Multimodal Information Processing Method of College English Course Online Education System

Baoling Feng(✉) and Linan Wang

Software Engineering Institute of Guangzhou, Guangzhou 510420, China
fbl5871@sina.com

Abstract. In order to process the information in the online education system more efficiently, this study designs a multi-modal information processing method for the online education system of college English courses. This study mainly deals with video information, image text information and audio information in the system. Firstly, based on structured processing, the video data stream is divided into organic whole with certain logical structure, and the visual feature information and visual invariant feature are extracted. The multi-branch convolutional neural network is designed and the text features are extracted. Convolutional neural network is used to extract audio features from the system. Finally, a functional model of multi-modal information fusion is designed to realize the fusion processing of multi-modal information. Experimental results show that this method has high data fusion efficiency and timeliness.

Keywords: College English courses · Online education system · Multimodal information processing · Structured processing

1 Introduction

With the further advancement of social development and reform, the construction of my country's college English curriculum system is facing major changes and transformations, and higher requirements have been put forward for the basic requirements of the curriculum, teaching content and teaching methods. At the same time, the need for the integration of computer network promotion and college English curriculum also makes the current college English curriculum face great challenges. How to realize the dynamic integration of information technology and English courses under the network environment, build a personalized and diversified college English ecological curriculum system, and further promote the improvement of the curriculum system, is an urgent problem to be solved at present.

The traditional college English teaching method is mainly based on teachers' explanation, students can only passively receive knowledge, and the classroom atmosphere is dull, the form is not novel enough, the students' learning motivation and interest are

W. Fu and G. Sun (Eds.): eLEOT 2022, LNICST 453, pp. 378–390, 2022.
https://doi.org/10.1007/978-3-031-21161-4_29

not enough, and the teacher-student interaction is not active enough. With the development of information technology and the general opening of online courses, the society's demand for talents with comprehensive practical ability is increasing, which also brings new challenges to college English courses [1].

The online education system of college English courses has gradually been promoted. This teaching mode conforms to the reform needs and can significantly improve the English learning efficiency of college students. Online courses have the advantage that they can be played repeatedly and are not limited by time and space. Students can study independently anytime and anywhere, which is very convenient. At the same time, online courses also put forward strong requirements for students to learn and control themselves. How to promote and innovate the construction of the online education system of college English courses has become a problem that the college English teaching system and English teachers should think about [2].

The development of network and computer technology has broken the space-time boundary of education and improved people's learning efficiency, learning initiative and learning interest. The introduction of information technology into College English teaching and the construction of College English course online education system make college English teaching a dynamic and open model. It can meet the individual needs of different majors and different college students, make the elements of College English curriculum interact and depend on each other, and change at any time according to the changes of the environment, so as to promote a dynamically balanced College English curriculum online education system. In the construction of College English course online education system, we must combine college English course with modern network information technology, create a natural and Harmonious English teaching environment, and transform modern information technology from auxiliary teaching means to leading teaching means.

In today's higher education, increasing the interaction between teachers and students through the promotion and application of online courses can improve students' learning initiative. However, in the process of curriculum construction, appropriate rules should be followed, through comprehensive consideration and mutual compatibility of various resources, to ensure unified standards in the implementation of the curriculum, and take into account restrictive factors such as attitudes, beliefs and skills towards students and teachers. The construction and application of College English online education system is conducive to the realization of collaborative learning. It is a new online interactive teaching mode combining teaching with learning. Therefore, the construction and promotion of College English course online education system is the inevitable development trend of College English education.

Based on the above analysis, this paper studies the multi-modal information processing method of college English course online education system. The specific research ideas are as follows:

(a) The multi-modal information in college English course online education system is divided into three categories: video information, image text information and audio information.

