Online-Journals.org

iJET International Journal of
Emerging Technologies in Learning

iJET | eISSN: 1863-0383 | Vol. 18 No. 22 (2023) | OPEN ACCESS

https://doi.org/10.3991/ijet.v18i22.44693

PAPER

A Classification and Retrieval System for Learning Resources of MOOC

Ye Zhang(✉)

School of Marxism,
Shijiazhuang University
of Applied Technology,
Shijiazhuang, China

2018010900@sjzpt.edu.cn

ABSTRACT

In this information age, massive open online courses (MOOCs) have become an integral component of modern education. These courses encompass a wide range of resources, such as videos, audios, texts, and other forms. An accompanying question is how to effectively organize, classify, and retrieve these resources. However, currently available classification and retrieval methods are mostly based on text retrieval technologies. As a result, multi-modal resources such as videos and audios are often ignored or incorrectly classified. Furthermore, more current methods exhibit low efficiency when processing the vast amount of data in MOOCs. To address and solve these issues, this study focuses on the extraction and fusion of multi-modal features of MOOC resources. It proposes an efficient classification and retrieval method based on 3D convolution, aiming to offer a more accurate and efficient approach for classifying and retrieving MOOC resources.

KEYWORDS

massive open online courses (MOOCs), multi-modal resources, attention mechanism, 3D convolution, classification, retrieval

1 INTRODUCTION

In the 21st century, the development of information technology, the internet, and mobile technology has propelled us into a new era of online learning. Now, massive open online courses (MOOCs) have become a hot spot in modern education, offering numerous learning opportunities and choices for learners worldwide [1–4]. The courses and textbooks of higher vocational education are sometimes limited by geographical, financial, and other factors. However, MOOCs can provide higher-level vocational students with high-quality courses designed by top universities and experts, thereby greatly increasing their learning resources. MOOCs are usually developed by top educational institutions and educators' course content and teaching methods are well-designed. With the help of MOOCs, higher-level vocational students can gain access to advanced educational concepts and have the opportunity

Zhang, Y. (2023). A Classification and Retrieval System for Learning Resources of MOOC. *International Journal of Emerging Technologies in Learning (iJET)*, 18(22), pp. 73–87. https://doi.org/10.3991/ijet.v18i22.44693

Article submitted 2023-08-03. Revision uploaded 2023-09-13. Final acceptance 2023-09-17.

to acquire more knowledge. The abundant content of MOOCs covers a variety of resources, including videos, audios, interactive exercises, and textual resources, constituting a multimodal learning environment [5, 6]. However, a key issue arising from such massive and diverse resources is how to effectively organize, classify, and retrieve them to meet the individual learning needs of different learners [7–10].

The classification and retrieval of MOOC resources are crucial for ensuring effective and efficient online learning [11, 12]. Properly organizing and providing these resources is conducive to helping learners find content that matches their learning objectives [13–16]. Besides, for educators, knowing which resources are popular with students and which ones should be upgraded or improved can help in formulating high-quality course programs. From a broader perspective, as society places more emphasis on lifelong learning and skill upgrading, establishing an efficient system for classifying and retrieving MOOC resources can greatly facilitate the dissemination and exchange of knowledge [17].

Despite the remarkable progress in the development of MOOCs in the past few years, there are a few limitations in the existing methods of classification and retrieval. First, most available methods still rely on conventional text-based retrieval techniques, which often leads to the oversight or misclassification of multi-modal resources such as videos and audios [18, 19]. Existing methods often fail to consider the complex hidden relationships among learning resources. For example, they may overlook the connection between speech and subtitles in videos or the association between speech content in audios and the accompanying visual elements. In addition, most methods are inefficient in dealing with the large-scale data of MOOC resources and cannot meet the requirement of real-time retrieval [20–23].

To address these issues, this study aims to discuss the method for extracting and integrating the multi-modal features of MOOC resources. The attention mechanism will be introduced to conduct in-depth analysis of the resources of each modality, including video, audio, and text, in order to extract their unique information and structure. Then, based on the distribution of attention weights for each modality, the multi-modal output will be weighted to create a comprehensive and representative vector of fused features. Further, this study proposes a new classification and retrieval method based on 3D convolution. The 3D convolution will be efficiently separated using the stream buffering technique. This will be coupled with the advanced causal convolution structure and the innovative context gated recurrent unit (GRU). The entire system is built with greater robustness and efficiency. These new methods not only greatly improve the accuracy of classifying and retrieving MOOC resources but also offer valuable references and insights for processing multi-modal data.

2 MULTI-MODAL FEATURE EXTRACTION AND FUSION OF MOOC RESOURCES

Massive open online courses usually offer learning resources in multiple formats, with videos and audios being the primary resources. Videos support visual instructions, such as demonstrations, diagrams, and animations, while audios provide learners with verbal explanations, discussions, or examples. Together, the resources from these two modalities can provide learners with a holistic learning experience. Video and audio modalities have their own distinct structures and information. For instance, the actions and expressions in videos, as well as the tones and rhythms in audios, can provide a wealth of information about the teaching content and methods. These features are crucial for the classification and organization of MOOC resources.

To extract accurate video and audio features, the first step is to preprocess the MOOC resources. This involves extracting frames from the learning resources and decomposing the video content into a series of static image frames. Typically, this extraction is done at specific time intervals. The audio of learning resources is subjected to segmentation processing in order to divide the audio signals into short-term segments with overlapping windows. For the video features of learning resources, the texture information of video frames is extracted using pixel values and color histograms, and the dynamic changes in videos are captured based on the differences between consecutive frames. When it comes to the audio features of learning resources, the Fourier transform is commonly used to obtain the spectral features of audio and analyze the rhythmic pattern, as well as the intensity or amplitude of sound.

2.1 Multi-modal feature fusion

Among existing methods for the problem of multi-modal feature fusion, the tensor-based fusion method is relatively effective. This is because tensors can intuitively represent high-dimensional data, providing a natural structural form for multi-modal data. Through tensor decomposition, the correlation and complementarity between different modalities can be effectively captured and utilized, resulting in more accurate fusion results. Assuming that X represents fused features, x^c represents video features, x^s represents audio feature, G represents the output vector of fused features mapped to a lower dimension, q represents the weight, and n represents the offset, there are:

$$X = \begin{bmatrix} x^c \\ 1 \end{bmatrix} \otimes \begin{bmatrix} x^s \\ 1 \end{bmatrix} \tag{1}$$

$$G = h(X; Q, n) = Q \bullet X + n \tag{2}$$

However, the tensor-based multimodal feature fusion method for videos has some defects. The computational complexity of tensor decomposition is very high, especially for large-scale and high-dimensional data. This can result in long computation times and may require significant computational resources. In light of this issue, the approach of utilizing modal-specific factors for conducting low-rank modal fusion can be employed to address it. The factor that decomposes weight into $\bar{Q}_j = \sum_{u=1}^{e} \bigotimes_{l=1}^{L} q_{l,j}^{(u)}$ is called an r-rank decomposition factor. Revised 2: To decompose the weight matrix based on modality and rank, we assume that the weight component decomposed based on modality is represented by ql. The dimension of the output tensor G is represented by g, the modality number is represented by l, and the Hadamard product (multiplying vectors by bits) is represented by Λ. Then, by performing *re*-rank factor decomposition, we obtain q(u) = [q(u)l, 1, q(u)l, 2, … ,q(u) l,g](u = 1, …, e). The output vector G can be expressed as:

$$G = \left(\sum_{u=1}^{e} \bigotimes_{l=1}^{L} q_l^{(u)} \right) \bullet X = \left(\sum_{u=1}^{e} \bigotimes_{l=1}^{L} q_l^{(u)} \right) \bullet \left(\bigotimes_{l=1}^{L} x_l \right) = \sum_{u=1}^{e} \left(\bigotimes_{l=1}^{L} q_l^{(u)} \bullet \bigotimes_{l=1}^{L} x_l \right) = \Lambda_{l=1}^{L} \left[\sum_{u=1}^{e} \bigotimes_{l=1}^{L} q^{(u)}{}_l \bullet x_l \right] \tag{3}$$

The conventional tensor-based fusion method may not effectively capture the weight distribution among different modalities, and there is a risk of losing key information from some modalities. In this paper, a novel multi-modal attention fusion module is proposed based on a conventional method. This module aims to assign appropriate weights to each modality during the fusion process by calculating attention weights. By doing so, it can effectively represent the features of learning resources, emphasize the key information in each modality, and ensure the retention of this information throughout the fusion process. Compared to the tensor-based method, the attention fusion module is more computationally efficient. It eliminates the need for complex tensor decomposition and instead directly utilizes the attention weight for feature fusion.

At first, this module combines the modal features of videos and audios into a 2048-dimensional vector using the modal-specific factor. The goal is to extract and fuse the key information from both modalities. Then, the module computes the attention weight distribution and performs normalization using the softmax function. This ensures that each modal feature is assigned an appropriate weight in the final fusion result. Next, after passing through two fully connected layers, the module transforms and compresses the original modal features. It then assigns weights based on the attention weights, ensuring that the modal features receive appropriate weights in the final fused features. Since the attention weight parameters are not fixed but learned by the network through training data, this ensures that the module can adaptively adjust the weights to optimize the fusion effect of multi-modal features. Figure 1 illustrates the structure of the modal feature fusion model for videos and audios.

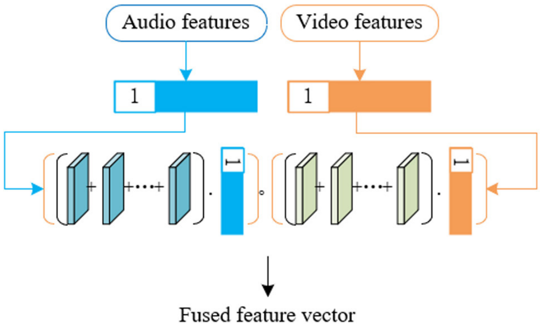

Fig. 1. Structure of modal feature fusion model of videos and audios

Assuming that $C \in E^{b \times 1024}$ represents video features of MOOC resources, $S \in E^{b \times 128}$ represents the audio features of MOOC resources, C represents the video features, S represents the audio features, $\beta \in E^{2048}$ represents the attention weights, and $G \in E^{b \times 2048}$ represents the fused feature vector of the modal-specific factor, the fused feature D of attention calculation is given by the following formula:

$$D = \mathrm{softmax}(q^{Y} G + n) \bullet \mathrm{concat}\left[DV(S), DV(C) \right] \tag{4}$$

