An Interactive System for Multimedia Retrieval in Video Collection with Temporal Integration

Gia-Kiet Pham*
Hai-Binh Tran*
Phi-Long Nguyen
phamgiakietk14@siu.edu.vn
tranlehaibinhk12@siu.edu.vn
nguyenhuynhphilongk14@siu.edu.vn
The Saigon International University
Ho Chi Minh City, Viet Nam

Song-Phuong Le
Hoang-Long Pham
Xuan-Tri Pham
letransongphuongk14@siu.edu.vn
phamhoanglongk15@siu.edu.vn
phamxuantri@siu.edu.vn
The Saigon International University
Ho Chi Minh City, Viet Nam

Ham-Duong Tran
Ngoc-Tin Huynh
Kiem Hoang
tranhamduong@siu.edu.vn
huynhngoctin@siu.edu.vn
hoangkiem@siu.edu.vn
The Saigon International University
Ho Chi Minh City, Viet Nam

ABSTRACT

Multimedia retrieval in computer science is the process of obtaining text, images, videos, and audio segments, all in digital form relevant to an information need from a collection of these resources. With the ever-growing amount of data, scalable and interactive retrieval systems that can efficiently work on extensive data collections while maintaining high precision are in high demand by industries and researchers. This paper presents the Pumpkin system, an interactive multimedia retrieval system first used in The AI Challenge Ho Chi Minh City 2023, an annual video event and moment retrieval competition. The system is built and set in motion to handle the retrieval task in a video collection of considerable size and complexity by three primary methods: visual-text association search, object-based search, and audio speech instances search. Additionally, the system has an integrated temporal workflow to search for conceptually related shots in a sequential motion, which removes out-of-context while leveraging suitable results as the user inputs more details to the system. Our system also puts great emphasis on user experience by cooperating with a clean and intuitive interface design with simplified user-side functionality, allowing a more efficient process of information retrieval, whether primary or complex, in a huge collection of multimedia data.

CCS CONCEPTS

• **Information systems**; • **Information retrieval**; • **Human-centered computing**; • **Interactive systems and tools**;

KEYWORDS

video event retrieval, interactive retrieval, information system, temporal search

*Both authors contributed equally to this research.

SOICT 2023, December 7–8, 2023, Ho Chi Minh, Vietnam

ACM ISBN 979-8-4007-0891-6/23/12…$15.00
https://doi.org/10.1145/3628797.3629019

ACM Reference Format:
Gia-Kiet Pham, Hai-Binh Tran, Phi-Long Nguyen, Song-Phuong Le, Hoang-Long Pham, Xuan-Tri Pham, Ham-Duong Tran, Ngoc-Tin Huynh, and Kiem Hoang. 2023. An Interactive System for Multimedia Retrieval in Video Collection with Temporal Integration. In *The 12th International Symposium on Information and Communication Technology (SOICT 2023), December 7–8, 2023, Ho Chi Minh, Vietnam.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3628797.3629019

1 INTRODUCTION

In an era characterized by the ever-expanding volume of video content, event retrieval within video collections has become paramount [19]. This task entails the meticulous process of searching for specific occurrences within vast video datasets, where insights are encapsulated in diverse forms of information, including visual concepts, audio transcriptions, object information, and temporal context. With the persistent interest of the research community in event retrieval, various competitions and tasks have been published to address the growing interest in this topic [16]. Recently, the Ho Chi Minh City Artificial Intelligence Challenge (HCMC AI Challenge) 2023 [1] has introduced a benchmark, serving as a platform to foster the development of innovative methods catering to the difficulties of event retrieval in news videos. The essence of the competition lies in developing an interactive retrieval system capable of retrieving keyframes that correspond to textual or visual descriptions. The queries encompass two categories: Video-Known-item Search (Video-KIS) and Textual-Known-item-Search (Textual-KIS). The final assessment is founded on a combination of correctness and followed by retrieval time for the submissions made by participating teams.

Due to the specialized nature of the up-to-date news programs, the surge in publicly available news videos, contributed by numerous publishers and spanning various facets of society, has underscored the significance of event retrieval in news video content. Its multifaceted applications encompass critical domains such as video surveillance, content management, news verification, and time-based event discovery. Thus, there is a need for a retrieval system that adopts a multifaceted approach, integrating audio transcription search, object positional retrieval, and a refined user interface while still maintaining meticulous attention to the need for speed.

[1] https://aichallenge.hochiminhcity.gov.vn/

As a participant in the HCMC AI Challenge 2023, the main contributions of our Pumpkin system revolve around enhancing the user's search experience.

- Our system empowers users to query video content through a wide range of information, including textual descriptions, audio transcripts, video metadata such as publishers and date-time information, object information within the contextual framework, or even an image from outside the dataset.
- We propose a temporal workflow to retrieve conceptually related scenes sequentially, which removes out-of-context results and gradually narrows down to the most suitable frames to the query.
- Furthermore, our user interface offers a range of utilities, including the ability to peruse the context surrounding a specific frame, commence video playback from the timestamp of a selected frame, and seek out similar frames within the dataset. This approach seeks to elevate scalable event retrieval systems and provide users with versatile tools even in the ever-expanding multimedia collections.

