ISSN 2663-8169 Vol. 7, Issue 1: 110-115, DOI: 10.25236/IJNDE.2025.070116

Knowledge-Based Teaching Video Retrieval and Localization Technology

Chao Pan[1,*], Yifeng Wang[1], Haihong Zheng[1], Fei Wang[2]

[1]*School of Computer Science and Technology, Xidian University, Xi'an, China*
[2]*School of Mechano-Electronic Engineering, Xidian University, Xi'an, China*
** Corresponding author*

Abstract: *The development of information technology has made 'Artificial Intelligence (AI)+ Education' widespread, generating a massive amount of classroom teaching videos and courseware text data. Knowledge-based multimodal teaching video retrieval and localization refers to the process of retrieving videos corresponding to specific knowledge points from teaching videos and locating fine-grained video moments based on query text related to those knowledge points. This approach can enhance teaching efficiency and support auxiliary teaching. However, current video retrieval and localization technologies are based on single-scene datasets and lack video-text datasets specifically for educational contexts. This research aims to develop knowledge-based teaching video retrieval and localization technologies. It involves constructing a video-text offline teaching dataset tailored for educational scenes, used for video retrieval and localization in educational contexts. A pre-trained model for the educational scene is employed to process data and extract features. By utilizing video-text retrieval and moment localization techniques, this study builds a smart education video retrieval and localization system. The system retrieves corresponding videos and locates the relevant knowledge point moments in the teaching videos quickly and accurately based on the input knowledge point query text. This contributes to both 'convenient teaching for teachers' and 'efficient learning for students.'*

Keywords: *AI+Education, Video Retrieval and Localization, Teaching Video*

1. Introduction

In recent years, the development of information technology has greatly promoted the transformation of learning methods[1]. With the rapid development of mobile internet, the widespread popularity of smart terminal devices, and the booming development and application of artificial intelligence, the education model has been continuously expanding[2]. The traditional offline teaching model, represented by classroom teaching, has begun to extend and develop into a smart education model. The teaching video data recorded in offline classrooms and the corresponding teaching courseware text data have accumulated rapidly. Classroom teaching videos, as one of the important forms of online education, have the advantages of rich content, high flexibility, and strong interactivity, and are widely used in educational activities. However, how to retrieve and locate valuable knowledge from the massive teaching data in order to meet the demand for diverse and personalized teaching services and knowledge navigation, and to better realize 'AI+ education', remains a problem that needs to be solved.

Video retrieval and localization technology provides a solution to the aforementioned problem. The goal of video retrieval and localization is to find the video segments that best match the description of a natural language query within a long video. Cross-modal retrieval utilizes deep learning, modality alignment, similarity metrics, and other techniques to extract features and match different modal data such as text, images, video, and audio[3]. It is a necessary approach to achieving accurate, comprehensive, and personalized retrieval. Chun et al. [4] calculated the matching distribution by computing the probability distribution between visual and textual embedding vectors to enable cross-modal retrieval; Wei et al. [5] used a weighted distance metric function to compute the similarity between visual and textual features; Lu et al. [6] proposed a collaborative dual-stream pre-training model that uses both masked language modeling and image-text matching tasks for pre-training. This technology has been widely applied in e-commerce, media, entertainment, and healthcare fields. However, it has not been extensively applied in the education field. The reason lies in the complexity of teaching video resources, which involve subject knowledge with complex data structures and high

ISSN 2663-8169 Vol. 7, Issue 1: 110-115, DOI: 10.25236/IJNDE.2025.070116

information density, as well as multiple modalities and fine-grained knowledge. These characteristics demand higher requirements for the in-depth exploration of modality information and the precise representation of modality features.

Teaching videos contain rich multimodal information, including visual, textual, and audio data. Therefore, understanding the information provided in the teaching videos and the textual information provided in the query, as well as aligning and interacting with cross-modal information, is a key issue in the task of retrieving and localizing teaching video segments.

(1) This research primarily focuses on the multimodal teaching video retrieval and localization technology based on knowledge points. Specifically, it aims to retrieve videos corresponding to knowledge points from a batch of teaching videos based on query texts, and localize fine-grained video moments, which can enhance teaching efficiency and promote auxiliary teaching.

(2) At the same time, addressing the issue that current video retrieval and localization technologies are based on datasets from a single scene and lack video-text datasets in educational contexts, this study will construct an offline teaching video-text dataset for the education scene, which can be used for video retrieval and localization in educational settings.