(b) For video information, based on structured processing, the video data stream is divided into organic whole with certain logical structure, and the visual feature information and visual invariant feature are extracted.
(c) Aiming at the image text information, the text multi-branch convolutional neural network is designed by using the multi-branch mode, and the image text features in the system are extracted.
(d) For audio information, convolutional neural network is used to extract audio features in the system.
(e) Design multi-modal information fusion function model to realize the fusion processing of multi-modal information.

2 Method Design

2.1 Video Visual Feature Extraction

Structured Processing

The feature extraction of video information is implemented for the online education system of college English courses. The video data is generally stored in the disk with unordered binary code. In fact, it is difficult to find more in-depth and valuable video information from this level. Therefore, it is necessary to divide the data stream of the video itself into an organic whole with a certain logical structure.

The general structural model is mainly hierarchical and upside-down tree structure. The most important logical order from top to bottom is the original video information, the video scene, and the key frames of the shot. The main technologies in the corresponding structured processing include shot segmentation, scene recognition and key frame selection [3]. Because the video itself has the characteristics of disorder, this disorder belongs to the unstructured data stream, but if the content analysis is carried out directly, the workload of the analysis will be greatly increased. Video materials with storylines and specific performance content generally take the shooting method of sub-shots and scenes during the shooting process. Therefore, the seemingly unstructured data has a strong causal relationship in its internal logical sequence.

Video structured processing is mainly to describe the video stream at all levels, so as to realize the low-level feature analysis of video and the establishment of semantic recognition model. In the research of multimedia such as video materials, the content of video itself is very rich, but the structure of video is often very complex. Nowadays, the general practice is to structurize the data first. Video structured processing adopts a hierarchical way to represent the content data of video, which greatly reduces the complexity of video [4]. Generally speaking, the video structure is structured according to the top-down logical relationship from the four levels of video, scene, shot and image frame of each content. The video hierarchy is shown in Fig. 1.

Based on video feature analysis, representative information abstracts are extracted at different levels. In the structured video processing, shot segmentation, shot key frame

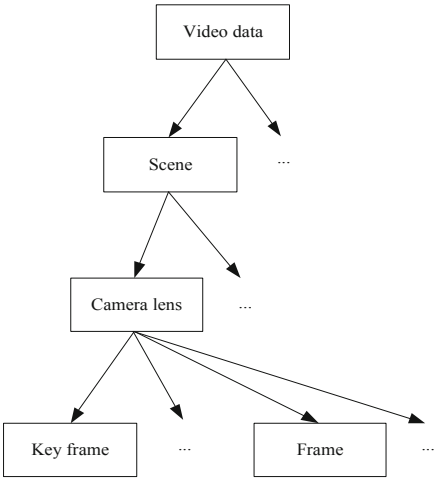

Fig. 1. Video hierarchy

extraction, scene segmentation and key frame extraction are the most important and important tasks in this stage.

Visual Feature Extraction

Visual feature extraction is the basis of video information representation, analysis and application [5]. In general, both images and videos contain low-level visual features and high-level semantic features. The semantic gap has always existed between low-level visual features $p + q$ and high-level semantic features. Since the problem of semantic gap cannot be well solved, so as to realize the automatic analysis and understanding of high-level semantics of video, we still use low-level visual features in our daily life.

The extraction of low-level visual features is to obtain a series of digital features that can characterize the content by analyzing the attributes of image pixels and the relationship between pixels, specifically including color features, texture features, shape features and motion features.

Color feature: color feature is the most widely used visual feature in image and video retrieval, mainly because of the strong correlation between color and the object or scene contained in the image. In addition, compared with other features, color features are easy to calculate, less dependent on the size, direction and angle of view of the image itself, and have better robustness. The color histogram is used to calculate its characteristics.