2 RELATED WORK

Nowadays, just a glimpse on social networking sites will give us a tremendous number of multimedia content, especially videos, that are being created and shared in every corner and aspect of life. Although these types of content are so simple today to create, store, and share, numerous challenges remain to access and manage these massive volumes of multimedia information dexterously. The demand for effective retrieval systems has been rising, dating back to 2001 with the first TRECVID[1] Video Event Retrieval Competition 2001. Nowadays, more workshops and competitions are still being held globally, attracting many participants, researchers, and cooperations, notably the Lifelog Search Challegene[4] LCS, Video Browser Showdown[12] VBS and the recently established AI Challenge Ho Chi Minh City in Vietnam AIC. These competitions all share a valuable task of video event retrieval, in which the participating team must retrieve the required answer from a massive multimedia collection, mainly video data.

The analysis of the comparison study by Heller, S., Gsteiger, V., Bailer, W. et al.[5] has given insight into previous iterations of VBS on its goals, tasks, scoring models, and notable participating systems. The analysis shows that recent trending answers to textual query tasks have been the effectiveness of embedding-based methods, with W2VV++[11] and its variations being the backbone of high-scoring systems, notably VIRET[13], and vitrivr[20]. At the same time, CLIP[18] was also utilizes by VIRET and several participated systems. According to Tran et al., [22], recent runs of LCS share the same trends, with the best three systems in LCS 2021 using CLIP for language-image embeddings (Voxento[2], LifeSeeker[17]). While there haven't been overviews of more recent systems, we have observed that there is a growing number of systems replacing W2VV++ and CLIP with BLIP in 2022 and 2023.

Both competitions feature the availability of Concept Detection in several teams with many different approaches to each other, for example using EfficientNet architecture[21] on different object datasets. Automatic Speech Recognition was also a powerful searching method for some participating systems with vitrivr, Exquisitor[9] in VBS, and Voxento in LCS relied on audio speech resources from the datasets' videos. Additionally, many systems were also capable of handling temporal context search, with some following an independent fusion approach, in which the score for an item is calculated by fusing its score with the best match for the second query result within a specified time range.

From the Overview of VBS and LCS of 2021 and before, not many systems emphasize user interface due to complex functions and the need for high retrieval speed during competitions. However, as pointed out by Heller, S., Gsteiger, V., Bailer, W. et al, a strong focus on the user interface is one of the significant keys to pursuing high-efficiency interactive search on extensive collections. In LCS'23, several high-scoring systems have proposed new and innovative user interface designs, notably LifeLens[6] with a minimal, consistent, and intuitive design that focuses on interactive user experience based on the user perspective of the systems through testing and participating in the LCS'23 competition.

As we capture an overview of the context, the Pumpkin system was conducted to enhance the user searching experience that is unified with a scalable retrieval system. Aside from the capability to perform image-text association retrieval, our system also enables users to perform searches by speech audio, object concepts, and metadata information—besides a convenient and persistent user interface with a cautious concern in accuracy and searching time.

3 METHODOLOGY

The Pumpkin system has been built to dexterously handle retrieval tasks from the AI Challenge 2023 dataset while providing an innovative and user-friendly interface (section 3.4). Figure 1 illustrated Pumpkin's central architecture comprising three main blocks: the content analysis backend, the data and queries management middleware, and the web-based user interface frontend, which enables users to interact with the data. This section will disclose each block in detail.

3.1 Content Analysis

Table 1: Number of duplicate and near-duplicate frames in 1,143,745 frames of AIC'23 dataset.

Max Hamming Distance	Duplicate Frames	Remaining Frames
6	277,822	865,923
8	353,128	790,617
10	423,884	719,861
12	489,827	653,918

The imminent goal of the Content Analysis block is to convert the raw dataset of a multimedia collection into a searchable database. Generally, individual frames from the result of the scene Splitting sub-block and Duplicate Removal sub-block will be fed through the Object Detection sub-block, Speech Extraction and Word Phrases Embeddings sub-block, Image-Text Embedding sub-block to generate object information, audio transcript information, and image-text embeddings.

