www.cad-journal.net
Computer-Aided Design AND APPLICATIONS™

Combined Application of Video Semantic Understanding Technology for Music Video Information Learning

Songhu Liu[1], Qi Yang[2] and Tianzhuo Gong[3]

[1]School of Music, The Chinese University of Hong Kong, Shenzhen, Guangdong 518172, China, songhuliu@163.com
[2]China National Opera & Dance Drama Theater, Beijing 100000, China, yangqi999992022@163.com
[3]Academy of music, Capital Normal University, Beijing 100040, China, gongtz@hrbnu.edu.cn

Corresponding author: Tianzhuo Gong, gongtz@hrbnu.edu.cn

Abstract. Using music practice in computer-aided design can make up for the technical advantages that traditional teaching cannot achieve. It can help teachers and students conduct theoretical research and music creation, integrate teaching resources through modern technological means, and reallocate and effectively utilize music teaching resources. This article mainly studies how to extract effective features from video data with the help of computers to catalog video data. According to the current research situation in this field, the audio features and subtitle features of video are analyzed theoretically. This paper proposes a short video classification and retrieval technology based on computer-assisted image semantic description, and expounds its application in music video information to promote the progress of computer-assisted short video semantic description technology and the development of music video information. The results show that the computer-aided algorithm is more accurate in feature recognition of short video images, 19.99% higher than the comparison algorithm, and can accurately locate the edge contours of short video images.

Keywords: Short video; Semantic features; Music video information; Computer Aided Design
DOI: https://doi.org/10.14733/cadaps.2023.S10.34-44

1 INTRODUCTION

Since its birth in the 1980s, computer music technology has spread rapidly to many fields, such as music therapy, recording industry, film and television industry and education. Computer technology is widely used in the field of education, especially in the field of music teaching, because of its convenient and flexible operation and immediate effectiveness [1]. Through computer assistance, students' sense of experience in traditional music teaching can be improved. It makes the traditional teaching method of orchestration course vivid, enhance the efficiency of students' learning and creation. It saves teaching costs, enhances students' intuitive learning, and improves students' interest in learning [2]. The emergence of Computer Aided Music Teaching (CAM) has brought new opportunities to music teaching, and also made music sense teaching a unique new way. In terms of computer-assisted music teaching, the influence of CAM strategy on education is superior to other educational means [3]. The use of music computer software can improve students' interest in learning and stimulate students' motivation for learning. CAM provides music cognition, skills, feelings and other teaching means because most CAM software provides music symbol related cognitive systems. Harmony analysis of music works Students can use computers to learn music analysis, modulation and other cognitive activities. In addition, listen through the series of tunes, rhythms and chords [4]. This gives students the opportunity to practice tone dictation, rhythm dictation and harmony dictation, thus developing students' ability of musical sense [5].

From the perspective of the industry, the growth of the computer-assisted video market will not stop here. Mobile media carrying computer-assisted short video transmission has occupied the center of people's daily life, and popular culture represented by original short video is still the mainstream of consumer culture market [6]. The rectification is conducive to promoting the demand for high-quality original content in the computer-assisted short video market. In the face of huge video data, from the perspective of users, the traditional video description method of browsing video through fast forward or rewind and other simple operations to obtain the main information in computer-aided video can no longer meet the needs of people to quickly obtain professional knowledge [7]. Through the accumulation of users' rich video resources, it is necessary to better manage user resources under automation technology. Therefore, computer-aided video resource indexing plays a very important role [8]. From the perspective of industry development, computer-assisted music teaching still has a broad audience base and large development space, and popular culture represented by it is still the mainstream of cultural market consumption. In computer-aided video image analysis, there is a correlation between image features, video clips and video semantics. These relationships can reduce the computational cost of semantic detection and improve the search quality. One is the main content of abstract video. The input of such tasks is usually video clips, while the output is one or several natural languages. The other is the dense description of video content, which usually requires a clear description of people, events, scene states, their relationships and change processes in video clips. The correlation between video semantics plays a very important role, and there will be problems of synonymy and polysemy between video semantic content, which is caused by ignoring the correlation of video semantics and will lead to a series of defects. This paper applies deep learning algorithm to semantic feature analysis of video, and combines it with the field of music education.

