Cross-Modal Hash Retrieval Model for Semantic Segmentation Network for Digital Libraries

Siyu Tang*, Jun Yin
College of Information Engineering, Shanghai Maritime University, Shanghai, 201306, China

***Abstracts*—The study optimizes the classic hash retrieval model by introducing Word2vec network and full convolutional neural network, with the goal of addressing the issues of low retrieval efficiency and poor retrieval accuracy of the traditional digital library retrieval system. The study extracts the secondary features in the graphic features and optimizes the loss function to design a cross-modal hash retrieval model based on semantic segmentation network. The experimental results indicated that the CPU utilization, memory utilization, number of concurrent operations, and operation success rate of this model were 45.4%, 36.3%, 92, and 96.7%, respectively. Its CPU utilization and memory utilization were significantly lower than other models, while the number of concurrent operations and operation success rate were significantly lower than other models. It proved its high efficiency of resource utilization and its reliability. When the amount of processed data was 100, the average response time of cross-modal hash retrieval model was 10.2s, which was significantly lower than other models, proving its high operational efficiency. The cross-modal hash retrieval model proposed in the study has better performance and provides technical support for cross-modal retrieval in digital library and also provides ideas for information retrieval in other fields.**

***Keywords*—Digital library; hash retrieval; semantic segmentation; word2vec; fully convolutional neural network**

I. Introduction

A distributed information system that processes and stores a range of textual and graphic multimedia creations is called a digital library. It makes use of digital technology. With the aid of digital technology, it may store information resources from many carriers and geographical locations, enabling querying and distribution throughout the area and object-oriented network [1-3]. As digital technology has advanced recently, public libraries now have access to intelligent service ways. In addition, digital libraries offer tailored suggestion services to cater to the varied demands of their patrons, substantially improving their overall service experience [4]. However, book information is constantly expanding, and users face problems such as inefficiency in graphic retrieval through traditional means. Therefore, many scholars have proposed a series of methods to address this problem.

Zeng Z et al. created a visual search model based on a bag-of-words model (BWM) and various semantic associations in order to build a mobile visual search service system for digital libraries. They extracted local and global features from an image using hue, saturation, and brightness variations that are scale invariant. The results showed that the model performs better in searches [5]. A base three indexing approach based on semantic visual indexing was created by Krishnaraj N et al. to address the issue of semantic gaps between queries and disparate semantics in large-scale databases. They demonstrated that the model was accurate to 83% by using an interactive optimization model to determine the joint semantic and visual descriptor space and by combining a design model with a semantic visual joint space model [6]. Khan U A et al. designed a hybrid feature descriptor-based retrieval method in order to reduce the reliance on text annotation based digital book retrieval system. It combined genetic algorithm and support vector machine for digital book retrieval in multi-category scenarios and the results showed that the method retrieved better on four standard datasets [7]. Yu H et al. designed an image retrieval method based on typical correlation analysis and domain adaptation for the poor retrieval performance of traditional digital library retrieval methods. It transferred classifiers trained on known datasets to new datasets and combined multi-source domain data to generate classifiers on the target domain, and the findings revealed that the approach was more effective in image retrieval [8]. To achieve better cross-modal retrieval of digital libraries, He S et al. developed a cross-modal retrieval method based on category-aligned adversarial learning. By using category information to create a shared identity space, it compared samples from several modalities. The strategy outperformed other approaches on four benchmark datasets, according to the findings [9]. Wang X et al. designed a deep learning based cross-media semantic search framework for digital library to improve and enhance the multimedia retrieval system in digital library. It integrated book processing, big data and deep learning and applied them to the integration of digital library, and the results showed that the framework improved the overall search performance by 11.53% over the suboptimal method [10]. Ahmadi A et al. designed an image semantic retrieval method combining ontology and composite modeling for applying ontology to information retrieval in digital library. It was able to handle linguistic errors such as ambiguity and structural errors in digital books using semantic web organization. The results indicated that the method was able to improve the accessibility and retrieval accuracy of digital library [11]. To address the semantic gap between multimodal input text and digital library images, Rong H et al. developed a multimodal retrieval approach based on contextually relevant and irrelevant attention alignment augmentation. To capture the shared semantics across several modalities, it employed a semantic alignment augmentation approach. The approach was able to successfully capture the shared semantics between text and images, according to the results [12]. Zhen Z et al. designed a shared cloud service platform based on big data in order to effectively integrate digital book resources. It analyzed the features of digital books

and the factors affecting these features. The findings revealed that the platform was more effective in the dissemination of digital book resources [13]. To demonstrate the role of visual and verbal pre-trained models in digital book retrieval, Baldrati A et al. designed a combinatorial retrieval method based on comparative learning and task-oriented features. It integrated bimodal information by training a combiner network and provided combinatorial features for performing retrieval. The results revealed that the method outperforms existing methods [14].

In summary, scholars have obtained many results in the field of digital library graphic retrieval. Nevertheless, none of these techniques account for the potential semantic association that may exist between various graphic data, which results in inadequate retrieval accuracy. Fully convolutional neural network (FCN) is the first network for pixel-level prediction. Since its proposal, it has become the basic framework for semantic segmentation [15]. However, Word2Vec is a popular natural language processing technique that is able to represent each word as a high-dimensional vector and represent the semantic relationship between words by the similarity between the vectors [16]. In view of this, the study uses Word2Vec with FCN to optimize deep cross-modal hashing (DCMH) and considers the extraction of secondary features. Concurrently, the research enhanced the efficiency and precision of retrieval by designing an IDCMH model and optimizing the loss function. The study is novel because it combines FCN with Word2Vec, which makes full use of both the semantic vector representation of word and the pixel-level semantic information of images. This improves the model's capacity to capture the semantic link between textual and graphic data. This enhances the model's capacity to identify complex data by accounting for the extraction of secondary features.

There are four parts to the study. An overview of the digital library retrieval model's background and the state of domestic and international research are covered in Section I. The building of a multi-modal digital library retrieval system and the functional design of an associated retrieval model comprise the Section II. The proposed model's performance analysis and the practical application effect analysis make up Section III. The study's weaknesses and discussions are highlighted in Section IV. Finally, