Construction of Sports Culture Recommendation Model Combining Big Data Technology and Video Semantic Comprehension

Bin Xie[1], Fuye Zhang[2]
Department of Physical Education, Henan Institute of Economics and Trade, Zhengzhou 450046, Henan, China[1]
Internet of Things College, Henan Institute of Economics and Trade, Zhengzhou 450046, Henan, China[2]

***Abstract*—Information blast makes it harder for clients to channel the substance they are keen on. This study aims to combine big data and video semantic comprehension technology to realize the recommendation of sports culture videos by exploring the semantics of video and taking advantage of multi-source heterogeneous information. The semantic structure of unstructured video data is defined first, and on this basis, Converse3D (C3D) - Connectionist Temporal Asifationon (CTC) is employed to complete the extraction of sub-action semantics and the integration of behaviour semantic sequences. In adjustment to break the botheration of low accurateness of the model for the semantic abstraction of unlabeled videos, this study proposes an unsupervised semantic abstraction adjustment based on Converse3D(C3D)-RAE, which completes the compression and affiliation of the semantic sequences and verifies the accurateness of both two models through experiments. In order to solve the problem of insufficient accuracy of video recommendation algorithms based on single video semantic similarity and topic similarity, this study comprehensively considers video semantic similarity and video topic similarity and proposes a multi-modal video recommendation algorithm. The experimental results show that the accuracy of the COMSIM-based algorithm is 7.8% higher than that of Video+ CNN + K-NearestNeighbor (KNN) and 15.9% higher than that of CLIP + CNN +Ncut+LDA.**

***Keywords*—Big data; video semantic comprehension; sports culture; semantic sequences; convolutional neural networks (CNN)**

I. Introduction

With the rapid advancement of information technology in recent years, the Internet has brought convenience to people's lives, providing them with a large amount of data and information resources, greatly satisfying people's demand for network information, and moving the Internet industry into the big data era [1], [2]. However, excessive information from multiple sources has also become an obstacle for people to enjoy convenience. Specifically, information in the new era is no longer delivered to users by websites or media in one direction, but more often than not, users and media have formed an effective two-way transmission, which has significantly expanded the existing data stock [3], [4]. In addition, the continuous sinking of the user market has led to the exponential expansion of the user community and the exponential growth of data [5]. However, instead of enjoying convenient services, the huge amount of data and information has plunged people into the vast data mud, a trap known as "information overload". For Internet users, it is undoubtedly difficult to find the data knowledge they need in the exponentially growing data information, and users usually need to spend a lot of time and effort to find this information, which leads to a very poor user experience [6], [7]. For service providers, it is difficult to fully explore users' interests and preferences due to the large amount and complex types of log information generated by users, so they cannot accurately determine the content of users' needs, which greatly reduces the service quality and may push inaccurate information to users, which destroys users' trust in the system and leads to user churn [8], [9]. Faced with the information overload network environment and multiple sources of heterogeneous data types, how to make users less frustrated when searching for information, how to deepen users' trust in service providers, and how to enable service providers to maximize the benefits of efficiency are hot issues that need to be solved in the current data mining field.

At present, there are two main ways to solve this problem: information retrieval technology and information filtering method [10], [11]. The content presented by search engines for different users is often the same, which cannot meet the personalized needs of each user. The personalized recommendation technology based on information filtering does not require users to give their specific needs [12]. It can mine users' interests and preferences through users' historical behaviour logs and browsing habits to generate relevant recommendations [13], [14]. In today's big data era, personalized recommendation technology can greatly improve the efficiency of processing problems and meet the personalized needs of users. This technology has been successfully applied to many fields and has become a research hotspot in the computer application field.

Digital video data contains abundant spatial and temporal information, but it is difficult to describe the video content because of its huge amount of data and the semantic gap between low-level features and high-level semantics [15], [16]. With the rapid development of deep learning, especially the development of cross-modal learning, people's understanding of video semantics has reached a new level. Therefore, based on the video semantic understanding model, this paper proposes a sports culture video recommendation method that integrates video semantics and video topic text similarity.

This article analyzes the construction technology of sports culture models that combine video semantics. The accuracy of the model was verified through data recommendation of diverse and heterogeneous information. The innovation points include:

1) By exploring the semantics of videos and utilizing heterogeneous information from multiple sources, sports culture videos can be recommended.

2) This article uses C3D-CTC to extract sub action semantics and integrate behavioral semantic sequences. In order to overcome the problem of low accuracy in unlabeled video semantic abstract models.

3) This article addresses the issue of insufficient accuracy in video recommendation algorithms based on single video semantic similarity and topic similarity.

Section I analyzes the background of uncertainty when users search for information in the face of information overload in the network environment and the multi-source nature of heterogeneous data types. How to deepen users' trust in service providers has become a current issue worth studying. Section II analyzes video recommendation algorithms as a popular component of recommendation systems. Considering the issue that topic modeling techniques cannot cluster short texts that ignore semantic relationships between words. Section III recommends methods for sports culture videos that consider multiple modes and analyzed the semantic extraction of sports videos in C3D. Section IV conducted validation analysis on the dataset used, including the UCF-12 dataset and the sports video dataset captured from the video website Vine.com. Section V summarizes the entire text. This study proposes an unsupervised semantic abstraction adjustment method based on C3D-RAE, which completes the compression and membership of semantic sequences, and verifies the accuracy of the two models through experiments.

II. Related Work

Video recommendation algorithms, as an important component of recommendation systems, have always been a hot research direction. Considering the disadvantage of topic modeling technology not being able to cluster short texts that ignore semantic relationships between words, M S Tajbakhsh et al. [17] proposed a topic modeling method for semantic relationships between words in Twitter social network tweets. Yaduv U et al. [18] proposed a recommendation system method based on linked open data and social network features. This method solves the problem of pure new user cold start by constructing user profiles based on collaborative features of linked public data and features of social networks. Nikolakopoulos et al. [19] proposed the EIGENREC recommendation model based on the existing PureSVD algorithm, which comprehensively utilizes multiple recommendation strategies. The created model can modify recommendation results in real-time based on the popularity of the project. Salah et al. proposed a weighted clustering method to address the issue of data sparsity in dynamic incremental collaborative filtering. The experimental results showed that the constructed model has fast computational speed and low computational cost [20]; Hewitt et al. classified eight types of emotions and proposed three improved convolutional neural networks to implement a music recommendation interface based on predicting user influence [21]. In order to solve the cold start problem caused by the lack of correlation score when adding new videos, Li Y et al. [22] proposed directly calculating video correlation from the content. And use deep convolutional neural networks to process video information, thereby constructing a video correlation table. Li X et al. [23] proposed a multi-directional pyramid common attention module for learning the attention values of two modalities in different dimensional spaces. Fan C et al. [24] viewed the semantic information of videos as a series of ordered events that occur in sequence as a whole and are continuously read and written to memory. The input sequence can be understood from a global perspective to prevent local capture due to information interference.