(1) This paper proposes a clustering method based on implicit semantic analysis. Through singular value decomposition, the word vector and document vector are projected into the low-dimensional space, and the initial center is determined by hierarchical clustering algorithm, and then the results are obtained by clustering.

(2) The model extracts the static and dynamic multi-dimensional features of video sequences through transfer learning, and extracts the semantic information of video key frames through image description algorithm to complete the feature representation of video information.

(3) In this paper, a multimodal information fusion method is proposed. This method describes the key frame of the video in language, and fuses the visual modal information with the language modal information to further enhance the diversity of the language generated by the model.

2 RELATED WORK

On the whole, the research on CAM started late, and the initial research focused on college students, and then gradually extended to primary and secondary students. Individual papers discussed the use of CAM by primary and secondary school students, but only emphasized the importance of application and did not give specific application strategies. There are also individual papers that only discuss the auxiliary music teaching using the multimedia function of a single computer, ignoring the application of various networks.

At present, content-based high-level content video retrieval has become a very important content of the Internet. Because many different video resources are transmitted in large quantities, video retrieval is necessary to manage these contents. Chen et al. [9] constructs a different neural network feature analysis framework. By analyzing the feature points of different video sources, a convolutional neural network is obtained. Zhang et al. [10] combining the advantages of convolution neural network and visual word packet module, a video frame information extraction and representation model is designed. Relevant practitioners need to deeply study and study the knowledge and topics of video semantic relevance, so that relevance can play an important role, and at the same time solve a series of unfavorable problems. Sikandar et al. [11] used convolutional neural networks to extract image features in videos, then used RNN-like methods to encode image features, and finally decoded to generate natural language descriptions of video content. However, the visual features extracted by this method are relatively single, and the language description of the video content is not rich enough. Jong-Chih et al. [12] proposed a recursive encoder combined with an attention model of video keyframes using a deep convolutional neural network pre-trained on Imagenet, and then input it for encoding in time series. introduces a temporal attention mechanism to build a global temporal structure, which enhances the appearance features by encoding the action features of the local temporal structure. Yang and Yu [13] proposed to extract static, dynamic and semantic information from videos by using models pre-trained in multiple source domains to improve the accuracy of video language description. Sun et al. [14] proposed a hierarchical RNN framework for describing long videos with multiplicity. The so-called hierarchical framework means that the framework includes two generators, one applies a recurrent neural network to generate sentences, and then uses the generated sentences as the input of another recurrent neural network, thus generating the entire paragraph. Extracted a new "consistent wire bundle" image feature, which classified the wire bundle by extracting the color, direction and spatial features of the line segment, and discriminated whether the image belonged to the category of building images. proposed a video description method based on multimodal fusion, which extracts dynamic features and static visual features in videos, and fuses audio features to generate language descriptions. However, for a single frame image in a video, the background and semantic information in the scene are not fully considered. Ran proposed a method to extract video keyframes to improve the accuracy of description language. However, this method also does not consider the multi-dimensional information of the video, such as object, background and space-time. Shen et al. proposed to pre-set the lexical and grammatical rules for generating sentences, and pre-define the visual categories of elements such as subject, predicate, and object, and map visual semantics to templates when the corresponding visual targets are detected.