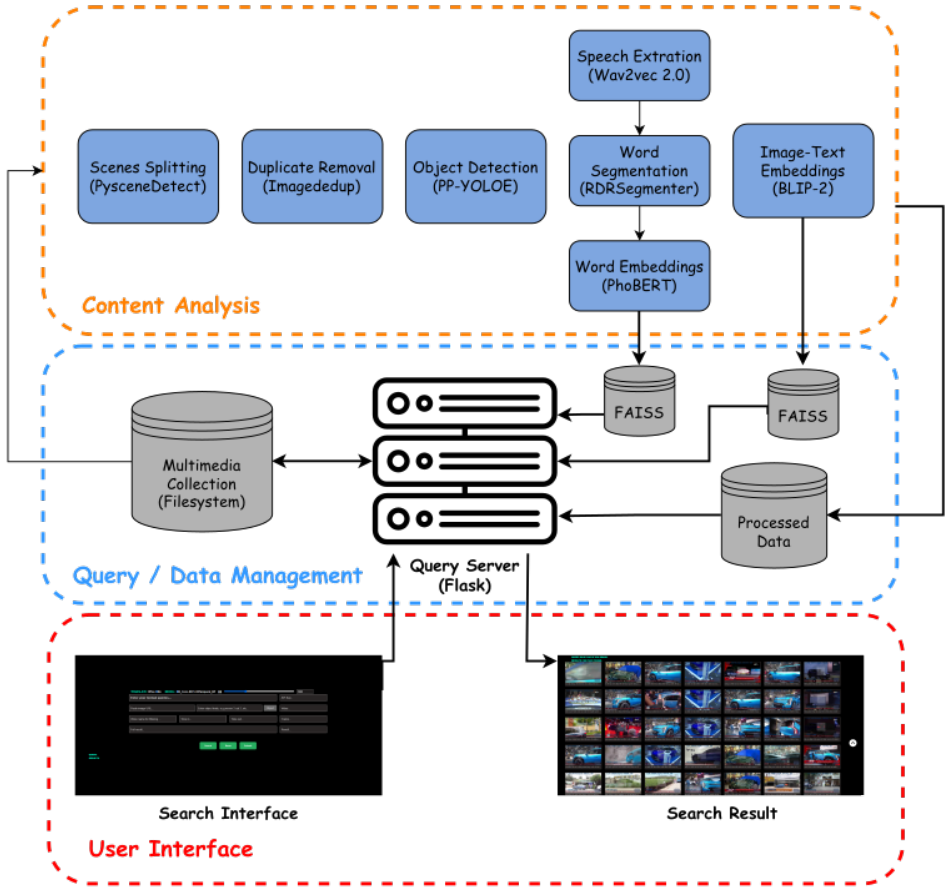

Figure 1: The System Architecture of Pumpkin.

Table 2: Overview of analysis components used by Pumpkin.

Type	Model/Method	Note
Scenes Splitting	PySceneDetect	Adaptive Detector with Detection Threshold T = 5.
Duplicate Removal	Imagededup	PHash metric is mainly used. CNN might yield better results, but it is much slower.
Object Detection	PP-YOLOE+	Pre-trained model from PaddleDetection Pipeline, trained on COCO dataset with 80 object categories.
Speech Extraction	Wav2vec2_base_vi_530h	Model trained by the authors on Vietnamese Speech Datasets.
Word Segmentation	RDRSegmenter	From VnCoreNLP Pipeline.
Word Embeddings	phobert-base-v2	Pre-trained model from PhoBERT, trained on 20GB of Wikipedia and News texts + 120GB of texts from OSCAR-2301.
Image-Text Embeddings	blip2_coco	Pre-trained model from LAVIS Pipeline, trained on COCO dataset.

At first, each video in the dataset video collection is split into a corresponding set of frames with the help of PySceneDetect [2], a video cut detection and analysis tool. PySceneDetect calculates the differences between scenes by three main detection algorithms: Content Detector, which detects shot changes by considering pixel changes in the HSV colorspace; Threshold Detector, which detects transitions below a set pixel intensity (cuts or fades to black); Adaptive Detector which is a two-pass version of Content Detector that in general handles fast camera movement better. These differences are denoted by percentage, called detection threshold T; a smaller T will result in PySceneDetect considering two frames belonging to different scenes, even if the visual differences are barely noticeable. From our observation, the detection threshold T of 20 would be sufficient for most normal videos, while T of 10 is more suitable for videos having fast-paced movements like sports events. For the AIC'23 dataset of all types of events, including fast-paced shots, low-quality videos, and sports events, we ran PySceneDetect through all

[2]https://github.com/Breakthrough/PySceneDetect

videos with the choice of Adaptive Detector and detection threshold T of 5%. The process created 1,143,745 individual frames from 300 hours of video from the dataset. However, the mentioned process can create up to many near-duplicate frames with the choice of a low detection threshold. On the one hand, these near-duplicate frames are vital for the performance of retrieval on scenes with fast camera movement or scenes that need attention to small details. On the other hand, most of the time, these types of scenes do not comprise a large proportion of the dataset. Thus, the remaining majority of the dataset will have many redundant duplicate and near-duplicate frames that provide little to no additional information compared to the original ones. To resolve this, Imagededup [7] is utilized to group similar frames across the entire frames collection, and from this, Pumpkin can perform duplicate removal and shots clustering of duplicate and near-duplicate frames. During AIC'23, we used the PHash method for its speed as we only had a limited amount of time to work on the final round dataset, but it is worth mentioning that the CNN method will result in the better group for its ability even to detect crops, resizes and rotations. Table 1 depicts our four choices of Max Hamming Distance for the PHash method during AIC'23 with the number of duplicate and remaining frames of 1,143,745 frames we generated from the Scenes Splitting sub-block.