In summary, the current research model cannot perform sparsity analysis on dynamic incremental collaborative data. This indicates that there are certain shortcomings in the video correlation of emotional users. For a long time, users have been plagued by low accuracy of unlabeled video semantic abstract models. On this basis, this article optimizes multi-source heterogeneous information. Using C3D-CTC to extract sub action semantics and integrate behavioral semantic sequences. It breaks the problem of low accuracy in unlabeled video semantic abstract models.

III. Sports Culture Video Recommendation Method Considering Multi-Modality

A. Semantic Extraction for Sports Videos based on C3D

1) Video behaviour semantic extraction based on supervised learning: Since video data is unstructured data, in order to analyze and process video more accurately, video data must be structured, which is the basis of video feature extraction and semantic extraction. Video is a continuous frame sequence with no obvious segmentation point. It isn't easy to extract the semantics of the entire video directly. Therefore, it is necessary to divide the video into multiple segments and extract the semantics of the segments respectively. Video can be divided into the following levels according to its physical level: video, scene sequence, shot sequence and frame sequence, as shown in Fig. 1.

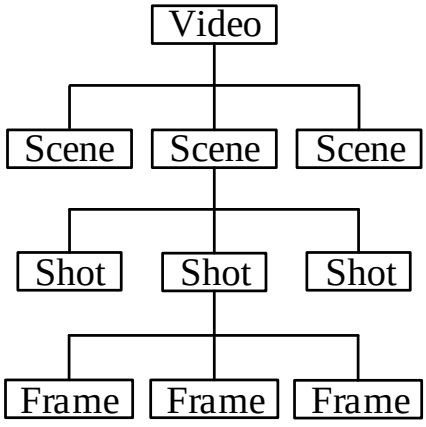

Fig. 1. Hierarchical structure of video.

Scene: A group of shots expressing the same theme. These shots record an event at the same time and place from different angles in space. Together, they describe the semantic concepts related to this event. For example, the scene of "air relay" can be composed of multiple shots such as "player passing", "another player is catching the ball in the air", "player shooting", or "player dunking", "ball in", etc. Shots: a recording process of a camera. Shots are the logical component unit of video and are also the smallest unit that can be used in indexing video. As the physical unit of video, a frame is defined as the frame is the smallest unit of video image - a single picture.

According to the physical hierarchy of video, this study defines the semantic structure of video as shown in Fig. 2.

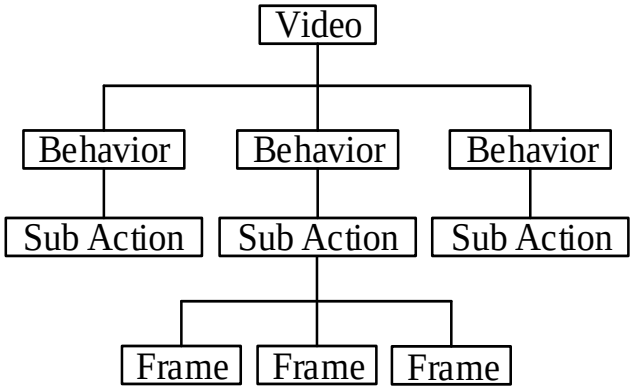

Fig. 2. Semantic structure of video.

Convolutional Neural Networks (CNN) have been extensively used in computer vision-related research in recent years, whose structure mainly includes a convolution layer, pooling layer and full connection layer. In image processing, CNN generally adopts the "planar convolution" with the dimension of convolution kernel of 2. However, when analyzing video, 2D convolution has a poor ability to capture timing information, leading to the loss of video timing information.

Therefore, 3D Convolutional Neural Network (C3D) with a convolution kernel dimension of three is adopted in this study, which can capture features in both temporal and spatial dimensions. The concrete implementation of 3D convolution operation is as follows: a three dimensions frame cube is obtained by stacking several consecutive frames, and the cube is convolved with the 3D convolution kernel, as shown in Fig. 3, where the connection of the same colour represents the shared weight, that is, there are weight values of three dimensions. Through such a convolution operation, the feature map obtained by convolution is connected with several consecutive frames of the previous layer to capture the timing information. D. Tran et al. [25] found that the same convolution kernel structure of 3*3*3 has the best accuracy on C3D. Therefore, the same convolution kernel of 3 * 3 * 3 is employed in this study.

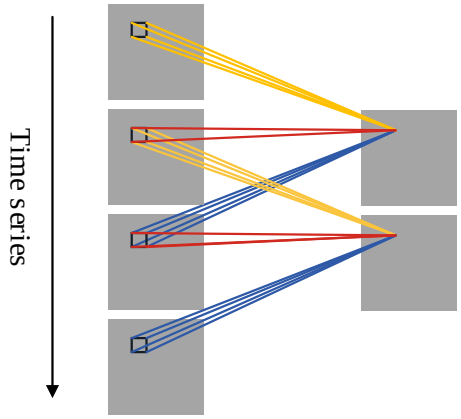

Fig. 3. 3D convolution.

Due to the large amount of redundant information between video frames and sub-actions, as well as the repetition of video frames in the time dimension, the extraction of video semantics will be affected. Therefore, this study puts to use the connectionist temporal classification (CTC) to solve the above problems. CTC algorithm was first applied in the acoustic training model, which is a complete end-to-end training without strict alignment of data in advance. For example, to understand intra coding, most areas of the image have the same color. False i will encode the red area, assuming that the colors in the frame remain consistent in the vertical direction. This means that the color of the unknown pixel is the same as that of adjacent pixels. Although this prior prediction technique (intra frame prediction) is used, the actual value is subtracted to calculate the residual. The residual matrix obtained in this way is easier to compress than the original data [26]. Yang improved the recognition algorithm by analyzing graph regularization and constructing a model, and compared and analyzed feature extraction methods. Meanwhile, the experiment aims to investigate the improvement of the improved recognition algorithm on English semantic translation after feature extraction [27]. Traditional content-based video retrieval algorithms typically only utilize the underlying features of video images, resulting in insufficient content description and unsatisfactory retrieval results. Guo studied how to combine the underlying features of videos with semantic features, improved the existing indexing structure, and designed an efficient sports video retrieval algorithm [28].

