Generating Summary Videos from User Questions to Support Video-Based Learning

KAZUKI KAWAMURA, The University of Tokyo, Tokyo, Japan and Sony CSL Kyoto, Kyoto, Japan

JUN REKIMOTO, The University of Tokyo, Tokyo, Japan and Sony CSL Kyoto, Kyoto, Japan

As the number of online educational videos continues to grow, learners find it increasingly accessible, yet challenging to find the information they need in this vast digital library. Current video search systems allow users to narrow down the videos of interest to some extent by entering appropriate search queries. However, the specific information the user is looking for is often scattered across multiple videos, and it is very inefficient for the user to watch them individually to gain the necessary knowledge. Furthermore, in order to identify the specific knowledge required within a given video, the user must devote a considerable amount of time and effort to carefully watching the video in its entirety. This is a time-consuming process that is not always feasible for learners. In this study, we present a method for generating concise lecture videos that address learners' questions and allow them to view knowledge scattered across a large number of videos in a unified manner. The system is designed to comprehend the user's query, select an appropriate number of videos from a range of available options, identify specific segments in the videos that correspond to the user's query, and generate a concise summary video comprising these selected segments. Our preliminary experiments will demonstrate that it is indeed feasible to create concise physical videos that address user queries by integrating multiple lecture videos.

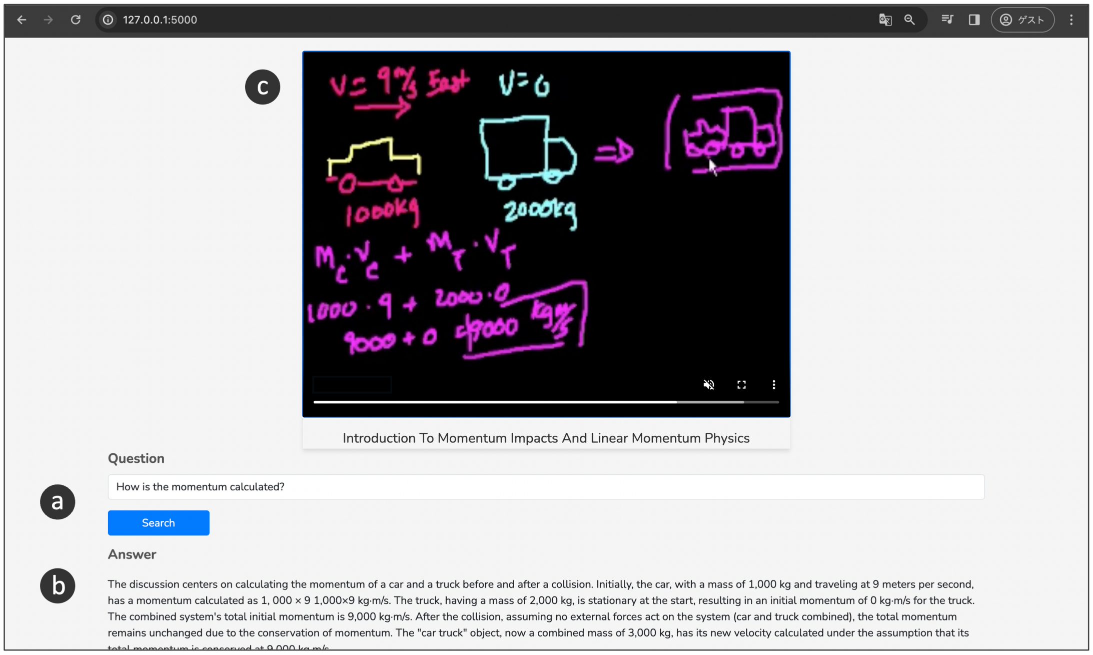
CCS Concepts: • **Human-centered computing** → **Interactive systems and tools**; • **Computing methodologies** → **Video summarization**.

Additional Key Words and Phrases: Video summarization, retrieval-augmented generation, e-learning, human–computer interaction, learning efficiency, large language model

1 INTRODUCTION

The proliferation of online educational videos has revolutionized the way learners access knowledge. Digital libraries of educational content continue to grow, giving learners easy access to vast amounts of information. However, this abundance of resources comes with its challenges. When learners search for the information they need, they often have difficulty finding the specific information they need because that information may be scattered across multiple videos. Furthermore, the necessity for learners to carefully watch the entirety of a video results in a significant investment of time and effort to identify the knowledge within a given video. Current video search systems allow users to narrow down the videos of interest to some extent by entering appropriate search queries. However, these systems still present the challenge that learners must navigate through multiple videos and manually locate the specific information they need within each video.