(3) Finally, the study will aim to develop a smart education system for teaching video retrieval and localization. Based on input knowledge point query texts, the system will accurately and quickly retrieve the corresponding videos and locate the relevant knowledge point video moments from a batch of teaching videos, thus enabling ‘convenient teaching for teachers’ and ‘efficient learning for students.’

2. Research Objectives and Content

This research aims to meet the construction needs of ‘AI+Education’ support systems. Based on massive multimodal classroom teaching resources, it focuses on the theory and methods for mining and analyzing classroom teaching resources. The study investigates language-guided multimodal joint representation methods to bridge the semantic gap between heterogeneous classroom data. It also explores teaching resource segmentation methods at the knowledge point granularity, enabling the organization and indexing of teaching resources at the knowledge point level to improve knowledge search efficiency and resource utilization. Furthermore, the research seeks to build an intelligent knowledge point teaching video retrieval system, which will provide students with learning video guidance at the knowledge point granularity and offer teachers intelligent organization of teaching resources, simplifying the lesson creation process. This will enhance students' learning abilities, strengthen classroom teaching effectiveness, and promote the intelligent development of school education. This research aims to combine the latest advancements in machine learning, information retrieval, data mining, and smart education to address the construction needs of ‘AI+Education’ support systems. Based on multimodal teaching video resources, the study focuses on knowledge-point-based teaching video retrieval and localization technology. The specific research contents are as follows:

2.1 Knowledge-point Granularity Teaching Video Segmentation

In the existing online learning platforms of schools, course resources need to be manually segmented and uploaded by teachers according to chapters or knowledge points, which increases teachers' workload. However, directly publishing teaching resources without segmentation would make it difficult for students to accurately find the resources related to specific knowledge points. To address this, the research will focus on segmenting and integrating teaching resources according to knowledge points, organizing and indexing them at the knowledge-point granularity.

2.2 Language-guided Multimodal Joint Representation

Classroom teaching resources include recorded videos, electronic handouts, electronic courseware, textbooks, syllabi, etc. The heterogeneity of the data and the differences in semantic granularity make it difficult to establish fine-grained semantic relationships between different modalities of data. Therefore, this research uses language information as a medium to extract correlated semantic features from multimodal data, reducing the heterogeneity between them, and studying knowledge-point association algorithms for multimodal data.

ISSN 2663-8169 Vol. 7, Issue 1: 110-115, DOI: 10.25236/IJNDE.2025.070116

2.3 Knowledge-point-based Video Retrieval Method

The study on knowledge-point-level video retrieval algorithms for classroom teaching videos focuses on how to automatically locate and retrieve video segments related to specific knowledge points from a large volume of video data. Specifically, this research will focus on: deep learning and computer vision analysis of teaching videos, extracting features and knowledge points from the videos, and constructing knowledge-point indexes. The retrieved video segments will then be presented to users, such as by playing the video or displaying the relevant knowledge points.

2.4 Intelligent Teaching Video Retrieval and Localization System

Based on the aforementioned research content, an intelligent video retrieval system based on knowledge points will be built. This system will provide students with knowledge-point-granularity learning resources to assist with reinforcing and consolidating knowledge after class. For teachers, the system will offer intelligent course construction services, enabling the automatic segmentation and organization of classroom teaching resources. The automatic segmentation simplifies the course creation process, while video retrieval avoids browsing large amounts of irrelevant content, achieving more precise teaching assistance and promoting the digital and smart transformation of educational teaching in schools.

3. Key Research Issues

The key scientific issues to be addressed in this research include:

3.1 Segmentation of Teaching Videos and Documents by Knowledge Points

Teaching videos are characterized by single scenes, long time spans, and low visual information density, making it difficult to use traditional video scene segmentation methods. Teaching documents, on the other hand, have long lengths and high keyword repetition rates, making it challenging to divide them based on knowledge point keywords. Therefore, the key scientific issue to be solved in this research is how to divide teaching videos and documents based on the knowledge points discussed, achieving knowledge-point granularity organization and indexing of teaching resources.

3.2 Semantic Alignment of Different Modal Data

Classroom multi-modal teaching data exhibit strong heterogeneity and large differences in semantic granularity, making direct analysis and processing difficult. Mapping different modal data to the same feature space and achieving semantic alignment across modalities is a prerequisite for fine-grained segmentation of teaching data at the knowledge point level. Effectively utilizing language information from different modalities helps establish relationships between different modal data. Therefore, the key scientific issue to be addressed is how to leverage language information from multi-modal teaching data to learn semantic features that are strongly correlated and have high representational power.