C3D is applied to model the sub-action features first, and different kinds of probability distributions of each sub-action are obtained. x represents the sub-action sequence, and y represents the output sequence of C3D.

The output of the C3N model is the probability of the category corresponding to each sub-action, which is a vector of N+1-dimensional probabilities, representing the different probabilities of N+ 1 category, N represents the number of sub-action categories, and 1 represents blank. We employ N_w to represent the convolutional neural network, and then the network output can be expressed as:

$$y=N_w(x) \quad (1)$$

For any sub-action input sequence of length T, whose corresponding label sequence is z, we can obtain that the

occurrence probability of label sequence z is the product of label probabilities at each moment.

$$P(\pi|x) = \prod_{t=1}^{T} P(\pi_t|x) \quad (2)$$

Where π denotes the decoding path and π_t represents the t-th sub-action label in the decoding path.

The probability of decoding path π can be calculated from the output of the C3D model as follows:

$$P(\pi|x) = \prod_{t=1}^{T} y_{\pi_t}^{t} \quad (3)$$

where, $y_{\pi_t}^{t}$ denotes the probability that the t-th sub-action label is π_t.

Define a many-to-one mapping B to remove all blank symbols and merge duplicate labels, transforming the decoding path π into label l. For example, (-,A,-,-,E,-) and (-,A,-,E,-,-) are both mapped to label (A,E). B^{-1} represents the inverse process of mapping B, which is a one-to-many mapping, that is, mapping label (A, E) into a sequence of labels with duplicate labels and blank symbols for all possible decoding paths so that the final decoding path π is the sum of the probability of each sequence with the probability of the label sequence given the input sequence x.

$$P(l|x) = \prod_{t=1}^{T} P(\pi|x), \pi \in B^{-1}(l) \quad (4)$$

Given an input sequence x and the corresponding label l, the loss function of the C3D-CTC model adopts the maximum likelihood error, that is, to minimize the negative logarithm of the probability:

$$\text{C3D-CTC}(x) = -log\, P\,(l|x) \quad (5)$$

In Eq. (4), the calculation of the objective function requires an exhaustive enumeration of all decoding paths, which is very difficult. In fact, only a small part of all paths are effective. Therefore, this study adopts the Forward-Backward Algorithm (FBA), a kind of dynamic programming algorithm, to calculate the objective function of the model.

For a given label l of length T, in order to find all paths π satisfying l = B(π), we need to construct an extended label l', whose length is 2T+1 by adding a blank at the beginning, end and middle of each character. For example, the sequence (A, E), whose extension label L 'is (-,A,-,E,-). The legal decoding path must meet the following conditions: (1) The path conversion can only be right or down; (2) There must be a blank between the same characters; (3) Only blank symbols can be skipped; (4) The path must start with the first two symbols; (5) The path must end with the last two symbols. The conversion process of all decoding paths of label l is shown in Fig. 4.

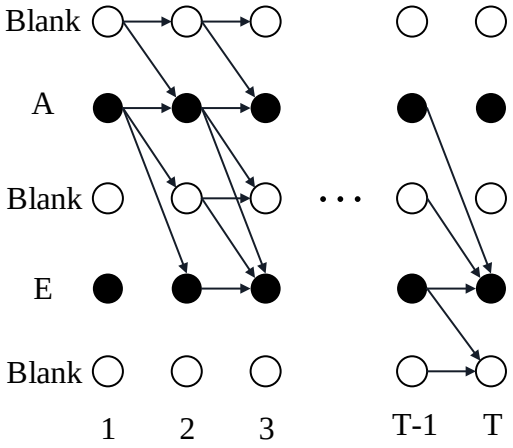

Fig. 4. All decoding paths of the label 'AE'.

To calculate the sum of the probabilities of all the above decoding paths, define the forward probability $\alpha_t(s)$ to denote the sum of the forward probabilities of all paths at the t-th input with s as the endpoint:

$$\alpha_t(s) = \sum_{B(\pi_{1:t})=l_{1:s}} \prod_{t'=1}^{t} y_{\pi_{t'}}^{t'} \quad (6)$$

Where s denotes the number of rows, the initial state of the forward probability $\alpha_t(s)$ can be calculated as follows:

$$\begin{cases} \alpha_1(1) = y_{blank}^{1} \\ \alpha_1(2) = y_{l_1}^{1} \\ \alpha_1(s) = 0, \forall s > 2 \end{cases} \quad (7)$$

The decoding path shows that: (1) there are only two possibilities from blank, either blank or l_i, i.e., the i-th element of label l; (2) there are threecases from l_i, that is, l_i, blank and l_i+1, with 3 possible output cases; (3) the input case of the blank is two and the input case of l_i is 3. We can obtain $\alpha_t(s)$ with the following iterative formula.

$$\alpha_t(s) = \begin{cases} (\alpha_{t-1}(s) + \alpha_{t-1}(s-1))y_{l_s}^{t}, if\ l_s = blank\ or\ l_{s-2} = l_s \\ (\alpha_{t-1}(s) + \alpha_{t-1}(s-1) + \alpha_{t-2}(s-1))y_{l_s}^{t}, otherwise \end{cases} \quad (8)$$

For the part without connection in the upper right corner, it represents the node that cannot be reached by the legal decoding path, so we can add restrictions and set the probability of these paths to 0.

$$\begin{cases} \alpha_t(s) = 0, \forall s<1 \\ \alpha_t(s) = 0, \forall s < |l'| - 2(T-t) - 1 \end{cases} \quad (9)$$

Therefore, the loss function of the C3D-CTC model can be expressed as follows:

$$-\ln(P(l|x)) = -[\ln(\alpha_T(2T+1) + \alpha_T(2T))] \quad (10)$$

The video behaviour semantic extraction model based on C3D-CTC is shown in Fig. 5.