To ensure the comprehensive extraction, integration, and evaluation of information from videos, audios, and their fused features, this study proposes a two-layer stacked attention fusion module. This module performs three-modal fusion of videos, audios, and fused features using low-rank decomposition. The objective of this step is to extract key information from videos, audios, and existing fused features. The next step is to perform two-stage enhanced fusion on multimodal features. In the first stage, the video and audio features are fused directly through low-rank decomposition, and their dimensions are standardized. In the second stage, the existing fused features are combined and integrated with video and audio features using low-rank decomposition to generate more comprehensive and detailed fused features. Furthermore, during the normalization operation, the module computes the attention weights of each modality during fusion using the fully connected layers and normalization operation. This ensures that the key information is assigned a higher weight in the final fused features. At last, after unifying the dimensions of video and audio features, the module combines the two and assigns weights to them using the computed attention weights. This is done in order to obtain more representative multi-modal fusion features. Assuming that "x_l" represents the modal feature and "D" represents the output of multi-modal fusion, the formula for calculating attention weights is:

$$D = \text{softmax}\left(q^{Y} \mathop{\Lambda}_{l=1}^{3} \left[\sum_{u=1}^{e} q_l^{(u)} \bullet x_l \right] + n \right) \bullet \text{concat}\left[DV(S), DV(C) \right] \tag{5}$$

2.2 Feature aggregation module

NetVLAD is a clustering method for deep learning that aims to extract a fixed-size descriptor from a set of features. It combines the *VLAD* (vector of locally aggregated Descriptors) method in computer vision with the capabilities of neural networks. In this paper, *NetVLAD* was used to combine video and audio features of MOOC resources. Figure 2 shows the structure of the *NetVLAD* model.

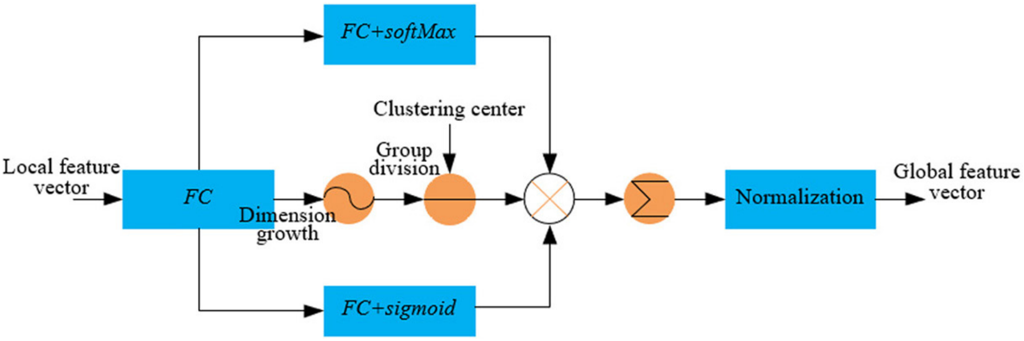

Fig. 2. Structure of *NetVLAD* model

In the first stage of *NetVLAD*, each input feature is assigned to a predefined cluster center. This is achieved by learning a weight set for each cluster center and using *Softmax* to determine the degree of correlation between each feature and its cluster center. Then, based on the relationship between each feature and its corresponding cluster center, the residual of each feature is determined. Next, all residuals are added together to generate a cumulative feature descriptor for each cluster center. After that, all cumulative descriptors are concatenated to form a single large vector, which is the output of *NetVLAD*. This provides a fixed-size representation of the entire input feature set. Assuming that "z_u" represents the F-dimensional feature vector of each image frame of MOOC resources, and $u \in \{1, \ldots, B\}$, then the following formula provides the global feature representation corresponding to each cluster center.

$$C(k,j)=\sum_{u=1}^{B}\beta_{j}(z_{u}(k)-v_{j}(k)) \tag{6}$$

$$\beta_{j}(z_{u})=\begin{cases}1 & z_{u}\in v_{j}\\ 0 & x_{u}\notin v_{j}\end{cases} \tag{7}$$

Due to the structure and computation of *NetVLAD*, particularly when the number of cluster centers is large, it may result in a significant number of parameters. This can make the model challenging to optimize and increase the risk of overfitting. To solve this problem, the input dimension needs to be reduced. Assuming that $h \in \{1, \ldots, H\}$ represents the number of divided groups, $u \in \{1, \ldots, B\}$ represents the number of image frames, $k \in \{1, \ldots, \eta F/H\}$ represents the feature dimension, $j \in \{1, \ldots, J\}$ represents the number of cluster centers, $\beta_{hj}(\dot{z}) = r^{qYhjz\cdot+nhj}/\sum_{a=1}^{J} r^{qYhjz\cdot+bhz}$ represents the weight of assigning group features to cluster centers, $\beta_{hj}(\dot{z}) = \delta(q_{h}^{y}\dot{z}_{u}+n_{h})$ represents the attention calculation of group h, where δ represents the *sigmoid* activation function. Then, there are:

$$t_{kj}=\sum_{u,h}c^{h}(k,j) \tag{8}$$

$$C^{h}(k,j)=\beta_{h}(z_{u})\beta_{hj}(z_{u})(z_{u}^{h}(k)-v_{j}(k)) \tag{9}$$

3 EFFICIENT MOOC RESOURCE CLASSIFICATION AND RETRIEVAL BASED ON 3D CONVOLUTION

When processing the content of MOOC resources, particularly videos, the data volume is typically substantial. Conventional convolutional neural networks often demand significant computational resources and memory, leading to reduced efficiency and inadequate real-time performance during practical operations. The stream buffering technique is an approach to handling large-scale continuous data streams. It can efficiently store and manage data, allowing the network to process data in blocks or batches instead of loading the entire dataset at once. By implementing this approach, the processing speed can be significantly increased, and the memory usage can be reduced. This enables the model to efficiently process MOOC video streams in real-time. Figure 3 illustrates the structure of the MOOC resource classification and retrieval model using the stream buffering technique.