3 METHODOLOGY

3.1 Short Video Feature Extraction Model

The features extracted by the image classification and retrieval system based on computer-aided design are actually potential visual features. However, because they are extracted from image databases with specific known semantic categories, these features already have semantic information. On this premise, it can be called the semantic feature of the image, thus realizing the mapping from the low-level feature to the high-level semantic. Since shots are the basic unit of video, most existing video classification methods use video segmentation technology to divide continuous video streams into shots with specific semantics as the basic unit of classification. The use of key frames can greatly reduce the amount of video data and provide an organizational framework for searching and retrieving video clips. Because the key frame can summarize the content of the video, the determination of the key frame can establish a video summary of the video clip, allowing users to quickly browse the entire clip by watching several limited key frames. The structure of short video depth learning is shown in Figure 1.

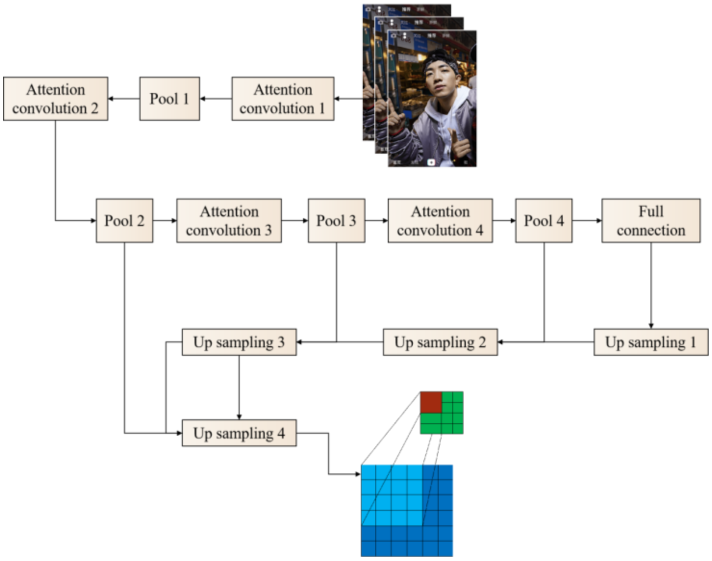

Figure 1: Short video image stitching detection model.

In the task of video natural language description generation, the representation and extraction of video features is the first step and an extremely critical link, which plays a vital role in the output results of subsequent natural language models. Yes, color is a color feature based on histogram. The input short video signal $I(X,t)$ is compared with N distribution models, and then the matching model is updated. If:

$$\left|I_j(X,t)-\mu_{ij}(X,t)\right|<\tau D_{ij}(X,t) \tag{3.1}$$

The $I(X,t)$ matches the model p_i . τ is a global threshold, i represents the i -th distribution model, similarity distance $d_i(X,t)$ is updated. $d_i(X,t)$ is defined as:

$$d_i(X,t)=\sum_{j=s,r,g}\frac{\left|I_j(X,t)-\mu_{ij}(X,t)\right|D_{ij}(X,t)}{h_{ij}(X,t)} \tag{3.2}$$

Update matching p_i as follows:

$$\mu_{ij}(X,t+1)=(1-\alpha)\mu_{ij}(X,t)\alpha I(X,t) \tag{3.3}$$

$$D_{ij}(X,t+1)=\min\left\{\left[(1-\beta)D_{ij}^2(X,t)+\beta\left(I(X,t)-\mu_{ij}(X,t)\right)_2\right]^{1/2},D_{\max}\right\} \tag{3.4}$$

If the color distribution in a picture conforms to a certain probability distribution, the moment of color can be used as a feature to distinguish different color distributions. Compared with color histogram, another advantage of this method is that there is no need to vectorize features, thus speeding up the processing speed. The three lower moments of color are mathematically expressed as:

$$\mu=\frac{1}{M\times N}\sum_{i=1,j=1}^{i=M,j=N}p_{i,j} \tag{3.5}$$

$$\sigma = \left(\frac{1}{M \times N} \sum_{i=1,j=1}^{i=M,j=N} \left(p_{i,j} - \mu \right)^2 \right)^{\frac{1}{2}} \tag{3.6}$$

$$S = \left(\frac{1}{M \times N} \sum_{i=1,j=1}^{i=M,j=N} \left(p_{i,j} - \mu \right)^3 \right)^{\frac{1}{3}} \tag{3.7}$$

Among them, $p_{i,j}$ is the color component value of the pixel located at the coordinate (i, j) in the image, and M, N are the length and width pixels of the image, respectively.