2) Video latent semantic extraction Based on unsupervised learning: For some videos with blurred boundaries between actions, the C3D-CTC model is difficult to get accurate labels by action decomposition, and its scalability for new types of videos is poor. In order to solve the above problems, this study proposes a Recursive Auto-Encoder (RAE) based video latent semantic extraction method.

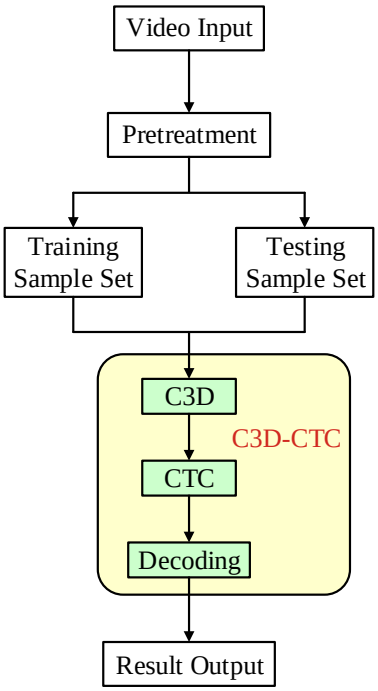

Fig. 5. C3D-CTC model.

RAE is an auto-encoder that searches for variable-length input structures, which is put to use in both NLP and computer vision. When the input is a sequence of word vectors, RAE uses neural networks to sense the score of all two adjacent word vectors. By measuring the probability of synthesizing a word for each pair of word vectors, the one with the largest probability is selected as the representative of the two-word vectors, the vector of the two words is removed from the sequence, and the synthesized vector is inserted into the position of the previous two words in the sequence. This is done recursively until the entire input statement is mapped to a vector at the root of the binary tree.

The input word vector sequence is (x_1, x_2, x_3, x_4, x_5), then the encoding process of word vector pair (x_1, x_2) transforming to parent node y_1 is defined as:

$$y_1 = f(W^{(1)}[x_1, x_2] + b^{(1)}) \quad (11)$$

where, $W^{(1)}$ represents the $n*n$ matrix parameters and $b^{(1)}$ is a bias term.

The decoding process of the parent node y_1 reconstructing x_1 and x_2 can be presented as follows:

$$[x_1', x_2'] = W^{(2)} y_1 + b^{(2)} \quad (12)$$

The reconfiguration error of the auto-encoder is as follows:

$$E_{rec}([x_1, x_2]) = \frac{1}{2}\|[x_1, x_2] - [x_1', x_2']\|^2 \quad (13)$$

Define A(x) as all adjacent node pairs of the input sequence x, and define T(y) as the case where the binary child node is transformed into the parent node. We define the objective function of the recursive auto-encoder as:

$$VE(x) = \underset{y \in A(x)}{argmin} \sum_{s \in T(y)} E_{rec}([x_1, x_2]_s) \quad (14)$$

When generating weights, in order to avoid weight skew, the reconstruction errors generated each time should be normalized.

$$E_{rec} = \frac{n_1}{n_1 + n_2}\|x_1 - x_1'\|^2 + \frac{n_2}{n_1 + n_2}\|x_2 - x_2'\|^2 \quad (15)$$

The binary tree recursive structure of the RAE is shown in Fig. 6.

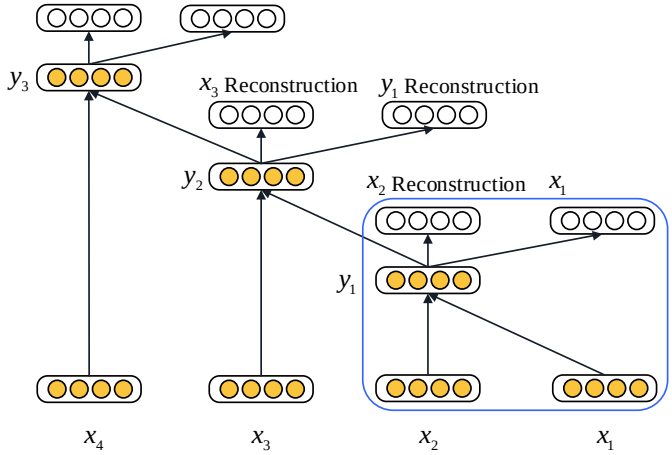

Fig. 6. Binary tree recursive structure of the RAE.

The input feature sequence is paired by (x_0, x_1),(x_1, x_2),..., (x_{n-1}, x_n), calculate the reconstruction error of all the paired nodes, select the pair with the smallest error value to generate the parent node into the input sequence, then remove the pair from the input sequence, and repeat the above steps until only one node remains in the input sequence. At this point, this node is the potential semantic feature of the input sequence.

The video latent semantic extraction model encodes the output of the fully-connected layer of 3D-CNN by unsupervised RAE to obtain video latent semantic features, and its specific model structure is shown in Fig. 7.

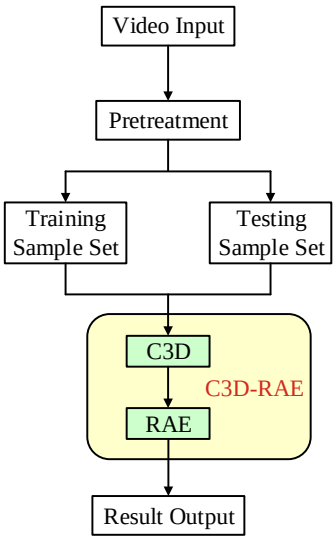

Fig. 7. C3D-RAE model.

B. Sports Culture Recommendation Model

1) Video recommendation based on text topic similarity: This research adopts the Latent Dirichlet Allocation (LDA) topic model to accomplish the topic extraction of video description information. The goal of the LDA topic model is to find the distribution of topics in each document and the distribution of words in each topic. How to calculate similarity specifically? This article calculates the similarity between two videos based on the metadata information of the videos. Simultaneously, using similarity ranking from high to low, obtain the most similar topN of a certain video as an association or similarity recommendation. Although both long and short videos use the same algorithm system, the front-end product form varies due to different video types. Due to the short duration of a single short video, it usually takes a few minutes to play. Therefore, the recommended method for associating short videos is to use information flow to play the original video. The videos associated with it will be played as information streams, which will greatly improve the overall user experience. When training the LDA topic model, the number of topics *K* needs to be given first, and all distributions are expanded based on *K* topics. The specific LDA topic model algorithm is shown in Fig. 8.