Assuming: Y^{CL} represents the length of each sub-clip of MOOC resources, z_{u}^{CL} represents the current sub-clip, N represents a buffer with a time dimension of n whose start is represented by a tensor initialized as zero, D_u represents the feature map of current sub-clip z_{u}^{CL}, z_{u}^{CL} represents the sub-clip connected along the time dimension, N_u represents the buffer. To perform computation with D_u, z_{u}^{CL}, and N_u as the combined input, let d represent 3D convolution and $[-n:]$ represent the last n frames of the combined input. Then, the stream buffering technique can be expressed by the following formulas:

$$\begin{aligned} D_{u} &= d(N_{u}\oplus z_{u}^{CL})\\ N_{u+1} &= (N_{u}\oplus z_{u}^{CL})_{[-n:]} \end{aligned} \tag{10}$$

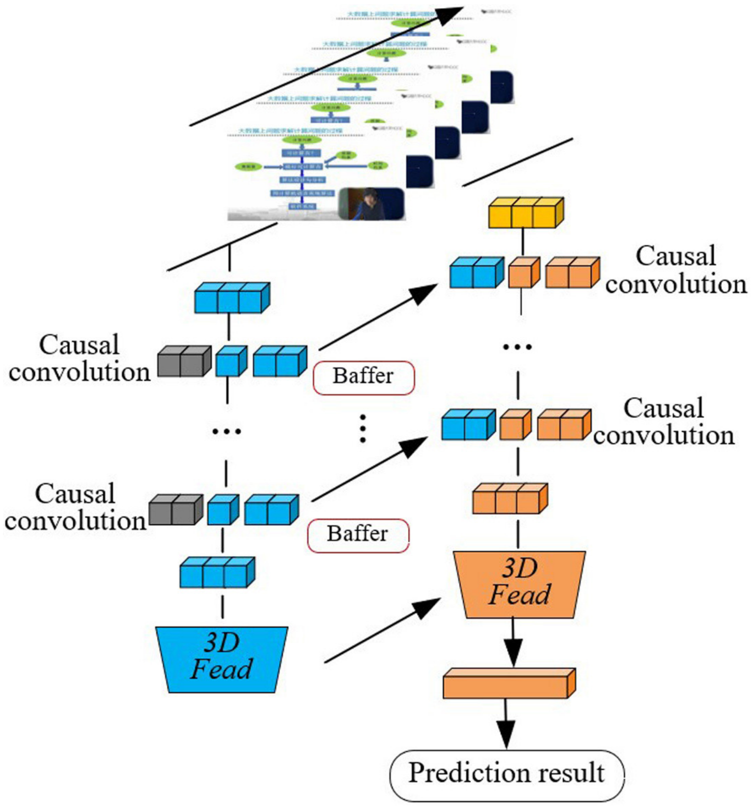

Fig. 3. Structure of classification and retrieval model using stream buffering technique

When processing time-series data, such as audio and videos, it is important to ensure that future information does not influence current decisions. Therefore, a mechanism is needed to ensure that the model is not affected by future data when processing the data of the current frame or time point. The causal convolution ensures that the model only utilizes past and current information at each time point. In contrast to conventional convolution, it does not take into account future data. As a result, it can provide more accurate predictions when processing sequential data. In terms of the video content of MOOCs, it means that when analyzing every video frame, the causal convolution ensures that decisions are made based only on past content. So, this study utilized causal convolution to replace all temporal convolutions in order to simulate the one-way time dimension. Specifically, the volume of the convolutions was calculated using the following formula, and all fillings were completed prior to the first frame. Assuming that o_u^{LE} and o_u^{RI} respectively represent left filling and right filling, the formulas for calculating the filling volume can be written as:

$$p_u^{LE}, o_u^{NI} = \begin{cases} \left(\dfrac{j_u - 1}{2}, \dfrac{j_u - 1}{2}\right) & \text{in case that } z \text{ is } odd \\ \left(\dfrac{j_u - 1}{2}, \dfrac{j_u}{2}\right) & \text{other cases} \end{cases} \tag{11}$$

Then, the filling volume of casual convolution can be expressed as:

$$o_y^{L-C}, o_y^{R-C} = (o_y^{LE} + o_y^{RI}, 0) \tag{12}$$

To align with the causal convolution in the time dimension, the original pooling operation has been updated to cumulative global average pooling. Assuming z represents the activation value in tensor form, then the formula is:

$$VHSO(z, Y') = \frac{1}{Y}\sum_{y=1}^{Y} z_y \tag{13}$$

The content of MOOCs usually contains rich contextual information. that fully capture or accurately classify using simple classification mechanisms. In order to enhance the classification performance, we require a method that can efficiently incorporate both the contextual information and excitation. The GRU, based on the squeeze and excitation context (SEC), can efficiently capture the long-term dependencies and contextual information of video and audio content. It controls the flow of information through a specific gating mechanism, thereby focusing more on the key contextual features. In terms of MOOC resources, this means that the classifier can better interpret the overall meaning and context of video or audio content, thereby making more accurate classification decisions. To achieve efficient classification and retrieval of MOOC resources, this paper proposes a model with a classification layer consisting of two parts: a gate control unit based on squeeze and excitation and context, and a classifier. At first, the features are passed through the SEC-based GRU. This introduces the attention mechanism on the channel dimension, allowing for better feature representations to be learned. These representations are then fed into the classifier to obtain the final result. Based on the following formula, the scaled features can be obtained by multiplying them with their respective weights.

$$x = \delta(h(d, Q)) = \delta(Q_2 \mathrm{ReLU}(Q_1 d)) \tag{14}$$

The structure of the MOOC resource classification and retrieval model, constructed in this paper, begins with the raw data of videos and audios. It utilizes a series of operations, including convolution, pooling, and full connections, to generate the final classification or retrieval results. Each operation targets a specific goal, such as feature extraction, dimensionality reduction, or classification. Modules contained in the model include: *Data, Convl, Pool2, Block3, Block4, Block5, Block6, Block7, Conv8, Pool9, FC10*, and *FC11*.

