Research and Implementation of Multi-modal Video Retrieval System Based on Deep Learning

Qi Bing
Hainan Technology and Business College, Hainan Haikou, 570100
tg667788@xzcstudio.com

Abstract: At present, keyword-based video retrieval is gradually difficult to adapt to the needs of the rapid development of the Internet due to its strong subjectivity and huge workload. As a result, multi-modal video retrieval based on deep learning has appeared. This retrieval method can conduct video retrieval through multiple methods such as text, image, and video, which fully meets the different retrieval needs of different users, and significantly improves the accuracy and effectiveness of video retrieval. Based on this, this article discusses in detail the design of a multi-modal video retrieval system based on deep learning, analyzes and designs each functional module of the system to provide reference for future related work.

1. Introduction
In recent years, with the development of the Internet and multimedia technologies, the number of online videos has continued to increase. The amount of video and the average number of views per day have reached hundreds of millions. Online videos have become an indispensable part of people’s daily lives. For video sharing websites, how to develop video retrieval services to deepen information exchange has become an important research. As a result, the optimization of the online video retrieval process and the innovation of online video retrieval solutions have become the research focus in related fields.

2. Overview of video retrieval
2.1. Basic principles of video retrieval
Video retrieval methods can be divided into text-based video retrieval and content-based video retrieval. At present, text-based video retrieval methods are more widely used, which manually mark the text of the video, and then use the marked videos to build a database. In this way, the user can search for matching videos by entering keywords. This retrieval method is easy to understand, simple to operate, and highly efficient. But the shortcomings of this retrieval method are also obvious. Because the video comprehension and annotation are made by manual methods, it is inevitable that the annotation of the video is not objective enough due to strong subjectivity. Therefore, this method still needs further improvement.

The content-based video retrieval method is different, which mainly analyzes the video content. According to the data information obtained by the analysis, the video feature database is constructed to realize the retrieval. This method avoids subjective influence, and its advantages are more obvious. Specifically, this retrieval method is mainly divided into several key steps. They are shot segmentation, key frame extraction, feature extraction and similarity measurement. Among which the most important are video feature extraction and video similarity measurement learning. Video feature extraction is
used to realize video processing and analysis, and similarity measurement learning has a vital impact on the final video retrieval results.

2.2. Development of video retrieval
Text-based video retrieval started relatively early, and its history can be traced back to the mid-19th century. After 1950, the information retrieval system was further developed. At present, many well-known online search engines at home and abroad have widely adopted this search method. However, due to the limitations of this retrieval method, the method that combines the image and the text has been researched and applied. The development of this retrieval method started late, and it has gradually developed with the development of Internet technology. It uses the information contained in images and videos to achieve related retrieval. At present, some related companies and research institutions are still conducting research on this and introducing advanced machine learning algorithms.

3. Multi-modal video retrieval scheme

3.1. Definition of background and social identity
In recent years, the number of online videos has shown a sustained and rapid growth. In theory, this makes the time and difficulty of searching videos significantly improved. For this reason, how to search quickly and accurately based on the relevance of videos has become crucial. Services such as video recommendation, video retrieval and video topic discovery have also become important. In fact, for video retrieval, the key is the issue of relevance measurement, which is to obtain retrieval results through the level of similarity. At present, researchers conduct video retrieval research by using elements such as text tags, text similarity, and visual feature similarity. But overall, most of its mainstream research focuses on calculating video similarity. For this reason, some researchers have focused their research on methods of multi-modal fusion of text, video and audio. These methods have achieved significant results in practice, but the social feature information in online videos is still ignored, and further research is needed.

The social feature information of online videos includes information such as video category, video upload time, and video author. This description information also reflects the video to a certain extent and analyzes user behavior based on this information. At the same time, because this information can be reflected from a specific perspective. Therefore, it be used to better measure the correlation between videos and provide ideas for multi-modal video retrieval.

3.2. Text filtering strategy
The video that the user wants to query is actually only a small part of the online video. Therefore, if retrieval is performed, the query time and difficulty will often be doubled due to video redundancy. For this reason, the application of a text filtering strategy is indispensable. This work is mainly to achieve text filtering through reliable keyword strategies. The principle is that only when a keyword appears in the surrounding text information of the two videos, the relevance calculation and screening begin. It has a significant effect on the later retrieval and classification efficiency, and irrelevant online videos will be streamlined on a large scale.

4. Design strategy of multi-modal video retrieval system under deep learning

4.1. System requirements analysis
Multi-modal video retrieval systems based on deep learning need to have video management functions and multi-modal video retrieval functions. Among them, the video management function mainly includes related operations performed by the user on the video, such as playback, pause, volume, and progress adjustment. The multi-modal video retrieval function supports users to enter keywords, open the image information to be retrieved locally, and select a video in different ways to match the videos
to achieve the purpose of retrieval. On the whole, the multi-modal video retrieval system should meet the following characteristics in order to effectively implement these functions. The retrieval results are comprehensive and accurate, and the system can run stably for a long time. The system operation interface is simple and easy to understand and its response time is short and the running speed is fast.