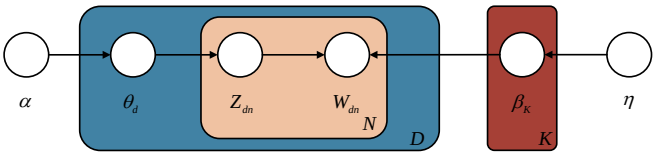

Fig. 8. LDA algorithm.

LDA assumes that the prior distribution of document topics is Dirichlet distribution; that is, for each document *d*, its topic distribution θ_d is:

$$\theta_d = Dirichlet(\alpha) \quad (16)$$

Where α is a hyperparameter in the distribution and is a *k*-dimensional vector. LDA has a premise that the prior distribution of the words in the topic satisfies the Dirichlet distribution; that is, for each topic *k*, the distribution β_k of its words is:

$$\theta_k = Dirichlet(\eta) \quad (17)$$

Wherein, η is a hyperparameter in the distribution and is a *v*-dimensional vector. v here stands for the number of words in the total vocabulary. For the *n*-th word in each document *d*, we can obtain the distribution of its topic number z_{dn} from the topic distribution θ_d as follows:

$$z_{dn} = multi(\theta_d) \quad (18)$$

And for that topic number zdn, the probability distribution of the word wdn is obtained as:

$$w_{dn} = multi(\beta_{z_{dn}}) \quad (19)$$

In the LDA model, there are *M* Dirichlet distributions of document topics, and the corresponding data have *M* multinomial distributions of topic numbers, so ($\alpha \rightarrow \theta_d \rightarrow z_d$) forms Dirichlet-multi conjugate. The posterior distribution of document topics based on Dirichlet distribution can be obtained by the Bayesian inference method. Defined in the *d*-th document, the total number of the *k*-th subject word is $n_d^{(K)}$, then the corresponding multinomial distribution can be expressed as follows:

$$n_d = (n_d^{(1)}, n_d^{(2)}, \ldots, n_d^{(K)}) \quad (20)$$

By using the Dirichlet-multi conjugate, the posterior distribution of β_k is obtained as follows:

$$Dirichlet(\beta_k|\eta + n_k) \quad (21)$$

The specific process of the recommendation algorithm based on LDA topic model is as follows: (1) select the initial topic number *K* value, train the LDA model, and calculate the similarity between each topic in the trained LDA model; (2) increase or decrease the value of *K* according to whether the topic similarity decreases, retrain the LDA model, and calculate the similarity between the topics again; (3) repeat the second step until the optimal *K* value is obtained when the similarity between topic is minimized.

2) Sports culture video recommendation considering multi-modal characteristics: In order to make full use of the multi-source heterogeneous information of videos, this study proposes a video recommendation algorithm considering multi-modal characteristics (RAMM) to improve the accuracy of video recommendations. The RAMM recommendation algorithm is a fusion of three different algorithms which are applied to different problems. When the video belongs to a labelled video, the recommendation based on C3D-CTC is employed. When the video belongs to an unlabelled video, the recommendation based on C3D-RAE is employed, and when the video has description information, the recommendation based on LDA is employed. The specific structure is shown in Fig. 9.

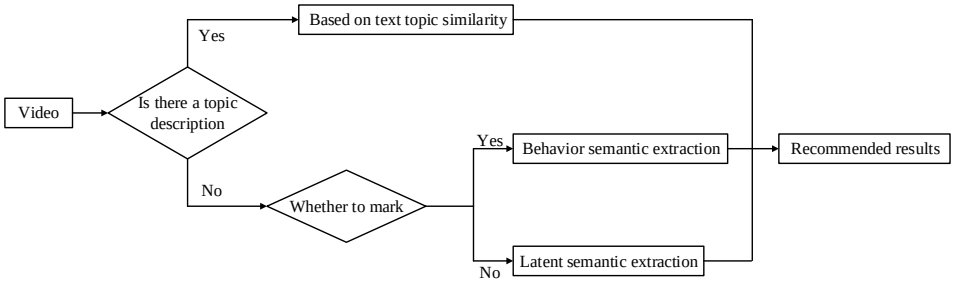

Fig. 9. Recommendation considering multi-modal characteristics.

This study constructs a comprehensive method for calculating video similarity considering the multi-modal characteristics ComSim.

$$Comsim(x,y) = \beta \cdot SimSemantic(x,y) + (1-\beta) \cdot SimTopic(x,y) \quad (22)$$

Where β is a hyperparameter, *SimSemantic* represents semantic similarity, and *SimTopic* represents text similarity. The larger β is, the higher the model's bias towards semantics and, conversely, the higher the bias towards topic information.

IV. Empirical Analysis

The datasets employed in this study contain the UCF-12 dataset and sports video set crawled from the video website Vine.co. The UCF-12 dataset is a specific set of 12 categories of videos excerpted from the UCF-101 video set, mainly including basketball shooting, basketball dunking, bowling, high jump, soccer free throw, cycling, skiing, volleyball dunking, table tennis hitting, pole vaulting, fencing, and rope skipping. Each of these categories consists of 25 groups, each group containing four to seven videos, with the shortest video lasting three seconds and the longest video lasting eight seconds, for a total of 1760 videos. The videos in this dataset have better stability and high similarity of actions in each category, with fewer interfering factors unrelated to the videos. The Vine dataset is crawled from the video website Vine.co sports category videos, consisting of a total of 2400 videos, and because this video set is directly crawled from the website and has not been processed manually, its content similarity is lower than the UCF-12 dataset. The videos contain some interfering factors that are not related to the video content.

Before training the model, the video data needs to be reprocessed. First, all test videos are converted into image sets at 20 frames per second, and for videos whose number of images does not meet a multiple of eight, the last frame is copied and added at the end. Then, to prevent spatial jitter, we scale the images to a uniform size of 112*112. We divide the dataset into a training dataset and a test dataset in a ratio of 9:1 for completing model training and model testing. The topic text similarity and potential semantic similarity of different videos are calculated by the cosine theorem, and the behavioural semantic similarity (*SimBs*) of different videos is defined as follows:

$$SimBs(m,n) = \begin{cases} ecp(m,n), |m| = |n| \\ ncp(m,n), |m| \neq |n| \end{cases} \quad (23)$$

When the lengths of semantic sequences *m* and *n* are equal, the value is the ratio of the total number of semantic equivalents to the length of *m*. When the lengths of semantic sequences *m* and *n* are unequal, the value is the ratio of the length of the common maximum continuous subsequence of *m* and *n* to min (*m*, *n*).