The frames are then fed through the most critical part of the system backend - Image-text Embeddings sub-block. With the state-of-the-art performance in image-text retrieval and zero-shot capacity of recent BLIP-2[10], the models show high robustness in image-text retrieval tasks even without the need for fine-tuning or retraining, making it our primary choice as the embeddings extractor of Pumpkin. We use a pre-trained BLIP-2 model provided by the authors, which was trained on the COCO dataset for the image-text retrieval task. The embeddings can later be used for free text and image similarity searches.

We used PP-YOLOE[23] to obtain object information from frames, which was pre-trained on COCO for 80 different object categories. PP-YOLOE has high accuracy speed and excels in detecting multiple objects in a single frame. We filtered detected objects with a pre-defined threshold above 0.5 and used the highest confidence of an object category if there were many detections of the same object category. For detecting multiple categories of objects, the system chooses the frame with the highest sum of the detected categories, each following the same principle. The relative positions in terms of an object's horizontal and vertical positions are also calculated based on its bounding box coordinates.

In the Speech Extraction sub-block, audio segments are directly extracted from the video collection of the dataset and then have their corresponding transcription predicted by a pre-trained Wav2vec 2.0[3] that we trained and fine-tuned on public Speech Recognition datasets of Vietnamese language. Each frame from the Scenes Splitting and Duplicate Removal sub-blocks are then matched with a transcript based on their time in their video. Transcript sentences are then segmented by RDRSegmenter[15], a Vietnamese Segmenter, and each segment has its embedding extracted by PhoBERT [14], pre-trained language models for Vietnamese word text embeddings.

Table 2 summarizes different components utilized in the Pumpkin content analysis layer with their employed models or methods with descriptions.

3.2 Query / Data Management

In the middle session, this system will obtain queries in several forms to specify the highest relevant results. The system can work well in separate or integrated retrieval descriptions based on intersecting algorithms to combine and narrow down search space. The system ranks the result with speech or audio content, visual-textual association description, and image URL by calculating the cosine similarity through defined indexes conducted through the FAISS [8] platform. Besides, other insights, such as object positional information and video metadata, have been used as additional queries for filtering results and enhancing the final submission. In addition, the system provides utils for tracking and logging the retrieving history and a function to combine results from multiple searching times. At each step, the system enables users to continue the search by selecting a particular item in response results.

3.3 Temporal Search

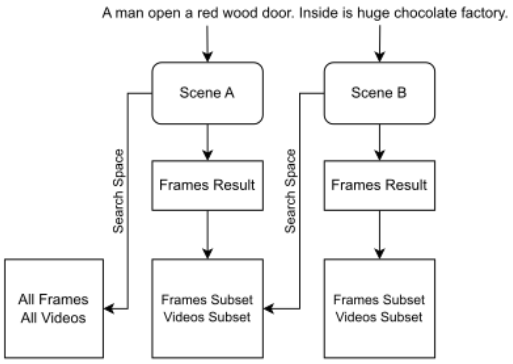

Figure 2: Temporal Workflow of Textual Queries.

Figure 2 depicts the temporal workflow of the Pumpkin system. At first, the retrieval space is the entire processed dataset. After input, the system splits the original query into many consecutive scenes. When a scene is processed, Pumpkin narrows down the entire processed dataset to only a subset of frames and videos based on the search result of the scene; this subset of frames and videos becomes the search space for the next scene and onwards. This process repeats indefinitely until there are no more unprocessed scenes from the original query, and the last scene's result is the temporal result of the entire query. To be specific in the narrowing process, we collect the frame information of each result frame, consisting of its video source V and the time (second) the frame appears in that video T; then we append the frames of V in the time range of (T, T+ delta) to the new search space. We chose a delta of 1000 for the AIC'23 dataset since each program in the news doesn't last longer than 40 seconds.

3.4 User Interface

To enhance the user experience and facilitate efficient retrieval of video events based on user queries, we have developed a user interface consisting of two main sections: the search section and the results section. Figure 3 demonstrates these sections of the user interface and the components contained within each section. This interface caters to the needs of both expert and novice users, providing simplified interaction and comprehensive search capabilities. The query section (1) encompasses features designed to accommodate user requirements. These features include: (A) Text-based search (B) Image-based search bar (C) Object recognition-based search bar (D) Video filtering bar.

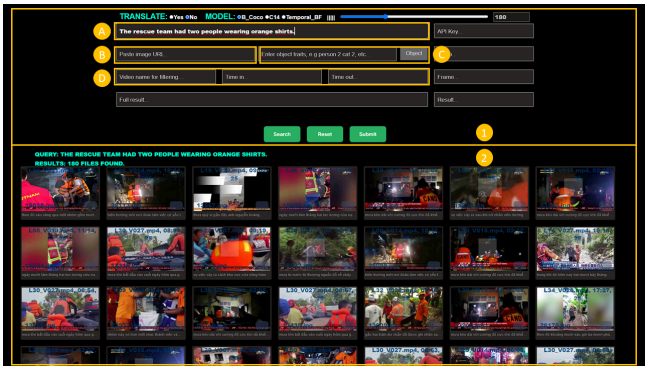

Figure 3: The Pumpkin's User Interface. The screen displays the query result, "The rescue team has two people wearing orange shirts."