The representation method of the color histogram feature is simple, and it is not sensitive to changes in the size and rotation of the image; only the brightness change has a greater impact on the resulting results. In short, the color histogram is a 1-D discrete function used to count the color values of all pixels in an image, defined as follows:

$$h(i) = \frac{m_i}{N^2} \tag{1}$$

In formula (1), m_i is the number of pixels whose color value is i in the image, and N is the total number of pixels. By calculating formula (1), a K-dimensional feature vector

can be obtained to represent the color histogram. Since there are different color models or spaces used to represent colors, color histograms can be calculated based on different color value spaces. Among them, RGB color space is the most common color space [6].

Shape feature is an important means of image content expression. It is usually associated with the target and is the main information used to identify the target. Human beings mainly recognize objects through shape. Shape features have certain semantic information to some extent, which makes them superior to other visual features such as color and texture in some aspects. Shape features can be represented by region features.

The regional feature of the image is related to the entire shape area, and the contour feature of the image is used to the outer boundary of the object, so the shape extraction methods are roughly divided into two categories: one is to first segment the image, so as to obtain a series of and then treat each area or several areas as an object, and the extracted shape can be represented by the area itself or its boundary curve; the other is to extract the edge points of the image first, and then connect these edge points, a series of curves are then obtained, and the extracted shape is represented by these curves. Shape-independent moments are typical representatives of region-based features. Assuming that R is an object represented by a binary image, the $p + q$ order central moment of shape R is:

$$\mu_{pq} = \sum_{R} \left[(x - x_c)^q (y - y_c)^p \right]^2 \tag{2}$$

In formula (2), (x_c, y_c) is the center of the object. By normalizing this central moment, scale-independent shape properties are obtained.

The shape representation and description methods also include rotation function method, finite element method and wavelet description method. The rotation function can be used to compare the similarity between concave and convex polygons. In order to describe how each point of the object is connected to other points, the finite element method defines a stiffness matrix. Wavelet descriptors can be used to describe the shape of objects and represent many characteristics of shapes, such as uniqueness, stability, spatial location and so on.

Motion is the most significant feature that distinguishes video from static images. It reflects the development and change of video content over time. Compared with other visual features such as color, texture and shape, the visibility of motion characteristics is low, which is often hidden in the changes of other visual features, and can only be extracted by using specific methods. Reflecting the motion information in video through motion estimation algorithm is widely used in current motion analysis [7]. From the space of motion estimation, it can be divided into two categories: two-dimensional motion estimation and three-dimensional motion estimation, of which two-dimensional motion estimation is used more in video. Commonly used motion estimation algorithms include block based motion estimation, optical flow algorithm and parameter model-based motion estimation.

According to the cause of the motion information, the motion information generated by the video object and the motion information formed by the movement of the camera are usually processed separately, which are called local motion and global motion respectively. Among them, the calculation method of local motion is more complicated,

and the commonly used method is parameterized global motion estimation and motion compensation to obtain the motion information of foreground objects [8]. In MPEG-7, four motion descriptors are specially defined in order to describe the motion characteristics in the video: namely, the moving object trajectory, the object parameter motion, the camera motion and the motion activity descriptor. The moving object trajectory and the object parameter motion descriptor both need Based on the detection of moving objects in the video, they belong to the local motion information descriptor; and the camera motion descriptor is used to describe the global motion information.

The extraction of visual invariant features mainly revolves around the extraction and calculation of local descriptors. The method used is a distribution-based descriptor. Distribution-based descriptors use histograms to represent different colors or representations. Scale-invariant feature transforms are the most famous among them.

2.2 Image Text Feature Extraction

Using the multi branch mode, a text multi branch convolutional neural network is designed to extract text features. The network parameters are shown in Table 1.

Table 1. Main parameters of text multi-branch convolutional neural network

Serial number	Network layer	Convolution kernel size/stride	Number of channels
1	Conv1	3 × 3/1	128
2	Branch1 (max+mean-pool)	6 × 6/2	128
3	Conv2	3 × 3/1	256
4	Branch2 (max+mean-pool)	6 × 6/2	256
5	Conv3	3 × 3/1	512
6	Branch3 (max+mean-pool)	6 × 6/2	512
7	Branch4 (max+mean-pool)	6 × 6/2	1024
8	ConvS	3 × 3/1	1024
9	Branch5 (max+mean-pool)	6 × 6/2	1024

Image text feature extraction is realized by text multi branch convolution neural network.