A. Training Parameter Selection

Parameter tuning is crucial to the final performance of the model, and appropriate parameter selection can effectively improve the generalization ability of the model. In order to measure the impact of various parameter changes, this study adopts the control variable method to select the learning rate and batch size of the model. The results are shown in Fig. 10 and Fig. 11.

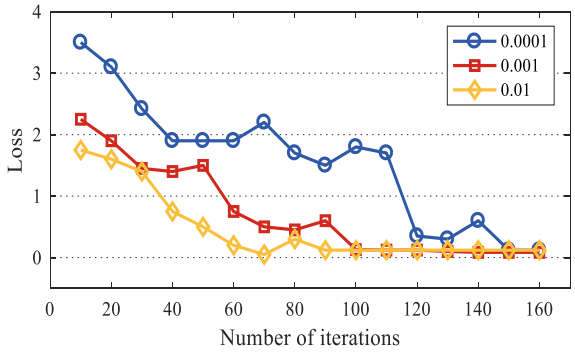

Fig. 10. The impact of learning rate on the Loss function.

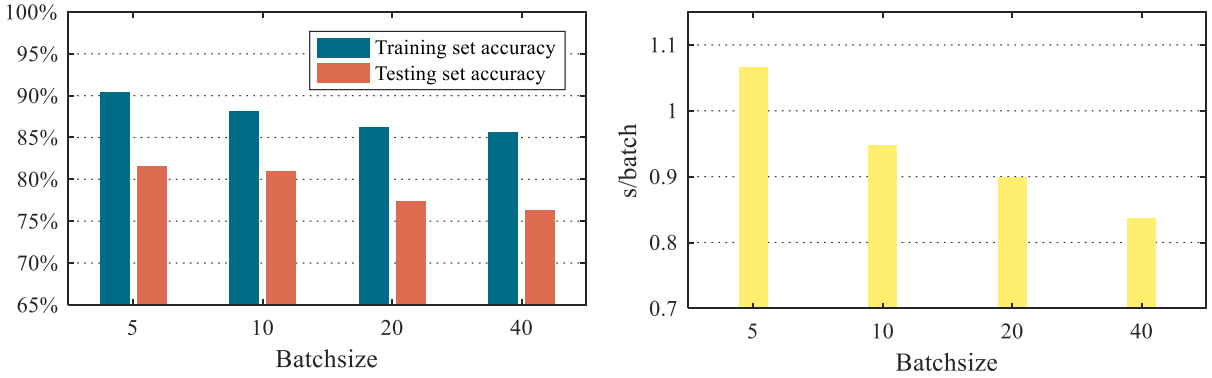

Fig. 11. The impact of batch size on model.

As it can be seen from Fig. 11, when the learning rate is 0.01 and 0.001, the model loss decreases rapidly for a period with the increase of training times and then gradually becomes stable at last. For the learning rate of 0.0001, when the number of training iterations reaches 160, the model is still in a state of oscillation and cannot converge, and the loss value is higher than the other two learning rates. When the learning rate is 0.01, the model reaches the convergence point at the fastest speed, and the loss value is lower than that of 0.0001 but higher than that of 0.001. That is, if the learning rate is too low, the trained model will be more reliable, but it will take longer for the model to reach the convergence point, and it still cannot converge on the basis of the same training time. If the learning rate is too high, the model will reach the convergence point very early. The learning rate is chosen as 0.001 for the model through the experiment of learning rate in this section.

It can be seen from Fig 12 that under the same model and the same data set, different values of Batchsize have an impact on the test accuracy of the model. It can be observed that the larger the value of Batchsize is, the less iteration will be required to process the same amount of data, and the less convergence time will be lost, but the accuracy will also be reduced. Through selection, the batch size value selected in this study is 10.

B. *Impact of Different Algorithms on the Accuracy of Video Semantic Comprehension*

Different algorithms lead to different accuracy of video semantic comprehension of the model, and this study compares the accuracy of the model behavioural semantic comprehension model and latent semantic comprehension model when adopting different algorithms, and the results are shown in Fig. 12 and Fig. 13 respectively.

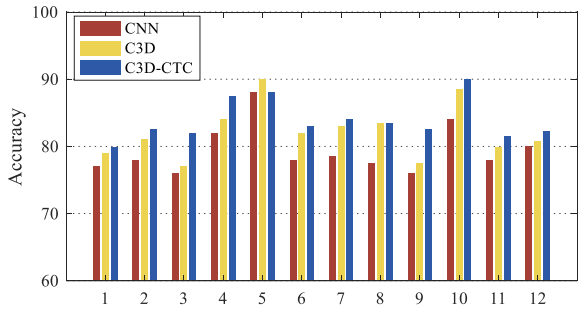

Fig. 12. Semantic comprehension accuracy of different algorithms.

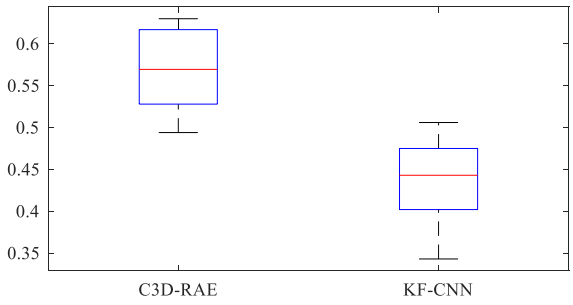

Fig. 13. Cosine similarity of different algorithms.

In Fig. 13, the horizontal coordinates indicate basketball shooting, basketball dunking, bowling, high jump, riding, soccer free throw, skiing, volleyball dunking, table tennis hitting, pole vaulting, fencing, and rope skipping, respectively. It can be seen that the accuracy of the C3D-CTC model is higher than the other algorithms except for the sport of riding, which is due to the fact that the result sequence of riding has a boosting behaviour, which affects the accuracy of C3D-CTC. By analyzing the test results, we found that the categories with no significant differentiation between sub movements achieved higher accuracy on the models in this section than those with significant differentiation between sub movements.