Figure 4 shows the object section activated when users click the "Object" button. This appears as a pop-up with two distinct areas: one for selecting the position within a 3x3 rectangle representing the relative position on the frames and another for selecting the specific object (e.g., airplane, apple, etc) base on 80 object categories of COCO dataset. Once users have made their selections, they can save the object and position description by clicking the Save button, which saves the object and position description and forward it to the input system. A comma separates each object description. When users click the "Close" button, the description shown in the pop-up is instantly displayed in the Object search bar located adjacent to the "Object" button.

Our user interface's result section (2) displays the top K (K adjustable) retrieval results sorted by descending scores of each modules. Each thumbnail represents a video event and provides a concise content preview, enabling users to make well-informed choices. To facilitate seamless exploration, the result section also includes playback controls that allow users to preview the video directly within the interface. Users can play the video backward or forward by simply dragging the cursor the left or right of the thumbnail, enabling quick assessments and evaluations without having to load the entire video. To enhance accessibility and ease of use, each video thumbnail is accompanied by an audio description of the speech event occuring during the time of the frame. Pumpkin has developed an audio description line for each video, which appears at the bottom of the thumbnail. This audio description provides a

Figure 4: The object search pop-up.

summary or critical details about the video event. Users can click on the audio description to expand its content, gaining access to more comprehensive information about the event while conducting their search. Moreover, each thumbnail represents a keyframe and displays its name and the corresponding video's name. Clicking on the video name opens a new page displaying the video for playback. An important feature is the ability to pause the video and click the submit button, which automatically fills the keyframe and video name in the submission section for convenient checking during competitions without the need to return to the home page. Additionally, when clicking on its name on the thumbnail, another page displays approximately 50 frames before and after the selected keyframe. In the bottom right corner, a button for quick filtering is available. Clicking on this button instantly populates the Video Filtering bar with the video name, facilitating the search for related keyframes of that specific video. Furthermore, users can control the number of thumbnails displayed in the result section using a Range slider at the top of the search section. This feature allows users to customize the number of results they wish to view, providing greater flexibility and control over the displayed content. Above the query section exists an the option to cycle between different types of visual-text association methods, namely single frame with BLIP-2 or temporal search. We also implemented a translate option which translate the query text from Vietnamese to English automatically as well as a results range slider. Initially, there were to be a lot of option to click through and as we recognized the tedious process, Pumpkin also have built-in keyboard shortcut combination for nearly all query operation, including method cycles, translate cycles, search bar focusing and page refresh etc.

3.5 Pumpkin at AIC'23

For the early AIC 2023 Public Test phases, Pumpkin achieved a high-score even when some retrieval functions had not been completed by that time. Since we had more time to refine the results before submitting, it was an easy task when it came to accuracy for our system. Figure 5 shows the Pumpkin performance in three phases of the Public Test with a total score of 86.8 and 96,4% accuracy.

Participating in the AIC Final Round, Pumpkin can compete with the top rankers at AIC 2023. A standout feature of Pumpkin's system is its robust Temporal Search capability, which accurately identifies specific events frame by frame. This proficiency enables our system to retrieve answers through video querying rapidly. Notably, our Temporal Search delivered an impressive performance,

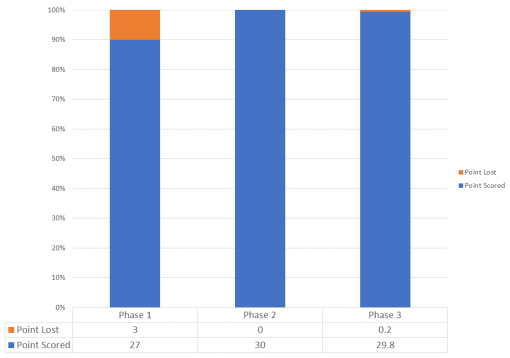

Figure 5: Pumpkin Performance at AIC Public Test 2023

with an average submission time of only 10 seconds after presenting the video query. However, during this round, we encountered a discrepancy in performance when it came to text queries. While our Temporal Search excelled in video queries, it proved less accurate and slower than our free text search module when presented with text queries. The description of text queries is only provided after 1 minute, requiring extensive waiting time to be displayed. In contrast, the free text search module showcased more substantial capabilities by quickly identifying the query within a single frame. Consequently, Pumpkin has decided to leverage the free text search module for text queries and continue utilizing Temporal Search for video queries. It is important to note that the performance of our system has been adversely affected by the utilization of an SSH server. The reliance on remote servers and poor internet signal at the competition have resulted in slower search processes for Pumpkin than teams that hosted their system locally. This drawback, stemming from the slower load times for image tasks due to remote server usage, poses a significant challenge in a competition that places emphasis on swift information retrieval and submissions.