3.2 Short Video Semantic Description Under Computer-Aided Design

Most of the short videos from media platforms are people and things that users share independently. They are simple in logic and more like graphic content. These short videos are different from the previous video research objects, because they have no accurate behavior range or role target, so they cannot only analyze human behavior, and the general feature extraction algorithm cannot fully represent all the information of these short videos. Because of the freedom and convenience of sharing, a large number of short videos are disseminated on the Internet every day, which urgently needs the assistance and management of intelligent analysis technology. Therefore, when studying the natural language description generation technology of short video from the media platform, it is necessary to consider the above features and design the model pertinently. The driving model of the short video image data mining function is shown in Figure 2.

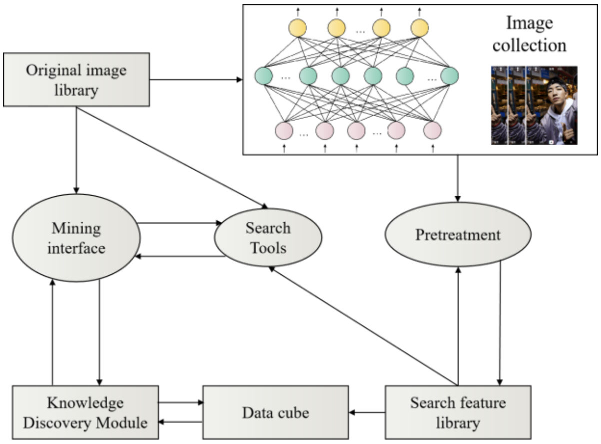

Figure 2: Function-driven model of short video image data mining.

The feature extraction model needs to ensure that the dimensions of various features are consistent, such as a two-dimensional matrix represented by $N \times F$. Suppose M_i is the $i(i \in [1, m])$-th modal feature, and the weighted weight vector is:

$$W = (w_1, w_2, \ldots, w_m) \tag{3.8}$$

And satisfy the sum of all weights equal to 1, that is:

$$\sum_{i=1}^{m} w_i = 1 \tag{3.9}$$

Then the fusion result obtained by the weighted summation of m features is:

$$M_{fusion} = WM^T = \left(w_1M_1 + w_2M_2 + \ldots w_mM_m\right) \tag{3.10}$$

Under the background of multi-core mapping, the high-dimensional space is decomposed into the combination of several low-dimensional feature spaces. Each feature space gives full play to its basic kernel feature mapping ability. Because these mappings are different, heterogeneous data with different feature components can be processed by corresponding kernel functions. The semantic feature analysis process of short video based on deep learning is shown in Figure 3.

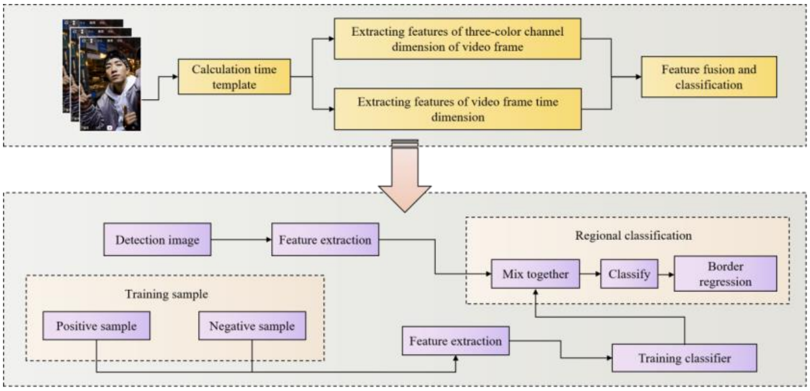

Figure 3: Analysis process of semantic features of short video.