The *Data* module takes in raw data from MOOC resources, which includes video sequence frames and related audio signals. Its output is the corresponding video and audio features. The input of the *Convl* module is the video and audio features output from the *Data* module. This module performs 3D convolution operations using the stream buffering technique. It is the initial convolution operation of the model and aims to extract primary spatio-temporal features. The output size represents the low-level mapping dimensions of spatio-temporal features. The *Pool2* module takes the feature mapping from the *Convl* module as input. It performs pooling operations to decrease the spatial dimensions of the feature mapping, which helps reduce the computational load. The output of the *Pool2* module is the downscaled feature mapping. The inputs of the *Block3* to *Block7* modules are the outputs of their respective previous modules. These modules consist of consecutive convolutional layers, with each block containing multiple convolutional layers, batch normalization, and activation functions. In light of the introduced causal convolution structure, one of these blocks has a specific convolution setup.

After passing through each block, the features will become more complex, but the spatial dimension will decrease.

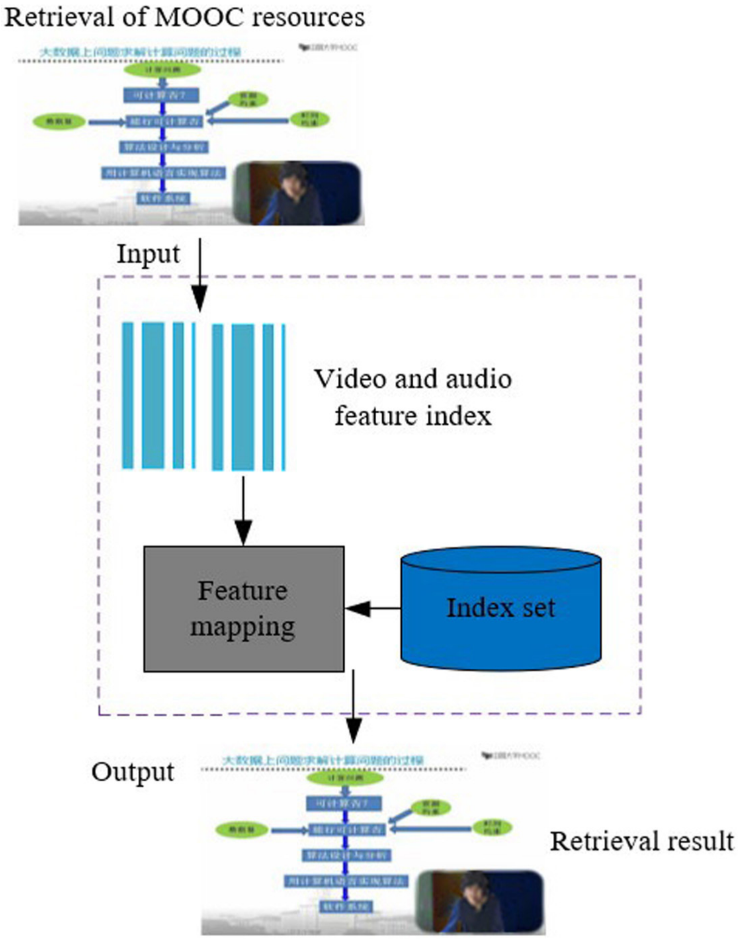

Fig. 4. Retrieval process of MOOC resources

The input of the *Conv8* module is the output of the *Block7* module. The purpose of the further convolution operations is to capture more complex features or prepare for subsequent pooling operations. This module increases the depth of the data, but the spatial dimension remains the same or may decrease slightly. The input of the *Pool9* module is the feature mapping of the output from the *Conv8* module. This module performs the final pooling operations to reduce the spatial dimension of the feature mapping to a smaller size. After undergoing downsampling in this module, the feature mapping will have a smaller spatial dimension. The input of the *FC10* module is the output of the *Pool9* module. The feature mapping will be flattened and passed through the fully-connected layer. The *NetVLAD method* is used for aggregation, resulting in a fixed-size descriptor. The output is a fixed-length feature vector. The input of the *FC11* module is the output of the *FC10* module. The fully-connected layer will be used again, and its purpose this time is to generate the final classification result. The SEC-based GRU classifier can be introduced at this time to enhance the model's classification performance. The output is the classification score or retrieval result of MOOC resources. Figure 4 illustrates the retrieval process for MOOC resources.

4 EXPERIMENTAL RESULTS AND ANALYSIS

Table 1. Accuracy of different models

Model	Feature	Top-1	Top-5
NetVLAD	Video	85.26	95.26
NetVLAD	Audio	39.24	65.22
NetVLAD + PFM	Video + Audio	88.15	95.16
NetVLAD + SFM	Video + Audio	87.43	95.14
NetVLAD + TFM	Video + Audio	88.26	96.68
NetVLAD + GFM	Video + Audio	88.63	96.88
The proposed model	Video + Audio	88.97	97.33

Table 1 compares the accuracy of various models under uni-modal and multi-modal conditions. For the *NetVLAD* model, when processing only the video features, the accuracy rates for Top-1 and Top-5 are 85.26% and 95.26%, respectively. These rates are significantly higher than the accuracy rates of 39.24% and 65.22% when for processing only the audio features. This suggests that in MOOC resources, the video content provides much more information than the audio content. Next, we compare the multi-modal feature fusion methods. For all fusion models (*PFM*, *SFM*, *TFM*, *GFM*), the accuracy rates of Top-1 and Top-5 are higher than those of the uni-modal methods. This further proves the significance of multi-modal feature fusion in improving classification accuracy. Among the various multi-modal fusion methods, *NetVLAD + GFM* achieved a high accuracy of 88.63% and 96.88% for Top-1 and Top-5, respectively. These results are only slightly lower than those obtained by the proposed method. According to the table, the proposed method achieved the highest accuracy for both Top-1 and Top-5, which were 88.97% and 97.33% respectively. This indicates that the strategy employed by the proposed method is more efficient and accurate in processing and combining video and audio features.