As it can be seen from Fig. 14, the box plot should start from the statistical points, and it can be seen that the C3D-RAE clustering accuracy used in this article is 0.57, while the statistical position of KF-CNN is 0.35. Therefore, C3D-RAE has a clustering accuracy of 22% higher than KF-CNN. Therefore, under unsupervised conditions, C3D-RAE has higher clustering accuracy than keyframe image semantic extraction algorithms and measured through average cosine similarity. Compared with the KF-CNN algorithm using video key frame image semantics, the C3D-RAE model has a larger average similarity between each data type and the center point in the clustering results, and the distribution of the clustering results is more uniform.

C. *Sports Culture Video Recommendation Considering Multi-Modal Characteristics*

Normalized Discounted Cumulative Gain (NDCG@N) is used as an evaluation indicator for sorting results to evaluate the accuracy of sorting. Firstly, to calculate NDCG, we need to calculate Gain, which is the definition of the quality of each result. NDCG adds all the results together to ensure that the higher the overall quality of the list, the larger the NDCG value. At the same time, the design of discounted results in higher weights for higher results. This ensures that the first, more relevant results ranked higher will have a larger NDCG value. From these two points of view, with NDCG as the optimization objective, it ensures that the search engine ranks higher quality results even though the overall quality of the returned results is good. Information retrieval metrics （ MAP@N ） actually' MAP@N The indicator is to measure the UUUU of all users AP@N Average the indicators. Overall, the MAP indicator takes into account both prediction accuracy and relative order, thus avoiding the disadvantage of traditional Precision indicators being unable to depict the relative position differences of recommended products. In the sports culture video recommendation model considering multi-modal characteristics proposed in this study, the top-T videos with the highest similarity are returned as the recommendation results. To select the appropriate β, this study adopts the variable control method to compare the values of NDCG and MAP under different T and β values, and the experimental results are shown in Fig. 14.

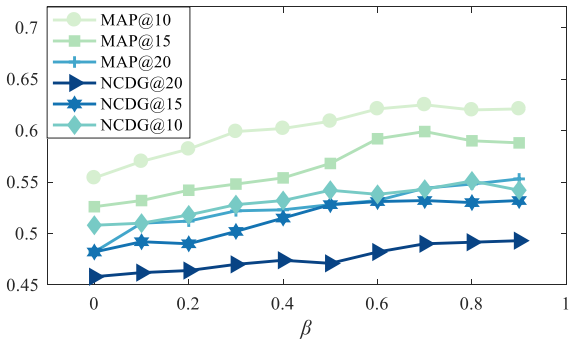

Fig. 14. NDCG@N and MAP@N under different T and β values.

When β is 0, the algorithm similarity calculation value only depends on the topic similarity, so the algorithm accuracy is the lowest. With the gradual increase of β, the recommendation algorithm accuracy gradually rises, indicating that the influence of semantic similarity on the algorithm accuracy is greater than the influence of topic similarity on the algorithm. When β is 0.7, and T is 10 and 15, the algorithm accuracy reaches the peak, and when β is 0.7, and T is 20, the accuracy of the algorithm increases only slightly as the value of β continues to increase, so the hyperparameter β is selected as 0.7. At the same time, it can also be seen that the integrated model considering multi-modal characteristics is more accurate than the video recommendation by only the video semantic model or only the text topic model.

The results of the comprehensive model considering multi-modal characteristics and the recommendation system using other models are shown in Fig. 15.

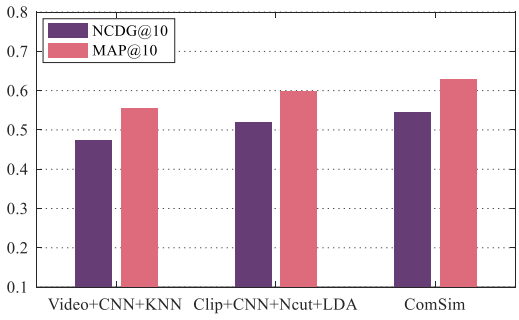

Fig. 15. Comparison of different recommended methods.

The horizontal coordinates from left to right represent Video+CNN+KNN, Clip+CNN+Ncut+LDA and the ComSim algorithm proposed in this paper. Video+CNN+KNN adopts a 3D convolutional neural network to extract the features of all frames of the video and uses the output of the fully connected layer as the video feature representative, then uses the KNN algorithm to find the nearest K neighbours to generate the recommendation list; Clip+CNN+Ncut+LDA adopts 3D convolutional neural network to extract the features of 16 randomly selected frames and adopts NormalizedCut clustering algorithm to generate clustering points, followed by Linear Discriminant Analysis to detect the top-T returned values. It can be seen the accuracy of ComSim is 7.8% higher than Clip+CNN+Ncut+LDA and 15.9% higher than Video+CNN+KNN.

V. Conclusion

After years of development, recommendation systems have made a number of research results in theoretical research and have been widely applied in industry. A well-performing recommendation system can accurately recommend the content that users are interested into users, avoiding a lot of time spent on searching and greatly improving the service experience of users, as well as maximizing the attraction and retention of users, improving the conversion rate of users, and helping the platform achieve profitability.

This study first defines the semantic structure of unstructured video data. On top of that, a 3D convolutional neural network and continuous temporal classification algorithm are used to complete the extraction of sub action semantics and integration of behaviour semantic sequences, and finally, a sports video semantic extraction model for specific sports categories is obtained. In order to solve the problem of the low accuracy of the model for semantic extraction of unlabeled videos, this paper proposes an unsupervised semantic extraction method based on a recursive self-encoder, which uses a recursive self-encoder to construct a semantic spanning tree to complete the compression and integration of semantic sequences and verifies the accuracy of the above two models through experiments. In order to improve the accuracy of the video recommendation algorithm based on single video semantic similarity and topic similarity of this paper, this study integrates video semantic similarity and video topic similarity. It proposes a video recommendation algorithm considering multi-modal characteristics. And it is experimentally demonstrated that the ComSim-based algorithm improves by 7.8% in accuracy over Video+CNN+KNN and 15.9% over Clip+CNN+Ncut+LDA.

However, this article has certain limitations. With the continuous growth of computing resources and dataset size, unsupervised semantics based on C3D-RAE have dominated many tasks. However, the continuous increase in unsupervised semantic depth has also introduced training challenges. Traditional supervised training methods only use supervision in the last layer, allowing errors to propagate from the last layer to the hidden layer, leading to intermediate layer optimization problems such as gradient vanishing. In the future, it is necessary to increase the amount of data, improve the robustness of the model, and avoid overfitting. At present, data augmentation is mainly applied to image data, and there are no good methods for other types of data such as text.