4 RESULT ANALYSIS AND DISCUSSION

4.1 Comparative Experiment of Video Classification and Retrieval Based on Computer Aided Design

Through audio classification, on the one hand, different types of voice signals can be processed in different ways, thus reducing the search space for further processing. Music elements can be analyzed in rhythm, notes and instruments. On the other hand, the results of audio classification, especially the extraction of silent segments, are of great significance for understanding the structure of video and scene segmentation. Therefore, audio signal classification technology is an important preprocessing technology in audio signal processing technology. The convergence index is used to compare the maximum clustering structure mining results of the method in the literature and the method in this paper. The convergence of training loss is shown in Figure 4.

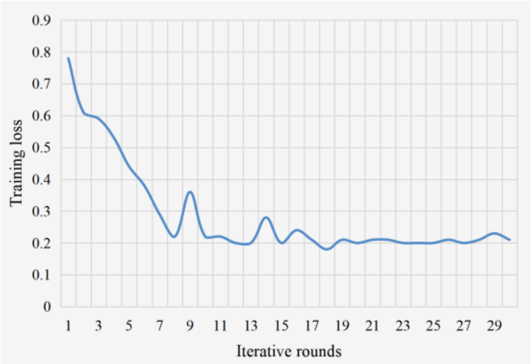

Figure 4: Convergence comparison results.

When using a video frame as a query unit, each video frame needs to be processed and matched, so the amount of computation is very large, which will lead to reduced efficiency. When a shot or key frame is used as a query unit, it is not only necessary to detect the shot boundary of the query video, but also to extract the key frame and features of the shot, and then match and query its features. The comparison of the average absolute error of the algorithm is shown in Figure 5.

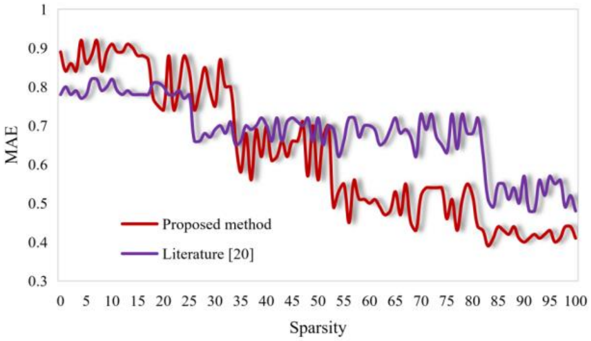

Figure 5: Comparison of the mean absolute error of the algorithms.

Compared with the method in literature, the method in this paper has obvious advantages in the later stage of operation, and the error is reduced by 40.66%. For video sequences, the audio signal is suitable to express descriptive semantics, and the visual signal is suitable to express mandatory semantics. Only by integrating vision and hearing can a complete and rich semantic information be expressed, and the separation of the two will make the complete semantic information lost. The accuracy of short video image feature recognition is taken as the test index, and the methods in literature. Table 1, Table 2 and Table 3 respectively describe the treatment results.

Sample size	*Accuracy of short video image feature recognition (%)*
15	98.65
30	98.28
45	97.69
60	97.74
75	96.82
90	96.28
105	96.07

Table 1: Methods in this paper.

Sample size	*Accuracy of short video image feature recognition (%)*
15	94.88
30	94.49
45	93.67
60	92.98
75	92.59
90	92.37
105	91.74

Table 2: Literature method.

Sample size	*Accuracy of short video image feature recognition (%)*

15	96.68
30	95.89
45	95.64
60	94.38
75	93.77
90	93.26
105	93.17

Table 3: Literature method.

The accuracy of short video image feature recognition by all methods will decrease. However, compared with the other two methods, the short video image feature recognition accuracy of this method is obviously higher. Figure 6 is an example of a video. There are 520 frames. Including one shot abrupt change at 256 frames. There is one transformation, at 42 -58 frames. For the continuous transformation around 200 frames and 420 frames, the translation of the lens leads to the continuous transformation.