So, it can be concluded that the performance of unimodal methods is significantly lower than that of multimodal methods in classification tasks, particularly for audio data. In terms of multi-modal feature fusion methods, both the gated fusion model (*GFM*) and the proposed method exhibited excellent performance. However, the proposed method achieved the highest accuracy rates on Top-1 and Top-5, demonstrating its effectiveness and superiority. These results provide an important reference for subsequent research. Specifically, when dealing with multi-modal data, developing a more comprehensive and precise fusion strategy is crucial for enhancing classification accuracy.

Figure 5 compares the Top-1 accuracy of various models. When the iteration epoch is 0, the accuracy rates of all models are close. However, the *NetVLAD* model outperformed the others with an accuracy of 82.1%. This suggests that the uni-modal *NetVLAD* performs well initially, even without any training. The accuracy of *NetVLAD + PFM*, *NetVLAD + SFM*, and *NetVLAD + GFM* increased rapidly after a few iterations in the early stage and then stabilized. This indicates that these multi-modal fusion models can converge quickly. After three iterations, the accuracy of the proposed model reached 85%, indicating rapid convergence. However, the rate of improvement has since slowed down. Next, the upper limit of model performance is being compared. After 20 iterations, the Top-1 accuracy of the proposed model reached its highest value of 86.9%. This was followed by *NetVLAD* + *TFM, which achieved* an accuracy of 86.7%. These results indicate that as the iteration epoch increases, the proposed model demonstrates excellent performance. As a unimodal model, *NetVLAD* also exhibited good performance with a Top-1 accuracy of 86%. Finally, the performance of various models is compared. The proposed model and the *NetVLAD* + *TFM* model both performed exceptionally well during most iteration epochs, while *NetVLAD* + *SFM* and *NetVLAD* + *GFM* performed slightly worse. Although the performance of the proposed model in this paper was not the best during the early iterations, its performance gradually outperformed other models with the increase of iteration epochs.

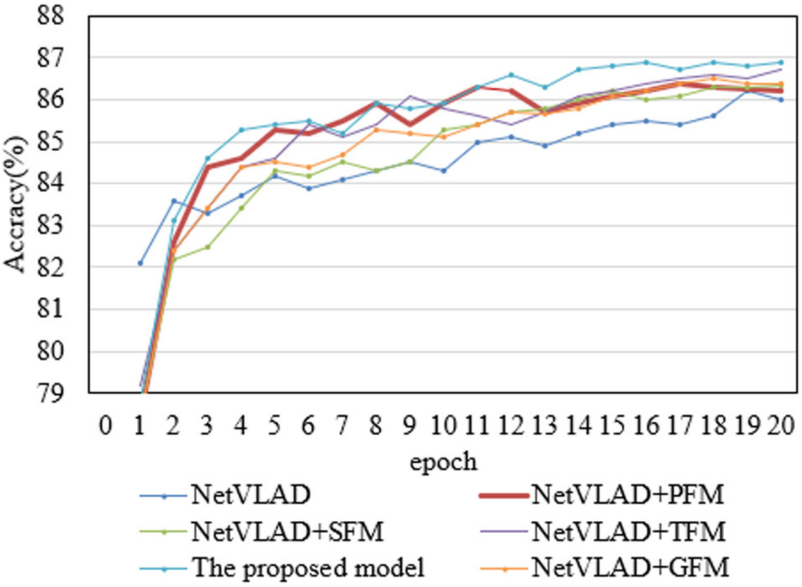

Fig. 5. Top-1 accuracy of different models

Thus, it can be concluded that during the iteration process, the proposed model demonstrated its superiority in multi-modal feature fusion and ultimately outperformed other models in terms of Top-1 accuracy. Although the uni-modal model *NetVLAD* performed well in the early stage, its performance ceiling is too low. Regarding the multi-modal fusion models, particularly the *NetVLAD* + *TFM* model and the proposed model, they exhibited superior performance and a higher upper limit during the iteration process. This further highlights the significance of multi-modal fusion methods in enhancing model performance. Additionally, the proposed method demonstrates higher efficiency and accuracy.

Again, a comparative analysis is conducted to determine the accuracy of different models based on the data provided in Table 2. Initially, the impact of feature type is discussed. It is evident from the table that the Top-1 accuracy of models utilizing solely video features (such as *TSCN*, 3*D-CNN*, *TCN*, *TSN*, *VTN*, *DVD*) varies from 71.2% to 84.2%. This suggests that relying solely on video features has limitations in this particular application. The models that combine videos and optical flows, such as *TS-CNN* and *FTSN*, showed some improvements compared to models that only use videos. The Top-1 accuracy of *FTSN* reached 76.1%. The models that combine videos, audios, and optical flows (such as *3S-CNN* and *MFN*) have further improved the classification accuracy. In particular, the accuracy of *MFN* reached 79.9%. Next, the complexity and performance of these models will be discussed. The simple *3D-CNN* model achieved a Top-1 accuracy of 73.5%, while more complex models like *DVD* achieved an accuracy of 84.2%. This suggests a positive correlation between model complexity and accuracy. According to the table, the proposed model achieved an 88.6% Top-1 accuracy and a 98.4% Top-5 accuracy using only video and audio features. These results are the highest among all models, emphasizing the effectiveness of the proposed method in feature fusion and classification. Except for the *VTN* model, the table lists the Top-5 accuracy rates of the other models. The top-5 accuracy of the proposed method is 98.4%, which further demonstrates its efficient feature extraction and classification capabilities.