3.3 Precise Retrieval of Knowledge Points from Teaching Videos

Identifying, extracting, and understanding knowledge points and key information from massive video data for intelligent indexing and retrieval. This involves converting users' natural language query requests into computer-understandable information and performing precise knowledge point-level retrieval. Therefore, the key scientific issue is how to use machine learning and other technologies to calculate the similarity between query requests and video segments, enabling efficient search and retrieval of relevant video segments from a vast amount of video data.

4. Knowledge-Based Teaching Video Retrieval and Localization Technology

This research aims to meet the needs of the campus online tutoring service system by studying knowledge-point-based classroom video segmentation and retrieval, based on vast amounts of multimodal classroom teaching resources. The research will be based on multimodal pre-trained models and follows a research approach along the lines of ‘knowledge-point granularity in teaching resource segmentation, language-guided multimodal joint representation, and the construction of an

ISSN 2663-8169 Vol. 7, Issue 1: 110-115, DOI: 10.25236/IJNDE.2025.070116

intelligent teaching resource retrieval system.' The research will fully leverage the strengths of the applicant and research team members in the fields of smart education and multimodal big data intelligent analysis. The research will be conducted in stages, gradually deepening and implementing the research content according to the proposed technical route, overcoming key challenges, and achieving the research objectives. The research plan is described using the most common classroom data—such as recorded classroom videos, PPT slides, and textbook texts—as examples. However, the research approach is not limited to the aforementioned three types of data.

4.1 Knowledge-point granularity teaching resource segmentation

In offline teaching scenarios, teaching videos are typically recorded by classroom-installed surveillance equipment and divided according to class start and end times, archived with other materials such as the teacher's lecture notes. When students access these resources, it is difficult for them to quickly and precisely find the relevant content related to a specific knowledge point, reducing the efficiency of resource utilization and the overall learning experience. Moreover, due to the characteristics of teaching scenarios, teaching videos typically only include internal classroom footage and intermittently appearing blackboard writing or presentation slides. These videos are often long, have a single scene, and contain a lot of irrelevant information. Therefore, conventional video scene segmentation algorithms based on scene transitions are not suitable for segmenting teaching videos.

This research will first integrate techniques such as object detection and speech recognition to remove irrelevant video frames. Then, based on the inter-frame pixel distances, the frames will be clustered to eliminate redundant frames. By utilizing text recognition and other techniques, key frames related to knowledge points will be identified and used as the basis for video segmentation. For document materials, key sentences will be identified based on paragraph structures to divide the documents. The multimodal joint representation model to be constructed will unify the feature representation of knowledge points in the syllabus and corresponding video/document segments. Through correlation matching and incorporating the temporal or spatial sequence of the segments, fine-grained associations between different types of teaching data and syllabus knowledge points will be achieved.

4.2 Language-Guided Multimodal Joint Representation Learning

In real teaching scenarios, there are significant heterogeneities and semantic granularity differences between different modalities of teaching data, making it difficult to mine and establish fine-grained semantic relationships between these data. The multimodal data in teaching resources are predominantly language-based. Extracting language information from different modalities as a medium to establish relationships between them helps reduce the heterogeneity between modalities and obtain more powerful semantic features. Therefore, this study proposes to use language information extracted from different modalities to design a multimodal pre-trained model, employing contrastive learning. The model will project different modality features of the same knowledge point into a common feature space to reduce heterogeneity, integrate multimodal teaching features, and construct a knowledge point text decoder based on a language model. By utilizing self-supervised text generation tasks, the model will enhance the correlation and semantic similarity between features from different modalities.

4.3 Cross-Modal Knowledge Point Video Moment Retrieval

A 'Video-Text Offline Teaching Dataset' will be constructed, using video segments at the chapter granularity and aligned teaching materials as training data. The complete videos will be uniformly sampled into multiple non-repeating segments, with a visual encoder from the pre-trained multimodal model used to extract features from different segments, and a text encoder used to extract corresponding text features. For features of multiple segments of a single complete video, a Transformer structure will be used for text-related pooling, where a single text feature is used as the query object, and multiple video segment features are used as the objects being queried. Cross-modal attention mechanisms will be employed to aggregate the visual features into a single feature vector representation. Cosine similarity will be used as a similarity measure, with contrastive learning optimizing the loss function to enhance cross-modal video retrieval. During model application, the knowledge point text input by students will be matched with the hierarchical directory of teaching resources using keyword matching. This will allow the retrieval of relevant teaching resources, such as videos and electronic course materials. Based on browsing records, teaching resources from different

ISSN 2663-8169 Vol. 7, Issue 1: 110-115, DOI: 10.25236/IJNDE.2025.070116

teachers explaining the same knowledge point will be ranked and recommended to the students, providing intelligent guidance for fine-grained learning resources.