The objective of this approach is to centralize dispersed knowledge across numerous videos and enable learners to access the information they require in a single concise video. Our system is designed to understand the user's query and select a number of relevant video segments based on the video image and audio information from a vast number of videos. It then uses these video segments to generate a summary video. To demonstrate the feasibility of the proposed method, preliminary experiments are conducted on a set of physics instructional videos. These experiments demonstrate the system's ability to integrate multiple lecture videos to generate concise physical videos that effectively respond to user queries. The system will allow users to easily pinpoint their concerns in the videos.



Introduction To Momentum Impacts And Linear Momentum Physics

Question

How is the momentum calculated?

Answer

The discussion centers on calculating the momentum of a car and a truck before and after a collision. Initially, the car, with a mass of 1,000 kg and traveling at 9 meters per second, has a momentum calculated as 1, 000 × 9 1,000×9 kg·m/s. The truck, having a mass of 2,000 kg, is stationary at the start, resulting in an initial momentum of 0 kg·m/s for the truck. The combined system's total initial momentum is 9,000 kg·m/s. After the collision, assuming no external forces act on the system (car and truck combined), the total momentum remains unchanged due to the conservation of momentum. The "car truck" object, now a combined mass of 3,000 kg, has its new velocity calculated under the assumption that its total momentum is conserved at 9,000 kg·m/s.

Fig. 1. Interface of our system, where a user question “How is momentum calculated?” is entered (a), prompting the system to select an appropriate instructional video from a large database of physics-related content. The selected video is then concisely summarized to specifically address the query, with the summary playback shown (c) and its transcription displayed (b).

2 RELATED WORK

Video summarization techniques focus on condensing lengthy videos into shorter summaries that highlight the essential content, aiding in quicker comprehension. A survey on video summarization methods underscores the importance of keyframe selection and video skimming as core approaches [15]. The field has seen advancements with the integration of deep learning [1, 4], particularly using reinforcement learning to fine-tune the generation of summaries to better match human preferences [18]. Furthermore, in recent years, video summarization that considers both audio and image information of the video, utilizing the technology of generative AI, has also emerged [5, 6].

Interactive video systems seek to improve user engagement and learning outcomes through enabling interactions with video content in various forms, from clickable video elements to more complex platforms incorporating quizzes, branching scenarios, and real-time feedback [7, 9]. Studies have demonstrated significant boosts in engagement and learning achievements in online educational settings [17], with applications extending into marketing, entertainment, and training, showcasing the versatility and potential of interactive videos to provide access to engaging and personalized content [13].

Large language models, such as those in OpenAI’s GPT series, have significantly enhanced the precision and relevance of video search through sophisticated natural language processing [2, 8, 10]. These models are adept at interpreting intricate queries and navigating information across multimedia formats, merging visual and textual data to optimize video retrieval efforts [11]. The effectiveness of multimodal transformers in processing and responding to mixed media inputs further exemplifies the adaptability of language models, leading to improvements in video search and summarization [16]. Moreover, these models contribute to video captioning and descriptions, elevating accessibility and discoverability [19], thereby revolutionizing the ways in which videos are interacted with and accessed.

3 OUR SYSTEM

3.1 User Interaction

Our system is crafted to empower users to query an extensive library of video content and receive tailored video summaries that directly address their inquiries. This approach streamlines the process of sifting through long videos for specific information, thereby saving users time and enhancing their learning efficiency.

Utilizing our system, as illustrated in Fig. 1, users are able to efficiently navigate through extensive video content to find precise answers to their questions. For instance, a student studying for a physics exam can type a specific question such as “How is momentum calculated?” into the interface (a). The system then analyzes the query, searches through a comprehensive database of physics videos, and identifies the segments most relevant to the topic. Leveraging advanced summarization algorithms, it compiles these segments into a succinct summary video that is displayed (c), along with a transcript of the content (b). This allows the student to quickly obtain a focused explanation and visual demonstration of the concept without the need to watch longer videos in their entirety.