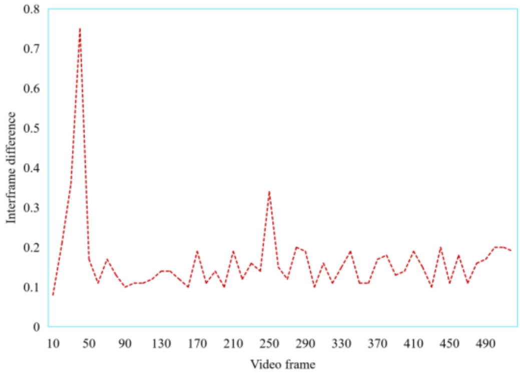

Figure 6: Video example frame difference.

Video content analysis under computer-aided design mainly studies the relationship between feature description and high-level semantic concepts. Its ultimate goal is to automatically extract video semantic concepts from various features and related original video data. Video is a kind of mixed media with multiple modal characteristics, and single modal characteristics can only reflect local and horizontal information. The test sample tested by the human motion recognition model using the short video description method in this paper is shown in Figure 7 and Figure 8.

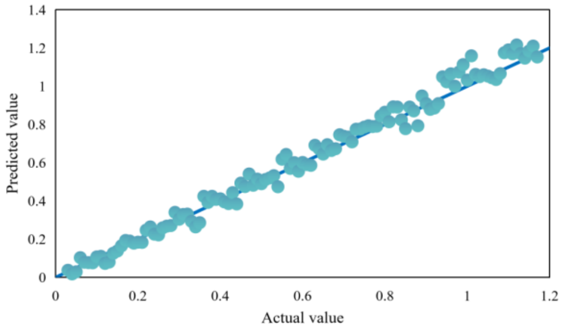

Figure 7: The scatter plot of the actual value and the predicted value of the literature.

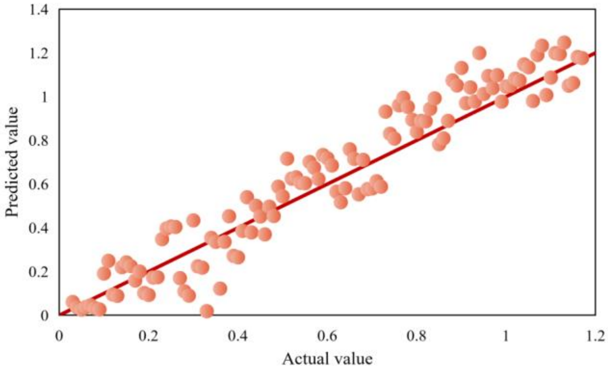

Figure 8: Scatter plot of actual value and predicted value of the short video description method in this paper.

It can be analyzed that the human motion recognition model based on the short video description method in this paper is better than that in literature in both accuracy and efficiency. A large quantity of data sets contains different kinds of data information, in which one kind of information or a different observation angle of data can be called a mode, and multiple modes form a multi-mode. In video processing applications, a large amount of video data in network resources have great complexity and diversity, so it is difficult to achieve ideal results if only a single mode of video processing is used. Therefore, it is need to use multimodal methods to process the appearance features and motion features in video data. Compare the accuracy of the algorithm in short video image feature recognition, as shown in Figure 9.

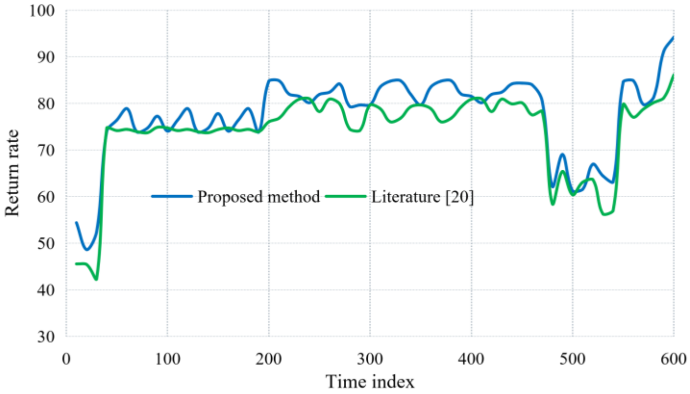

Figure 9: Accuracy comparison of short video image feature recognition.