4 AIC'23 DATASET

4.1 Dataset Description

The AIC'23 dataset was collected from YouTube channels such as 60 Giay Official[3], HTV Sport[4], HTV Entertainment[5], and Tuoi Tre[6]. Approximately 1213 videos were collected, with a total duration of 296.3 hours. The videos have varying bitrates, primarily with a frame rate of 25fps, with some at 23-30 fps. The videos can be categorized into local and international news with genres such as news, sports, entertainment, culture, travel, and education. Of these videos, the genres of entertainment, culture, travel, and education each account for about 10 percent, sports videos only account for about 5 percent, and the rest make up about 55 percent, as shown in Figure 6. The duration of the videos ranges from 6 to 8 minutes for HTV channels and from 14 to 26 minutes for channels like 60 Giay Official and Tuoi Tre.

[3]https://www.youtube.com/@TintucthoisuVietnam
[4]https://www.youtube.com/@ThethaoHTV
[5]https://www.youtube.com/@EntertainmentHTV
[6]https://www.youtube.com/@baotuoitre

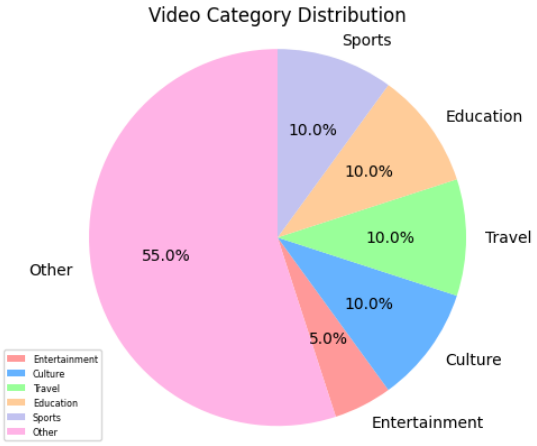

Figure 6: Video Category Distribution

4.2 Challenges

While constructing and testing video retrieval systems for AIC'23, we encountered many challenges related to image quality and missing metadata in the dataset. Some specific challenges are:

- There are many video genres, but they are not arranged. This requires systems to process and store data effectively and quickly.
- Metadata is missing or inaccurate in some videos, such as missing some frames. This makes filtering and sorting videos by various criteria more difficult.
- The image quality of the videos is inconsistent, with some videos being blurry or noisy. This affects the ability of systems to recognize faces, objects, and emotions.
- Some scenes are obscured by objects (according to our statistics, not many), such as trees, people, or signs. This reduces the accuracy of systems in detecting and recognizing essential elements in the video.
- Videos have a lot of content; in addition to the main news, there is also an introduction and ending that causes confusion in finding the exact video segment. This requires systems to segment and classify video segments according to different content and purposes.

5 EVALUATION

To evaluate the system's effectiveness, we conducted a survey based on some of the queries of the AIC'23 competition. The query set of the AIC'23 competition is very diverse, but we will only use a few representative queries to conduct our system survey.

For query requests involving consecutive events that do not occur in the same frame, the effectiveness of models like CLIP or BLIP will be lower than the Temporal model, as depicted in Figure 7. For example, in the case of *"A video with the appearance of a man in a white shirt and a girl in a dress at the park, and then a bicycle appears amongst them later"*, when using the Temporal model, the system will first query the initial description in the sentence, *"a man in a white shirt and a girl in a dress at the park"*, with the BLIP-2-COCO model. Subsequently, the search space within the

resulting video clips will be constrained to query the *"next scene: a bicycle appears amongst them."* If both descriptions exist within the narrowed search space and the scene with the bicycle appears after the initial scene, videos containing both descriptions will be selected and ranked based on cosine similarity scores.

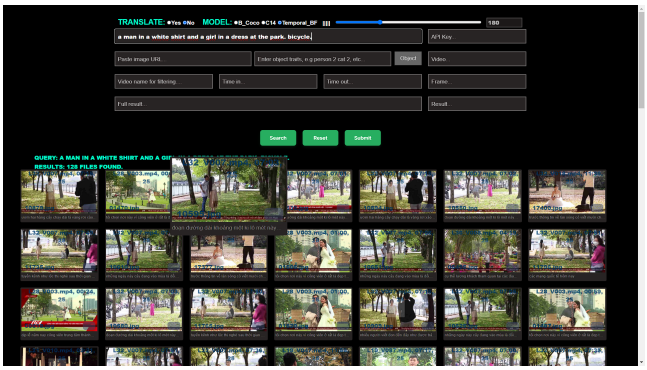

Figure 7: An example using Temporal Search to retrieve a video event with multiple consecutive event descriptions.

We need to consider the audio component to identify the exact video when dealing with queries in which the desired scene appears in multiple videos. Therefore, our system utilizes speech-to-text technology to transcribe the audio descriptions. This process allows us to precisely determine the video's description and the event's location within the video, for example, with a description like *"a scene of the city before the earthquake in Turkey, with the appearance of a car at the beginning of the video, making a right turn at an intersection"*. It is evident that the specific detail *"a car at the beginning of the video, making a right turn at an intersection"* serves as a helpful cue for the system to search for visually matching images in the description. However, there could be numerous frames containing such descriptive content. To pinpoint the exact video we are looking for, we review the audio transcript associated with each frame on the system's user interface to confirm the match. This ability to access the audio component of the video streamlines the process. It enables us to find the correct video more rapidly without the need to watch each video individually.