Similarly, a professional learning a new programming language can enter the query, “What is object-oriented programming?” The system processes this inquiry, sifts through various instructional videos, and creates a summary that distills the essence of object-oriented programming into a short, accessible format. This not only expedites the learning process but also provides the professional with a tailored educational experience, presenting complex information in a clear and manageable way.

3.2 Implementation

The implementation of our system comprises several key phases: understanding user queries, selecting relevant videos, segmenting chosen videos, and synthesizing a coherent audiovisual summary.

Processing Questions. The system uses a large-scale language model, such as GPT (Generative Pre-trained Transformer) [10], to understand the nuances of the user query Q and generate an initial answer A. The answer A assists in the video selection and segment selection by providing a context-rich representation of the user’s intent.

Video Selection. Initially, when a user question Q is input into the system, the Video Selection phase begins. The system evaluates a large database of educational video content and computes embeddings for each video V to find those that are most relevant to the query. The relevance is determined by calculating the cosine similarity between the embedding of the user’s question and the embeddings of the videos, expressed by the equation $\text{VideoScore}_i = \cos(\text{embed}(Q), \text{embed}(V_i))$ for each video V_i. The highest-ranking videos based on their similarity scores are selected for further processing.

Segment Selection. After selecting the relevant videos, the system then segments these videos into smaller parts S and performs a more fine-grained selection to identify the top K segments that are semantically closest to the user’s query. This is accomplished using the equation $\text{TopK} = \text{top-K}\{\cos(\text{embed}(Q), \text{embed}(S))\}$.

Summary Generation and Speech Synthesis. With the top K video segments identified, the system generates a written summary based on the transcripts of these segments. It does this by converting the visual and auditory information of the videos into text and then using ChatGPT [12] to create a summary narrative. By compiling multiple video segments

Table 1. Evaluation Results: Comparison of User Questions, Selected Video Segment, and Generated Summary Video

User Question	Selected Video Segment	Generated Summary Video
How is the momentum calculated?	**Introduction to momentum**: “So what’s the mass of the car? That’s 1,000. What’s the velocity of the car? It’s 9 meters per second. So as you can imagine, the unit of momentum, which is the ...”	“The discussion centers on calculating the momentum of a car and a truck before and after a collision. Initially, the car, with a mass of 1,000 kg and traveling at 9 meters per second, has a momentum ...”
Why is Kelvin used as the temperature scale?	**Thermodynamics part 3: Kelvin scale and Ideal gas law example**: "You don’t have 100 times the kinetic energy. So this is a bit of an arbitrary scale. So the actual interval might, you know, the interval’s arbitrary. You could pick the 1 degree ..."	"The video discusses the concept of temperature in relation to kinetic energy, explaining that the scale used to measure temperature can be somewhat arbitrary. It points out that the Celsius scale, for instance, starts at ..."

into a single summary, the system provides an efficient means of information absorption, superior to viewing multiple individual segments. Additionally, the system synthesizes narration in the same voice as the original video’s audio. This synthesized voice is synchronized with the playback of the selected video segments, providing users with an integrated audiovisual summary that directly answers their query.

4 PRELIMINARY EVALUATION

This preliminary evaluation is a investigation of how the system behaves in response to some questions related to video content. We fed the system with about 156 lecture videos on physics collected from Khan Academy[1] to see which videos and their segments in the video are extracted and what summary videos are generated when the user ask questions. Some of the results are shown in Table 1.

Our approach demonstrated promising results in several instances, where it could accurately identify the portions of a video that directly answered specific questions, thereby generating effective summaries. For example, when asked about the calculation of momentum, the system pinpointed the exact segment discussing the mass and velocity of a car, followed by a succinct summary highlighting the process of momentum calculation before and after a collision. Similarly, in response to a question about the Kelvin temperature scale, it located a segment that discussed the arbitrariness of temperature scales and provided a summary explaining the rationale behind using Kelvin as a measure.

However, our evaluation also found a challenge in that the system generates some kind of summary video for questions that are unrelated to the group of videos being targeted. This indicates a limitation in the system’s ability to identify the relevance of the query to the video content possessed by the system and may generate a summary that is not contextually appropriate.