The detection results show that this algorithm is more accurate in feature recognition of short video images, which is 19.99% higher than the comparison algorithm, and can accurately locate the edge contour of short video images. After the overall framework design of the video description system, as well as the design of each sub-module and its corresponding neural network structure, it is need to build a framework on the system platform and write a program to realize the system. After the realization, it is also need to train the model by using training data sets, so that the model can keep learning and update the network parameters, thus obtaining a mature description system that can directly input and output correct video descriptions.

4.2 Development of a Concert Video Retrieval System Prototype Based on Computer Aided Design

The research of semantic based concert video retrieval involves many fields such as computer vision, machine learning, pattern recognition, and human-computer interaction. Currently, video retrieval systems are not yet mature. There is no fully efficient universal video database retrieval system, except for a specific field of specialization. Video has its own characteristics compared to images. Video has a large hierarchy, structure, and complexity, so it is necessary to process video files hierarchically. The design of query systems in video retrieval is much more complex than character-based databases. Concert video database features should have the following characteristics:

(1) The system should have the general functions of a general multimedia data warehouse management system, and also serve as a dedicated system for managing music and video, with various dedicated functions.

(2) The system should have data sharing to reduce the redundancy of video data.

(3) The system should be extensible, allowing users to add special application modules to the system.

(4) The system should be built on top of the existing database management system and have relative independence, and can be connected to other current general-purpose databases.

In addition to the system's database, the design of the video retrieval interface is also important. It needs to follow certain human-computer interaction principles, such as when a user submits a search request, the user has only a vague understanding of their search objectives. Therefore, user retrieval interfaces are used to help users understand and express retrieval needs. At the same time, a good search interface is also used to help users further define their search needs, select appropriate results from the obtained searches, and initiate new searches. Therefore, the retrieval interface must have feedback, interactivity, and visualization, otherwise it will have no practical application value. Moreover, different user interfaces must be targeted, but data consistency must be maintained. The task of the interface is divided into input modules to provide valuable feedback. Establish a link between the user's search requirements and the results obtained, and establish a link between the feedback search results themselves. Allow to undo previous results. If the user is not satisfied with the results of the last retrieval, they can have the right to undo and retain the intermediate results from the retrieval.

Semantic concept-based retrieval models essentially support customized OSQL query languages. The language is easily converted from the user's natural language, is simple and convenient, and conforms to the high-level semantics of the human brain, and is also suitable for computer reading. The prototype system for concert video retrieval can meet the requirements of small batch ontology based semantic video retrieval. Three retrieval modes are supported simultaneously. They are: native semantic tree based, semantic concept based, and keyword-based retrieval methods. The core is semantic video retrieval, which can fully exploit the advantages of various retrieval modes. We call it concert video retrieval based on multimodal retrieval. The system can simultaneously support any combination of three retrieval modes.

5 CONCLUSIONS

Video data contains rich semantic scenarios, and advanced semantic information is closer to people's thinking and understanding. Therefore, more semantic content can be extracted to make the search results more responsive to user needs. Organize and classify network video data effectively through computer-aided design retrieval. This paper proposes a short video classification and retrieval technology based on computer-aided design (CAD) image semantic description, and describes its application in music video analysis. Promote the progress of short video semantic description technology and the development of music video education. This algorithm is more accurate in feature recognition of short video images, 19.99% higher than the comparison algorithm, and can accurately locate the edge contours of short video images. Compared with traditional methods, this method has significant advantages in later operations, reducing errors by 40.66%, and is conducive to sharing high-quality music video semantic analysis course resources.