Table 2. Comparison with other models

Model	Feature	Top-1	Top-5
TSCN	Video	71.2	88.5
3*D-CNN*	Video	73.5	91.2
TS-CNN	Video + Optical flow	74.5	91.3
FTSN	Video + Optical flow	76.1	93.4
TCN	Video	78.9	94.5
3*S-CNN*	Video + Audio + Optical flow	79.1	94.7
MFN	Video + Audio + Optical flow	79.9	95.6
TSN	Video	81.5	95.5
VTN	Video	81.9	–
DVD	Video	84.2	96.3
The proposed method	Video + Audio	88.6	98.4

Thus, it can be concluded that the proposed method is significantly better than other methods in terms of both Top-1 and Top-5 accuracy. This indicates that the strategy of the proposed method is efficient in feature extraction and fusion. Furthermore, it can also be observed from the table that models that combine multiple features (audio, video, and optical flow) tend to have better performance. However, an effective fusion mechanism is still needed to achieve optimal performance, as demonstrated by the proposed method.

At last, the video retrieval results of different models are compared based on the data provided in Table 3. The proposed model achieved a Mean Average Precision (*MAP*) of 0.9515 and an Average Recall Precision (*ARP*) of 1.528 under a dimensionality of 63. The performance was excellent. The (*MAP*) and (*ARP) without* the implementation of the stream buffering technique are both slightly higher than those of the proposed model. In the absence of introducing causal convolution, the dimensionality increased significantly to 26358. This led to a significant decrease in performance compared to other models, particularly in terms of the MAP value. This suggests that causal convolution plays a crucial role in improving the model's performance. In the case of not introducing the SEC-based GRU, although the dimensionality is the same as that of the proposed model, the performance declined slightly, especially on *MPA*. This indicates that the SEC-based GRU has a positive effect on the model's performance. Under the condition of relatively high dimensionality (524), the model's performance was the lowest. This suggests that the conventional convolution model may not be the optimal choice for this type of application.

Table 3. Video retrieval results of different models

Method	MAP	ARP	Dimension
The proposed method	0.9515	1.528	63
Without stream buffering technology	0.9617	1.539	63
Without causal convolution	0.9238	1.951	26358
Without SEC-based GRU	0.8566	1.754	63
Conventional convolution model	0.8317	1.826	524

Thus, it can be concluded that the proposed model can provide high performance even under low dimensionality, demonstrating its effectiveness in feature extraction and video retrieval. Causal convolution and SEC-based GRU both play key roles in enhancing model performance, while the conventional convolution model is not suitable for these types of applications. Therefore, when designing a video retrieval system, choosing the appropriate combination of models and techniques is crucial.

5 CONCLUSION

This study extracted the features of multi-modal information from MOOC resources, such as videos and audios. The study adopted the *NetVLAD* technique, which is a method suitable for capturing long time span features, to aggregate video and audio features. Additionally, the paper discussed several multi-modal feature fusion models, including *PFM*, *SFM*, *TFM*, and *GFM*, in order to achieve richer and more accurate representations from different modalities. This paper proposes an efficient 3D convolutional neural network that aims to process MOOC resources with high efficiency. The network utilizes the stream buffering technique and combines the causal convolution structure with SEC-based GRU to enhance the model's classification and retrieval ability.

Through multiple tests, we compared the performance of different combinations of models and techniques on various indicators, including Top-1 accuracy, Top-5 accuracy, *MAP*, and *ARP*. The proposed method demonstrated strong performance across most indicators, particularly when considering multi-modal information. It resulted in a significant improvement in performance compared to unimodal or conventional methods. Based on the test results, the proposed method has achieved excellent performance on multiple evaluation indicators, thus verifying its effectiveness. These findings provide a new and efficient method for classifying and retrieving learning resources on MOOC platforms, aiming to enhance the user experience and learning outcomes.

6 REFERENCES

[1] S. Urooj, A. Sajjad, N. Bano, and M. Mukarram, "Gutenberg and the MOOC," *International Journal of Emerging Technologies in Learning*, vol. 17, no. 12, pp. 93–105, 2022. https://doi.org/10.3991/ijet.v17i12.31357

[2] Y. Q. Zhang and A. Mangmeechai, "Exploring the factors of undergraduate learners' engagement and knowledge sharing for sustainable MOOC learning," *International Journal of Sustainable Development and Planning*, vol. 17, no. 3, pp. 1007–1015, 2022. https://doi.org/10.18280/ijsdp.170332

[3] K. Najmani, E. H. Benlahmar, N. Sael, and A. Zellou, “A systematic literature review on recommender systems for MOOCs,” *Ingénierie des Systèmes d’Information*, vol. 27, no. 6, pp. 895–902, 2022. https://doi.org/10.18280/isi.270605

[4] N. Mohamad, A. Othman, T. S. Ying, N. Rajah, and N. Samsudin, “The Relationship between Massive Online Open Courses (MOOCs) content design and students’ performance,” *International Journal of Interactive Mobile Technologies*, vol. 15, no. 4, pp. 4–15, 2021. https://doi.org/10.3991/ijim.v15i04.20201

[5] A. K. Singh, S. Kumar, S. Bhushan, P. Kumar, and A. Vashishtha, “A proportional sentiment analysis of MOOCs course reviews using supervised learning algorithms,” *Ingénierie des Systèmes d’Information*, vol. 26, no. 5, pp. 501–506, 2021. https://doi.org/10.18280/isi.260510