5 DISCUSSION

Our preliminary evaluations suggest that the proposed system has the ability to provide concise video responses that directly address user queries. Future research will investigate the pedagogical impact of the system’s summary videos on the learning process. Specifically, we aim to investigate the pedagogical advantages of producing short summary videos, as facilitated by our methodology, over simply navigating to the relevant segments of the original video that

[1] https://www.youtube.com/watch?v=ihNZlp7iUHE&list=PLAD5B880806EBE0A4

are relevant to the user's question. Further, we pose a research question for future investigation: Is it conceivable to develop a system that, by pre-training on a wide range of videos, can emulate similar functionality without the need for real-time search? The current system, which relies on an RAG-like framework [3], can experience delays in search operations as the volume of videos to be searched escalates. Consequently, we are intrigued by the possibility of circumventing these limitations through the development of a learning-based generative model [14], and the potential impact such an approach could have on improving educational outcomes for learners engaging with video content.

ACKNOWLEDGMENTS

This work was supported by JST Moonshot R&D Grant JPMJMS2012, JST CREST Grant JPMJCR17A3, and the commissioned research by NICT Japan Grant JPJ012368C02901.

REFERENCES

[1] Evlampios Apostolidis, Eleni Adamantidou, Alexandros I Metsai, Vasileios Mezaris, and Ioannis Patras. 2021. Video summarization using deep neural networks: A survey. *Proc. IEEE* 109, 11 (2021), 1838–1863.

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Proc. of Advances in neural information processing systems* 33 (2020), 1877–1901.

[3] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint* (2023).

[4] Longlong Jing and Yingli Tian. 2020. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence* 43, 11 (2020), 4037–4058.

[5] Kazuki Kawamura and Jun Rekimoto. 2024. FastPerson: Enhancing Video-Based Learning through Video Summarization that Preserves Linguistic and Visual Contexts. In *Proceedings of the Augmented Humans International Conference 2024 (AHs '24)*. 205–216.

[6] Kazuki Kawamura and Jun Rekimoto. 2024. QA-FastPerson: Extending Video Platform Search Capabilities by Creating Summary Videos in Response to User Queries. In *Proceedings of the Augmented Humans International Conference 2024 (AHs '24)*. 290–293.

[7] Peter H Martorella. 1983. Interactive Video Systems in the Classroom. *Social Education* 47, 5 (1983), 325–27.

[8] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* 35 (2022), 27730–27744.

[9] Linda C Petty and Ellen F Rosen. 1987. Computer-based interactive video systems. *Behavior Research Methods, Instruments, & Computers* 19, 2 (1987), 160–166.

[10] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training. (2018).

[11] Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, AmirAli Bagher Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. 2020. Integrating Multimodal Information in Large Pretrained Transformers. In *Proc. of the Annual Meeting of the Association for Computational Linguistics*. 2359–2369.

[12] Partha Pratim Ray. 2023. ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems* (2023).

[13] Catharyn Shelton, Annie Hale, and Leanna Archambault. 2016. Exploring the Use of Interactive Digital Storytelling Video: Promoting Student Engagement and Learning in a University Hybrid Course. *TechTrends* 60 (2016).

[14] Aditi Singh. 2023. A Survey of AI Text-to-Image and AI Text-to-Video Generators. In *International Conference on Artificial Intelligence, Robotics and Control (AIRC)*. 32–36.

[15] Ba Tu Truong and Svetha Venkatesh. 2007. Video abstraction: A systematic review and classification. *ACM Trans. Multimedia Comput. Commun. Appl.* 3, 1 (2007), 3–es.

[16] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proc. of the conference. Association for Computational Linguistics. Meeting*, Vol. 2019. 6558.

[17] Sirui Wang and Huei-Lien Chen. 2016. Video That Matters: Enhancing Student Engagement Through Interactive Video-Centric Program in Online Courses. *thannual* (2016), 136.

[18] Kaiyang Zhou, Yu Qiao, and Tao Xiang. 2018. Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward. In *Proc. of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[19] Luowei Zhou, Yingbo Zhou, Jason J Corso, Richard Socher, and Caiming Xiong. 2018. End-to-end dense video captioning with masked transformer. In *Proc. of the IEEE conference on computer vision and pattern recognition*. 8739–8748.