Songhu Liu, https://orcid.org/0009-0008-0027-8287
Qi Yang, https://orcid.org/0009-0002-6997-8915
Tianzhuo Gong, https://orcid.org/0000-0001-8551-4657

REFERENCES

[1] Tran, M.-T.; Lee, G.-S.: Staff-line Removal for Music Score Images using U-net, KIISE Transactions on Computing Practices, 26(1), 2020, 35-47. http://doi.org/10.5626/KTCP.2020.26.1.26

[2] Qin, L.; Kang L.: Application of Video Scene Semantic Recognition Technology in Smart Video, Tehnicki Vjesnik, 25(5), 2018, 1429-1436. https://doi.org/10.17559/TV-20180620082101

[3] Aakur, S.-N.; Souza, F. D.-D.; Sarkar, S.: Generating open world descriptions of video using common sense knowledge in a pattern theory framework, Quarterly of Applied Mathematics, 77(2), 2019, 1. https://doi.org/10.1090/qam/1530

[4] Castellanos, F.-J.; Gallego, A.-J.; Calvozaragoza, J.: Automatic scale estimation for music score images, Expert Systems with Applications, 158(38), 2020, 113590-113612. https://doi.org/10.1016/j.eswa.2020.113590

[5] Zhao, Z.; Wu, Q.: Computer-Aided Recognition and Analysis of Abnormal Behavior in Video, Computer-Aided Design and Applications, 18(S3), 2020, 34-45. https://doi.org/10.14733/cadaps.2021.S3.34-45

[6] Lee, S.; Kim, I.: Video captioning with visual and semantic features, Journal of Information Processing Systems, 14(6), 2018, 1318-1330. https://doi.org/10.3745/JIPS.02.0098

[7] Lv, W.; Ji, S.: Atmospheric environmental quality assessment method based on analytic hierarchy process, Discrete & Continuous Dynamical Systems - S, 12(4), 2018, 941-955. http://doi.org/10.3934/dcdss.2019063

[8] Liao, W.-C.; Hsieh, J.-C.; Wang, C.-M.: Compressed sensing spectral domain optical coherence tomography with a hardware sparse-sampled camera, Optics letters, 44(12), 2019, 2955-2958. http://doi.org/10.1364/OL.44.002955

[9] Chen, H.; Hu, C.; Lee, F.: A Supervised Video Hashing Method Based on a Deep 3D Convolutional Neural Network for Large-Scale Video Retrieval, Sensors, 21(9), 2021, 3094. http://doi.org/10.3390/s21093094

[10] Zhang, C.; Lin, Y.; Zhu, L.: CNN-VWII: An efficient approach for large-scale video retrieval by image queries, Pattern Recognition Letters, 123(5), 2019, 82-88. https://doi.org/10.1016/j.patrec.2019.03.015

[11] Sikandar, T.; Rabbi, M.-F.; Ghazali, K.-H.: Using a Deep Learning Method and Data from Two-Dimensional (2D) Marker-Less Video-Based Images for Walking Speed Classification, Sensors, 21(8), 2021, 2836. http://doi.org/10.3390/s21082836

[12] Han, L.; Ge, Z.: Design of psychology experiment teaching system based on CAD virtual reality technology, Computer-Aided Design and Applications, 20(S1), 2023, 76-85. https://doi.org/10.14733/cadaps.2023.s1.76-85

[13] Yang, H.; Yu, N. Y.: A fast algorithm for joint sparse signal recovery in 1-bit compressed sensing, AEU - International Journal of Electronics and Communications, 138(4), 2021, 153856. https://doi.org/10.1016/j.aeue.2021.153856

[14] Sun, X.; Long, X.; He, D.: VSRNet: End-to-End Video Segment Retrieval with Text Query, Pattern Recognition, 119(4), 2021, 108027. http://doi.org/10.1016/j.patcog.2021.108027