4.4 Construction of an Intelligent Teaching Video Retrieval System

This research proposes to establish a knowledge-point-based teaching resource retrieval system, which will include the following functionalities: intelligent organization of teaching resources and knowledge-point granularity learning resource guidance to enhance students' learning abilities and improve classroom teaching effectiveness.

4.4.1 Intelligent Classification and Organization of Teaching Resources

First, a knowledge-point hierarchy directory will be established based on the teaching syllabus. Second, after the class, teaching resources such as recorded classroom videos, electronic course materials, and textbooks will be finely segmented based on knowledge points and automatically uploaded to the system. Finally, using the multimodal representation model, the system will calculate the similarity between teaching resources and knowledge points, and intelligently associate the resources with the directory, enabling students to browse and access materials quickly and easily.

4.4.2 Knowledge-Point Granularity Learning Resource Guidance

The system will perform keyword matching between the knowledge point text input by students and the hierarchical directory of teaching resources. It will then retrieve teaching resources such as explanatory videos and electronic courseware related to the specified knowledge point. Based on browsing history, the system will rank and recommend teaching resources from different teachers explaining the same knowledge point, providing intelligent guidance for fine-grained learning resources.

5. Conclusion

This research focuses on the construction needs of the smart campus education service system. Based on vast classroom teaching resources, it studies the analysis theories and methods of multimodal classroom big data, designs knowledge-point-based teaching video retrieval methods, and addresses issues that are both innovative and forward-looking. The main innovations of this study can be summarized as follows:

5.1 Proposing an adaptive knowledge-point-level video segmentation algorithm

By combining the characteristics of teaching videos and documents, key frames from the video and key sentences from the document are extracted. These key frames and sentences are used to segment candidate video and document fragments. The similarity between the candidate fragments and knowledge points is then calculated, enabling the segmentation of teaching resources based on the content of the knowledge points. This lays the foundation for the knowledge-point granularity guidance of teaching resources.

5.2 Constructing a language-guided multimodal pre-training model

Given the characteristics of classroom multimodal teaching data, this study uses language information from different modalities to establish the relationships between different modalities. Through self-supervised contrastive learning and text generation methods, a model is trained using large-scale classroom data to learn multimodal semantic features that are strongly correlated and well-represented, providing a solid foundation for future research.

5.3 Designing a knowledge-point-based intelligent video retrieval method

By uniformly sampling multiple non-repeating fragments from the entire video, the visual encoder in the cross-modal pre-training model is used to extract features from different fragments, and the text encoder is used to extract corresponding text features. For the features of multiple fragments from a single video, the Transformer structure is used for text-related pooling, improving the cross-modal teaching video retrieval performance.

ISSN 2663-8169 Vol. 7, Issue 1: 110-115, DOI: 10.25236/IJNDE.2025.070116

5.4 Building an intelligent retrieval and localization system for teaching videos

Based on the knowledge-point segmentation method and intelligent retrieval methods for multimodal teaching resources, this research constructs a teaching video retrieval system. The system provides knowledge-point granularity learning resource guidance, intelligent organization of teaching resources, and other functionalities, aiming to enhance students' learning ability, strengthen classroom teaching effectiveness, and empower 'AI+Education.'

Acknowledgement

This research was supported by the Fundamental Research Funds for the Central Universities, Xidian University, Grants No. QTZX24063 and No. QTZX23084.

References

[1] Chen L, Chen P, Lin Z. Artificial intelligence in education: A review[J]. Ieee Access, 2020, 8: 75264-75278.
[2] Cantú-Ortiz F J, Galeano Sánchez N, Garrido L, et al. An artificial intelligence educational strategy for the digital transformation[J]. International Journal on Interactive Design and Manufacturing (IJIDeM), 2020, 14: 1195-1209.
[3] George B, Wooden O. Managing the strategic transformation of higher education through artificial intelligence[J]. Administrative Sciences, 2023, 13(9): 196.
[4] Chun S, Oh S J, De Rezende R S, et al. Probabilistic embeddings for cross-modal retrieval[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 8415-8424.
[5] Wei J, Yang Y, Xu X, et al. Universal weighting metric learning for cross-modal retrieval [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 44(10): 6534-6545.
[6] Lu H, Fei N, Huo Y, et al. COTS: Collaborative two-stream vision-language pre-training model for cross-modal retrieval [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 15692-15701.