text-to-video retrieval via multi-modal multi-tagger pre-screening," *Visual Intelligence*, vol. 2, 2025.

[10] D. Jeong *et al.*, "Learning audio-guided video representation with gated attention for video-text retrieval," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2025.