4.2. Hardware and software environment and basic system framework
In the design of multi-modal video retrieval system, there are relatively high requirements for the hardware and software environment. In terms of hardware, the computer of the system should have at least 8G of memory, with an i7 CPU and GTX 1070 graphics card. At the same time, in order to ensure that the system can run safely and stably, the operating system uses 64-bit Windows 10 Professional. In terms of software, MySQL8.0 (fig.1) is used for database design, PyQt4 is used for development, and Python language is used for programming.

![MySQL8.0 Operating interface](image)

Figure. 1 MySQL8.0 Operating interface

In the overall frame structure design of the system, the three input modes of text, image and video should be fully considered to realize the output of the corresponding video. On the whole, the frame structure of the entire system mainly includes offline library construction and online retrieval. Among them, the offline library includes a video characterization module and a database. The main function of the video characterization module is to extract the descriptors of the videos and construct the corresponding index vector. This function is mainly embodied from the following two aspects. One is to extract the static features of the video to obtain the classification label. The second is to input key information into a dual-stream convolutional network to extract video features. The online retrieval part includes several modules of human-computer interaction, data characterization and similarity evaluation to realize video query and retrieval result feedback. This process is mainly realized by the similarity evaluation module’s feature matching and similarity calculation functions.

4.3. Database design
In order to realize the multi-modal retrieval of text, image and video, it is necessary to design two core data tables. They are used to store the information characteristics of the video image and the related information and characteristics of the video respectively. The information in the data table mainly includes serial number, classification label, name, path and feature, etc. The video frame information table usually includes relevant information about the key images extracted from each video. In order to ensure the realization of feature extraction, an index number must be established first, and the video category, name and path must be marked, so that a large amount of data information can be obtained to form a static feature database. The feature extraction is similar and the step of establishing an index
is still adopted, and the feature database is formed by marking the relevant feature information.

4.4. System detailed design and implementation

4.4.1 System interface module design. In order to ensure that users can intuitively understand the functions of the system and facilitate retrieval, the basic principle of system interface design should be concise and generous. Generally speaking, the main interface of the system should display the title bar, menu bar, part of the playback window, and display of search results.

4.4.2 Video management function module design. The video management function module should include basic functions such as retrieving videos, opening local videos and images. These functions should be selected directly through the operation menu. At the same time, the related buttons of play, pause and reload should also support keyboard shortcut control. During video playback, the video playback window should be displayed in a large window, and the video list column on the right side should be listed at the same time.

4.4.3 Design of multi-modal video retrieval function module. As the core module of this system, the multi-modal video retrieval function module includes multiple retrieval based on text, image and video. For this reason, in the system interface, it should include retrieval type button, text box display, retrieval button and close retrieval button. In the search type, it should include three functions, including text search, image search and video search. In the text search function, it needs to try to ensure that the target video can be searched more accurately without entering the complete search keywords.

4.5. System testing

4.5.1 Function testing. Function test includes the video management function module and retrieval function module. In the video management function module test, it mainly includes testing file browsing, video control components and window adjustments. If the function can be completed, the test passes. In the video retrieval function module test, if a keyword enters and the relevant video snapshot is displayed in the retrieval result list box, the test is passed.

4.5.2 Performance testing. The performance test is mainly carried out by Euclidean distance and cosine distance, and the P-R curve (as shown in figure 2) is drawn according to the corresponding data information. If a system bug is found, it must be solved and improved in time.
Figure. 2 PR Curve

5. Conclusion
In the context of the rapid growth of video resources, how to effectively utilize the massive video resources is an issue of constant concern. As a result, the video retrieval function came into being. Due to the diversity of user retrieval requirements for video resources, it is necessary to use multi-modal video retrieval methods and design a multi-modal video retrieval system based on deep learning. The current system still has a lot of room for improvement, which requires further research and development and innovation to create a more complete video retrieval system.

Acknowledgments
Fund Project: Natural Science Foundation of Hainan Province
Project Name: Research on Video Semantic Analysis and Retrieval Technology Based on Multi-modal
Project Number: 618QN255

References
[1] Gao Wenhui. Design and implementation of mobile video retrieval system based on multi-level feature fusion [D]. Beijing University of Posts and Telecommunications.
[2] Dian Yujie, Jin Qin. Multi-modal concept detection related to audio-visual [J]. Computer Research and Development, 56(05): 1071-1081.
[3] Wen Youfu. Research on multi-modal fusion method for network video retrieval[D]. Beijing Jiaotong University.
[4] He Liyuan. Research and implementation of audio and video retrieval system[J]. Digital Media Research, 35(11): 44-46.