In another scenario, object-based search enables us to quickly obtain results when the description demands precision in arranging objects within the frame. For instance, consider the description: *"In the frame, with the right half obscured by a white wall, on the left, a girl leans against a house pillar, there is a bright blue painting next to the girl, and multiple woven mats lie in the lower part of the frame."* The system divides the frame into nine equal parts with a 3x3 grid, as depicted in Figure 4. In each grid cell, the system allows users to select and arrange multiple objects in their appropriate positions, thus facilitating the swift identification of frames that match the description accurately.

For queries that refer to events that happened before and after, when an event can appear in multiple videos, it is more accessible to know which video is correct by clicking to watch it. For such challenges, our system can still solve the problem by allowing users to see forward and backwards. In this way, the system helped us identify the correct video faster, saving more time than having to click on the video to watch the entire video. For example, *"A blue KIA car is displayed, followed by a scene where the car opens its two doors and the interior is visible"*. Using forward/backwards, we can determine the event that took place later and find the exact video. Specifically, there are many videos with a blue KIA car on display. With our system, we do not need to watch each video to find a video with this car opening its two doors and being able to see the interior. We can find the exact video faster using the forward/backward viewing feature.

6 CONCLUSION

In conclusion, the paper presents our efforts in building a scalable, multimodal, and efficient system for multimedia retrieval in video collections while maintaining an intuitive and convenient user interface. Our system empowers users to query video content through a wide range of information, including textual descriptions, audio transcripts, video metadata such as publishers and date-time information, object information within the contextual framework, or even an image from outside the dataset. Furthermore, our user interface offers a range of integrated temporal utilities to boost the performance of the search process by gradually narrowing down the search space in a query. This approach seeks to elevate scalable event retrieval systems and provide users with versatile tools even in the ever-expanding multimedia collections in the competition and practical scenarios.

REFERENCES

[1] 2018. *ITI-CERTH participation in TRECVID 2017.* Zenodo. https://doi.org/10.5281/zenodo.1183440

[2] Ahmed Alateeq, Mark Roantree, and Cathal Gurrin. 2021. Voxento 2.0: A Prototype Voice-Controlled Interactive Search Engine for Lifelogs. In *Proceedings of the 4th Annual on Lifelog Search Challenge* (Taipei, Taiwan) *(LSC '21)*. Association for Computing Machinery, New York, NY, USA, 65–70. https://doi.org/10.1145/3463948.3469071

[3] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *CoRR* abs/2006.11477 (2020). arXiv:2006.11477 https://arxiv.org/abs/2006.11477

[4] Cathal Gurrin, Björn Þór Jónsson, Klaus Schöffmann, Duc-Tien Dang-Nguyen, Jakub Lokoč, Minh-Triet Tran, Wolfgang Hürst, Luca Rossetto, and Graham Healy. 2021. Introduction to the Fourth Annual Lifelog Search Challenge, LSC'21. In *Proceedings of the 2021 International Conference on Multimedia Retrieval* (Taipei, Taiwan) *(ICMR '21)*. Association for Computing Machinery, New York, NY, USA, 690–691. https://doi.org/10.1145/3460426.3470945

[5] Silvan Heller, Viktor Gsteiger, Werner Bailer, Cathal Gurrin, Björn Þór Jónsson, Jakub Lokoč, Andreas Leibetseder, František Mejzlík, Ladislav Peška, Luca Rossetto, Konstantin Schall, Klaus Schoeffmann, Heiko Schuldt, Florian Spiess, Ly-Duyen Tran, Lucia Vadicamo, Patrik Veselý, Stefanos Vrochidis, and Jiaxin Wu. 2022. Interactive video retrieval evaluation at a distance: comparing sixteen interactive video search systems in a remote setting at the 10th Video Browser Showdown. *International Journal of Multimedia Information Retrieval* 11, 1 (March 2022), 1–18.

[6] Maria Tysse Hordvik, Julie Sophie Teilstad Østby, Manoj Kesavulu, Thao-Nhu Nguyen, Tu-Khiem Le, and Duc-Tien Dang-Nguyen. 2023. LifeLens: Transforming Lifelog Search with Innovative UX/UI Design. In *Proceedings of the 6th Annual ACM Lifelog Search Challenge* (Thessaloniki, Greece) *(LSC '23)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3592573.3593096

[7] Tanuj Jain, Christopher Lennan, Zubin John, and Dat Tran. 2019. Imagededup. https://github.com/idealo/imagededup.

[8] Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data* 7, 3 (2019), 535–547.

[9] Omar Shahbaz Khan, Björn Þór Jónsson, Mathias Larsen, Liam Poulsen, Dennis C. Koelma, Stevan Rudinac, Marcel Worring, and Jan Zahálka. 2021. Exquisitor at the Video Browser Showdown 2021: Relationships Between Semantic Classifiers. In *MultiMedia Modeling*, Jakub Lokoč, Tomáš Skopal, Klaus Schoeffmann, Vasileios

Mezaris, Xirong Li, Stefanos Vrochidis, and Ioannis Patras (Eds.). Springer International Publishing, Cham, 410–416.