2.3 Audio Feature Extraction

Extracting audio features of an online education system for college English courses using CNN.

CNN is a structure of deep neural network, which can be regarded as a locally connected network. Compared with the fully connected network, its biggest features are: local connectivity and weight sharing. That is to say, for a point P in a two-dimensional

grid, the closer the point is to the point P, the greater the influence on it; according to the statistical characteristics of natural images, a local weight can also be used for Another partial analysis. The weight sharing here is the sharing of convolution kernel parameters. Two convolution kernels with unequal weights can extract two kinds of features on a two-dimensional grid.

First, the spectrogram is calculated, and the spectrogram is selected as the object of our analysis. The following is the calculation process of the spectrogram: first, the frame-by-frame windowing process is performed. The window function here selects the hamming window. During the frame-by-frame process, the frame stack is set to half the window length. In this way, the original timing signal $x(n)$ can be represented as $(m)nx$, wherein the variable n represents the frame indication sequence number, and the variable m represents the time indication sequence number in the corresponding frame. Then discrete Fourier transform is performed to obtain short-term amplitude spectrum data. The spectrogram can be obtained by expressing the value of the short-term amplitude spectral data as a two-dimensional image composed of gray levels, and the spectrogram can be represented by transformation [9]. It can be seen that although the spectrogram is a two-dimensional image, it represents three-dimensional information. The shades of color on the image indicate the amount of energy at the corresponding time and frequency. The spectrogram is selected as the object of our analysis because the spectrogram reflects information in two directions, but this also requires us to have a corresponding method with three-dimensional information analysis capabilities. The convolutional neural network has its unique The properties are exactly what we want. The following is the theoretical basis for CNN to perform deep feature analysis on spectrograms.

Step 1: Local features are extracted by local perception. The object of ordinary neural network is the whole input variable every time. When we analyze in the spectrogram, we need to pay attention to the local features, so the object of our analysis can only be a part of the spectrogram. The convolution operation of convolution neural network is just carried out through the convolution filter and the convolution operation of a small area in the spectrogram. Specifically, the convolution filter moves forward with overlapping in the time domain and frequency domain. The object of each analysis is the local information of the subband with a certain central frequency in a long time period. It is equivalent to time-frequency filtering, which will retain the information of frequency direction in a long period of time. At the same time, it also has the ability to enhance local features and reduce noise.

Step 2: Downsampling. Downsampling is performed by efficient decimation and retention operations on smaller local information, which makes the CNN more robust to the displacement phenomenon of the payload in the signal. For example, if the sound segment corresponding to the basketball hitting the ground in the stadium scene spectrogram appears whenever it appears, the corresponding scene category is the basketball court, and whether it appears sooner or later is not so important. This confusion position processing method has strong anti-distortion ability to the phenomenon of signal deformation or distortion. In addition, the downsampling operation can effectively reduce the interference of other irrelevant information while retaining useful information. For example, the speech segment of the basketball hitting the ground interval in the stadium

scene will be gradually ignored during the downsampling process. From the above analysis, we can see that in the process of convolution and downsampling, CNN not only retains long-term structural features containing frequency information, but also reduces the influence of noise. The position produces a certain anti-distortion ability. Compared with the previous short-term features and long-term statistical feature values, this long-term structural feature can better reflect the relatively long effective content of the audio scene sound segment, and thus can better perform audio scene recognition [10].