References

[1] Y. Deldjoo, M. Elahi, P. Cremonesi, F. Garzotto, P. Piazzolla, and M. Quadrana, "Content-based video recommendation system based on stylistic visual features," J Data Semant, vol. 5, pp. 99–113, 2016.

[2] P. Covington, J. Adams, and E. Sargin, "Deep neural networks for youtube recommendations," in Proceedings of the 10th ACM conference on recommender systems, 2016, pp. 191–198.

[3] Y. Zhang, C. Yin, Q. Wu, Q. He, and H. Zhu, "Location-aware deep collaborative filtering for service recommendation," IEEE Trans Syst Man Cybern Syst, vol. 51, no. 6, pp. 3796–3807, 2019.

[4] J. Qi, J. Du, S. M. Siniscalchi, X. Ma, and C.-H. Lee, "On mean absolute error for deep neural network based vector-to-vector regression," IEEE Signal Process Lett, vol. 27, pp. 1485–1489, 2020.

[5] H. Liu, X. Zhao, C. Wang, X. Liu, and J. Tang, "Automated embedding size search in deep recommender systems," in Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2020, pp. 2307–2316.

[6] K. K. Jena et al., "Neural model based collaborative filtering for movie recommendation system," International Journal of Information Technology, vol. 14, no. 4, pp. 2067–2077, 2022.

[7] X. Zheng, G. Zhao, L. Zhu, J. Zhu, and X. Qian, "What you like, what I am: Online dating recommendation via matching individual preferences with features," IEEE Trans Knowl Data Eng, vol. 35, no. 5, pp. 5400–5412, 2022.

[8] Q. Li, S. Chu, N. Rao, and M. Nourani, "Understanding the Effects of Explanation Types and User Motivations on Recommender System Use," in Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, 2020, pp. 83–91.

[9] S. Natarajan, S. Vairavasundaram, S. Natarajan, and A. H. Gandomi, "Resolving data sparsity and cold start problem in collaborative filtering recommender system using linked open data," Expert Syst Appl, vol. 149, p. 113248, 2020.

[10] J. Bobadilla, Á. González-Prieto, F. Ortega, and R. Lara-Cabrera, "Deep learning feature selection to unhide demographic recommender systems factors," Neural Comput Appl, vol. 33, no. 12, pp. 7291–7308, 2021.

[11] S. Tao, C. Shen, L. Zhu, and T. Dai, "SVD-CNN: A convolutional neural network model with orthogonal constraints based on SVD for context-aware citation recommendation," Comput Intell Neurosci, vol. 2020, 2020.

[12] S. Meng et al., "Privacy-aware factorization-based hybrid recommendation method for healthcare services," IEEE Trans Industr Inform, vol. 18, no. 8, pp. 5637–5647, 2022.

[13] G. Xu et al., "TT-SVD: An efficient sparse decision-making model with two-way trust recommendation in the AI-enabled IoT systems," IEEE Internet Things J, vol. 8, no. 12, pp. 9559–9567, 2020.

[14] X. Zhou, W. Liang, I. Kevin, K. Wang, and S. Shimizu, "Multi-modality behavioral influence analysis for personalized recommendations in health social media environment," IEEE Trans Comput Soc Syst, vol. 6, no. 5, pp. 888–897, 2019.

[15] C.-C. Hsu and M.-Y. Yeh, "A general framework for implicit and explicit social recommendation," IEEE Trans Knowl Data Eng, vol. 30, no. 12, pp. 2228–2241, 2018.

[16] W. Liu, Z. Wang, X. Liu, N. Zeng, Y. Liu, and F. E. Alsaadi, "A survey of deep neural network architectures and their applications," Neurocomputing, vol. 234, pp. 11–26, 2017.

[17] M. S. Tajbakhsh and J. Bagherzadeh, "Semantic knowledge LDA with topic vector for recommending hashtags: Twitter use case," Intelligent Data Analysis, vol. 23, no. 3, pp. 609–622, 2019.

[18] U. Yadav, N. Duhan, and K. K. Bhatia, "Dealing with pure new user cold-start problem in recommendation system based on linked open data and social network features," Mobile Information Systems, vol. 2020, pp. 1–20, 2020.

[19] A. N. Nikolakopoulos, V. Kalantzis, E. Gallopoulos, and J. D. Garofalakis, "EigenRec: generalizing PureSVD for effective and efficient top-N recommendations," Knowl Inf Syst, vol. 58, pp. 59–81, 2019.

[20] A. Salah, N. Rogovschi, and M. Nadif, "A dynamic collaborative filtering system via a weighted clustering approach," Neurocomputing, vol. 175, pp. 206–215, 2016.

[21] C. Hewitt and H. Gunes, "Cnn-based facial affect analysis on mobile devices," arXiv preprint arXiv:1807.08775, 2018.

[22] Y. Li, H. Wang, H. Liu, and B. Chen, "A study on content-based video recommendation," in 2017 IEEE International Conference on Image Processing (ICIP), IEEE, 2017, pp. 4581–4585.

[23] X. Li et al., "Learnable aggregating net with diversity learning for video question answering," in Proceedings of the 27th ACM international conference on multimedia, 2019, pp. 1166–1174.

[24] C. Fan, X. Zhang, S. Zhang, W. Wang, C. Zhang, and H. Huang, "Heterogeneous memory enhanced multimodal attention model for video question answering," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 1999–2007.

[25] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in Proceedings of the IEEE international conference on computer vision, 2015, pp. 4489–4497.

[26] L. Lu, X. Liang, G. Yuan, L. Jing, C. Wei, C. Cheng, "A study on the construction of knowledge graph of yunjin video resources under productive conservation," Heritage Science, 11, pp. 1-16, 2023

[27] L. Yang, "Feature Extraction of English Semantic Translation Relying on Graph Regular Knowledge Recognition Algorithm," Informatica, 2023, vol. 47, no. 8, pp. 1.

[28] C. Guo, "Research on sports video retrieval algorithm based on semantic feature extraction", Multimedia Tools and Applications, 2023, vol. 82, no. 14, pp. 21941-21955.