[10] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597* (2023).

[11] Xirong Li, Chaoxi Xu, Gang Yang, Zhineng Chen, and Jianfeng Dong. 2019. W2VV++: Fully Deep Learning for Ad-hoc Video Search. https://doi.org/10.1145/3343031.3350906

[12] Jakub Lokoč, Patrik Veselý, František Mejzlík, Gregor Kovalčík, Tomáš Souček, Luca Rossetto, Klaus Schoeffmann, Werner Bailer, Cathal Gurrin, Loris Sauter, Jaeyub Song, Stefanos Vrochidis, Jiaxin Wu, and Björn þóR Jónsson. 2021. Is the Reign of Interactive Search Eternal? Findings from the Video Browser Showdown 2020. *ACM Trans. Multimedia Comput. Commun. Appl.* 17, 3, Article 91 (jul 2021), 26 pages. https://doi.org/10.1145/3445031

[13] Jakub Lokoč, Gregor Kovalčík, Tomáš Souček, Jaroslav Moravec, and Přemysl Čech. 2019. VIRET: A Video Retrieval Tool for Interactive Known-item Search. 177–181. https://doi.org/10.1145/3323873.3325034

[14] Dat Quoc Nguyen and Anh Tuan Nguyen. 2020. PhoBERT: Pre-trained language models for Vietnamese. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. 1037–1042.

[15] Dat Quoc Nguyen, Dai Quoc Nguyen, Thanh Vu, Mark Dras, and Mark Johnson. 2018. A Fast and Accurate Vietnamese Word Segmenter. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*. 2582–2587.

[16] Thao-Nhu Nguyen, Tu-Khiem Le, Van-Tu Ninh, Cathal Gurrin, Minh-Triet Tran, Thanh Binh Nguyen, Graham Healy, Annalina Caputo, and Sinead Smyth. 2023. E-LifeSeeker: An Interactive Lifelog Search Engine for LSC'23. In *Proceedings of the 6th Annual ACM Lifelog Search Challenge* (Thessaloniki, Greece) *(LSC '23)*. Association for Computing Machinery, New York, NY, USA, 13–17. https://doi.org/10.1145/3592573.3593098

[17] Thao-Nhu Nguyen, Tu-Khiem Le, Van-Tu Ninh, Minh-Triet Tran, Nguyen Thanh Binh, Graham Healy, Annalina Caputo, and Cathal Gurrin. 2021. LifeSeeker 3.0: An Interactive Lifelog Search Engine for LSC'21. In *Proceedings of the 4th Annual on Lifelog Search Challenge* (Taipei, Taiwan) *(LSC '21)*. Association for Computing Machinery, New York, NY, USA, 41–46. https://doi.org/10.1145/3463948.3469065

[18] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020 [cs.CV]

[19] Jérôme Revaud, Matthijs Douze, Cordelia Schmid, and Hervé Jégou. 2013. Event Retrieval in Large Video Collections with Circulant Temporal Encoding. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*. 2459–2466. https://doi.org/10.1109/CVPR.2013.318

[20] Luca Rossetto, Ralph Gasser, Silvan Heller, Mahnaz Parian-Scherb, Loris Sauter, Florian Spiess, Heiko Schuldt, Ladislav Peška, Tomáš Souček, Miroslav Kratochvíl, František Mejzlík, Patrik Veselý, and Jakub Lokoč. 2021. On the User-Centric Comparative Remote Evaluation of Interactive Video Search Systems. *IEEE MultiMedia* 28, 4 (2021), 18–28. https://doi.org/10.1109/MMUL.2021.3066779

[21] Mingxing Tan and Quoc V. Le. 2020. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. arXiv:1905.11946 [cs.LG]

[22] Ly-Duyen Tran, Manh-Duy Nguyen, Duc-Tien Dang-Nguyen, Silvan Heller, Florian Spiess, Jakub Lokoč, Ladislav Peška, Thao-Nhu Nguyen, Omar Shahbaz Khan, Aaron Duane, Björn þór Jónsson, Luca Rossetto, An-Zi Yen, Ahmed Alateeq, Naushad Alam, Minh-Triet Tran, Graham Healy, Klaus Schoeffmann, and Cathal Gurrin. 2023. Comparing Interactive Retrieval Approaches at the Lifelog Search Challenge 2021. *IEEE Access* 11 (2023), 30982–30995. https://doi.org/10.1109/ACCESS.2023.3248284

[23] Shangliang Xu, Xinxin Wang, Wenyu Lv, Qinyao Chang, Cheng Cui, Kaipeng Deng, Guanzhong Wang, Qingqing Dang, Shengyu Wei, Yuning Du, and Baohua Lai. 2022. PP-YOLOE: An evolved version of YOLO. arXiv:2203.16250 [cs.CV]