A very important part of CNN for audio scene feature extraction is the design of convolution filter size. We have mentioned above that the convolution filter is used for long-term local feature analysis, so the size of the convolution filter in the time domain direction is generally designed to be equivalent to the duration of the effective content, and the size in the frequency axis direction is generally designed to be frequency. Subband width size. In this way, the elements in the feature map obtained by the convolution operation correspond to the features of the long-term subband. We evenly divide the entire frequency band into 6 sub-bands, and initially set the size of the convolution filter in the direction of the frequency axis to be the width of each sub-band. Through our human ear analysis, the duration of effective content in most audio scene corpora is about 1/8 of the duration of the entire speech segment, so we set the size of the convolution filter in the time domain direction to 1/8 of the entire duration.

The design principle of convolution filter and its size in the experimental system have been described above. Next is the audio scene feature extraction and the final audio scene classification process. The training and classification of convolutional neural network are carried out at the same time, that is, it is an end-to-end generation model. The advantage of this is that it can adjust adaptively through self-learning and find the best classification features by active learning, rather than relying too much on subjective experience. The following is the process of feature extraction and final classification:

1) The audio signal is preprocessed to obtain the time-frequency joint analysis data;
2) The initial characteristic matrix is obtained by convolution operation of convolution layer;
3) The final characteristic matrix is obtained through the down sampling operation of the down sampling layer;
4) Train the final classifier, that is, the weight parameters between the classification layer and the output layer, and classify the final audio scene.

In the training process, we use the classical optimization algorithm batch gradient descent method to update the parameters. If the parameter of learning rate is difficult to adjust, we can also try to use optimization algorithms such as Newton method or quasi Newton method. In the whole process, the first step is to train the final classification layer parameters. First, take the difference between the output scene category and the real scene category as the value of the loss function, and then calculate the gradient of the loss function with respect to the weight matrix of the full connection layer; The second step is to train the parameters of the lower sampling layer, that is, update according to the recursive formula of interlayer gradient; Finally, the parameter training of convolution layer is carried out, which requires an up sampling operation; All parameters of the network are updated once, which is recorded as an iteration. After several iterations, the

finally convergent convolutional neural network model is obtained. The feature extraction of audio scene is carried out on the above convergent network model, then the category prediction is carried out, and finally the final audio scene recognition rate is obtained.

2.4 Feature Fusion

Design a multimodal information fusion function model, implement multimodal information fusion, and realize multimodal information processing in the online education system of college English courses.

The multimodal information fusion functional model serves to achieve the basic goal of fusion, so more specifically, the multimodal information fusion functional model needs to meet the following functional requirements: First, to provide users with complete and comprehensive information in a specific field, that is, there is no information loss; the second is to provide users with real and objective information, that is, there is no information distortion; the third is to provide multi-level relevant information, that is, to meet the needs of different user groups; the fourth is the interactive function, which can integrate the results with users. The requirements are matched and fed back multiple times to optimize the results.

The multimodal information fusion function model mainly includes six parts: acquisition module, preprocessing module, fusion module, information center, service module human-computer interaction interface and management interface for coordinating system functions.

The functions of each module are as follows:

Acquisition module: This module is the input of the system and undertakes the collection of external information. The module is composed of two parts. “Demand collection” is to collect and integrate the user’s information needs and determine the information subject, “information collection” is to obtain the multi-modal information content of a specific subject. The starting point and foothold of multimodal information fusion processing are to serve users. Multimodal information can reflect its deep information utility only based on users’ specific information needs. Therefore, the collection of users’ information needs is the focus of the module and even the whole system. This part of information can be called “prior knowledge”, which provides prediction or limits the search scope for the data association of specific observations. The information collection part collects the corresponding multi-modal information based on this. Considering the subject consistency and modal diversity of object information, two aspects need to be paid attention to in the collection process: one is the matching and screening of information subjects; The second is the selection of corresponding extraction methods for different modal information.