[6] Y.-A. Bachiri and H. Mouncif, “Artificial intelligence system in aid of pedagogical engineering for knowledge assessment on MOOC platforms: Open EdX and moodle,” *International Journal of Emerging Technologies in Learning*, vol. 18, no. 5, pp. 144–160, 2023. https://doi.org/10.3991/ijet.v18i05.36589

[7] S. Sraidi, E. M. Smaili, S. Azzouzi, and M. E. H. Charaf, “A neural network-based system to predict early MOOC dropout,” *International Journal of Engineering Pedagogy*, vol. 12, no. 5, pp. 86–101, 2022. https://doi.org/10.3991/ijep.v12i5.33779

[8] F. Y. Cao, Y. T. Feng, and B. Wei, “Practical research of online teaching platform on reform of computer network with flipped classroom,” *Review of Computer Engineering Studies*, vol. 9, no. 2, pp. 67–70, 2022. https://doi.org/10.18280/rces.090204

[9] S. Assami, N. Daoudi, and R. Ajhoun, “Implementation of a machine learning-based mooc recommender system using learner motivation prediction,” *International Journal of Engineering Pedagogy*, vol. 12, no. 5, pp. 68–85, 2022. https://doi.org/10.3991/ijep.v12i5.30523

[10] A. M. Nidhom, A. B. N. R. Putra, A. A. Smaragdina, G. D. K. Ningrum, and J. M. Yunos, “The integration of augmented reality into MOOC’s in vocational education to support education 3.0,” *International Journal of Interactive Mobile Technologies*, vol. 16, no. 3, pp. 20–31, 2022. https://doi.org/10.3991/ijim.v16i03.28961

[11] R. Vankayalapati, K. B. Ghutugade, R. Vannapuram, and B. P. S. Prasanna, “K-means algorithm for clustering of learners performance levels using machine learning techniques,” *Revue d’Intelligence Artificielle*, vol. 35, no. 1, pp. 99–104, 2021. https://doi.org/10.18280/ria.350112

[12] Y. Yang, D. Zhou, and X. Yang, “A multi-feature weighting based K-means algorithm for MOOC learner classification,” *Computers, Materials and Continua*, vol. 59, no. 2, pp. 625–633, 2019. https://doi.org/10.32604/cmc.2019.05246

[13] Y. Zhang and J. Gao, “The current situation of medical information retrieval teaching under the background of MOOC,” *Journal of Physics: Conference Series*, vol. 1744, no. 4, p. 042152, 2021. https://doi.org/10.1088/1742-6596/1744/4/042152

[14] S. Bhoir, T. Ghorpade, and V. Mane, “Semantic search over MOOC aggregator using query expansion,” in *International Conference on Energy, Communication, Data Analytics and Soft Computing*, Chennai, India, 2018, pp. 2237–2240. https://doi.org/10.1109/ICECDS.2017.8389849

[15] X. Yusheng, “Research on the construction of mixed teaching resources in higher vocational colleges based on MOOC,” in *5th International Conference on Smart Grid and Electrical Automation*, Zhangjiajie, China, 2020, pp. 519–523. https://doi.org/10.1109/ICSGEA51094.2020.00118

[16] J. J. Williams, “Improving learning in MOOCs with cognitive science,” *CEUR Workshop Proceedings*, vol. 1009, pp. 49–54, 2013.

[17] P. M. Ashok Kumar, R. R. Ambati, and L. Arun Raj, "An efficient scene content-based indexing and retrieval on video lectures," *Advances in Intelligent Systems and Computing*, vol. 1171, pp. 521–534, 2021. https://doi.org/10.1007/978-981-15-5400-1_53

[18] Neha and E. Kim, "A classification method of the learners' queries in the discussion forum of MOOC to enhance the effective response rate from instructors," in *23rd HCI International Conference, Virtual Event*, 2021, pp. 109–115. https://doi.org/10.1007/978-3-030-78645-8_14

[19] S. Bouzayane and I. Saad, "A preference ordered classification to leader learners identification in a MOOC," *Journal of Decision Systems*, vol. 26, no. 2, pp. 189–202, 2017. https://doi.org/10.1080/12460125.2017.1252233

[20] X. Wei, H. Lin, L. Yang, and Y. Yu, "A convolution-LSTM-based deep neural network for cross-domain MOOC forum post classification," *Information*, vol. 8, no. 3, p. 92, 2017. https://doi.org/10.3390/info8030092

[21] H. L. Pinto and V. Rocio, "Combining sentiment analysis scores to improve accuracy of polarity classification in MOOC posts," in *EPIA Conference on Artificial Intelligence*, Vila Real, Portugal, 2019, pp. 35–46. https://doi.org/10.1007/978-3-030-30241-2_4

[22] Z. Kastrati, A. S. Imran, and A. Kurti, "Integrating word embeddings and document topics with deep learning in a video classification framework," *Pattern Recognition Letters*, vol. 128, pp. 85–92, 2019. https://doi.org/10.1016/j.patrec.2019.08.019

[23] A. Boughoula, C. Geigle, and C. X. Zhai, "A probabilistic approach for discovering difficult course topics using clickstream data," in *Proceedings of the 4th ACM Conference on Learning @ Scale*, Cambridge Massachusetts, USA, 2017, pp. 303–306. https://doi.org/10.1145/3051457.3054010

7 AUTHOR

Ye Zhang, holds a master's degree from Tianjin Normal University and is currently employed at Shijiazhuang University of Applied Technology. Her primary research areas include ideological and political education (E-mail: 2018010900@sjzpt.edu.cn; ORCID: https://orcid.org/0009-0001-3275-9826).