Preprocessing module: This module preprocesses the information collected in the early stage to prepare for the next fusion work. For the user’s information needs, according to the standardized task attribute description, analyze the user’s cognitive network or establish a cognitive model. Based on this standard, the collected multimodal information is refined, conflict resolution and classification. Finally, the preprocessing results are stored in the fusion object database.

Fusion module: this module is the core part of the whole system, and its function is to fuse multi-modal information at different levels. According to the relevant theories

of information fusion level division and the needs of different types of users, the fusion module is divided into three levels: data fusion, feature fusion and decision fusion. After using specific fusion methods and technologies for information fusion, the fusion results of the three levels are stored in the corresponding database.

Information Center: it mainly includes two types of data warehouses that support data and fusion result data. Among them, it supports the database to store the preprocessing results for other modules to call, and provides the basis for the feedback and correction of the final fusion results, including environment database, doctrine database, technology database, algorithm database, observation database and archive task database; the fusion database stores the fusion results of three levels: data, characteristics and decision-making, which can be output according to different needs, including target location identity database, situation assessment database, threat estimation database, etc. In order to ensure the normal operation of database management, it is necessary to use high-speed parallel reasoning mechanism and imprecise reasoning to deal with the mass and uncertainty of data. In addition, it also includes model base, rule base, knowledge base and so on.

Service module: this module is the human-computer interaction end of the system. According to the form required by users, this module will publish and provide the final fusion results to users in a unified form, and can provide multi-level and personalized result publishing services for different users. In addition, active push service can be provided according to the user's subscription, which supports the user to generate personalized information products and provide them for use according to the service components (combinations) freely selected by the user's information. If users are not satisfied with the results, they can also re input their opinions into the module and match and analyze with the user task description established in advance to correct and optimize the results. On the other hand, users can associate and initiate association queries according to the results.

3 Experimental Tests

In order to verify the feasibility of the above-designed multimodal information processing method for the online education system of college English courses, the following experiments are designed.

3.1 Experimental Environment

1) The hardware environment of the experiment
   The experiments were performed on a desktop computer with a main frequency of 2.66 GHz, a Core™ 2 quad-core processor, and a memory capacity of 2 GB.
2) Experimental software environment
   The host runs on the Windows 7 Ultimate operating system, and the experiment is built on the simulation experiment platform built by MATLAB R2009a. software.
3) The online education system for college English courses used in the experiment is an ios-based online education system. The system is to provide a mobile online education platform that both student users and teacher users can use. Student users

can use the fragmented time to learn the content of interest through the system and improve their competitiveness. At the same time, teacher users and institutions Users can show their skills through this platform and get considerable benefits.

In order to avoid too single experimental results, the traditional information processing method of online education system based on cloud computing was compared to complete performance verification together with the method in this paper.

3.2 Data Fusion Performance Test

Firstly, the data fusion performance of different methods is tested, mainly for the fusion efficiency of video visual features, image text features and audio features. The specific test results are shown in Table 2.

Table 2. Data fusion efficiency test results/%

Number of experiments	Amount of data/GB	Method of this paper	Traditional method
100	20	96.36	85.41
200	20	96.35	86.23
300	20	96.30	88.47
100	50	96.28	81.59
200	50	96.24	80.20
300	50	96.21	53.71
100	100	96.20	86.65
200	100	96.18	86.92
300	100	96.15	85.76

According to the fusion efficiency test results in Table 2, as the amount of data increases, the data fusion efficiency of the proposed method decreases slightly, but is overall higher than 96%. However, the data fusion efficiency of traditional methods has not reached 90%. In contrast, the data fusion efficiency of the proposed method is higher.

3.3 Information Processing Time Test

Then taking the information processing time as an indicator, the application performance of different methods is verified, and the results are shown in Fig. 2.

The analysis results shown in Fig. 2 show that: after the application of the method in this paper, the information processing time is always kept below 10 min. However, the information processing time of the traditional method is obviously higher than that of the method in this paper. It can be seen that the method in this paper has a higher timeliness in information processing for the online college English course education system.

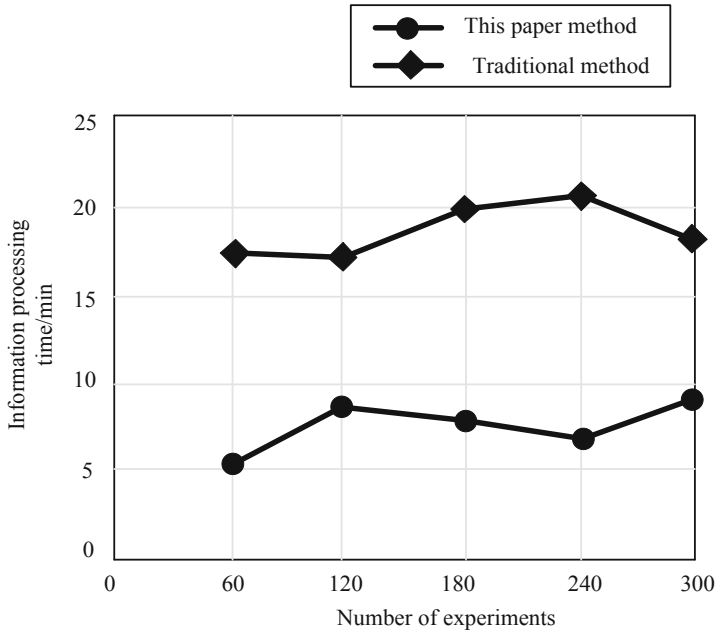

Fig. 2. Comparison of information processing time between different methods

4 Conclusion

Aiming at the multi-modal information in the online education system of College English courses, this study designs an effective processing method, which is of great significance to the development of the online education system.

In this paper, multi-modal information is divided into video information, image text information and audio information. On the basis of designing targeted feature extraction methods, information processing is realized through feature fusion. Compared with traditional methods, this method has higher data fusion efficiency and timeliness.

References

1. Behmanesh, M., Adibi, P., Chanussot, J., et al.: Geometric multimodal learning based on local signal expansion for joint diagonalization. IEEE Trans. Sig. Process. **66**(5), 129–141 (2021)
2. Khan, A., Maji, P.: Selective update of relevant eigenspaces for integrative clustering of multimodal data. IEEE Trans. Cybern. **37**(8), 1–13 (2020)
3. Yuan, M., Li, C.: Research on global higher education quality based on BP neural network and analytic hierarchy process. J. Comput. Commun. **09**(06), 158–173 (2021)
4. Li, G., Tan, X., Xiao, H.: Research on knowledge fusion model of decision support system for higher vocational education. Educ. Vocat. (10), 84–91 (2020)
5. Wang, L., He, Y., Tian, J.: Constructing and verifying a model of integrating multimodal data from online learning behaviors. Distance Educ. China (06), 22–30+51+76 (2020)
6. Behmanesh, M., Adibi, P., Chanussot, J., et al.: Geometric multimodal learning based on local signal expansion for joint diagonalization. IEEE Trans. Sig. Process. **35**(9), 391–405 (2021)

7. Jiménez-Bravo, M., Marrero-Aguiar, V.: Multimodal perception of prominence in spontaneous speech: a methodological proposal using mixed models and AIC. Speech Commun. **124**(11), 28–45 (2020)
8. Wei, X., Zhao, H.: Research on multi-source and multi-modal big data retrieval method based on mapreduce. Comput. Simul. **38**(4), 422–426 (2021)
9. Li, J., Peng, H., Hu, H., et al.: Multimodal information fusion for automatic aesthetics evaluation of robotic dance poses. Int. J. Soc. Robot. **12**(2), 5–20 (2020)
10. Gao, Y., Chang, H.J., Demiris, Y.: User modelling using multimodal information for personalised dressing assistance. IEEE Access **16**(5), 1–12 (2020)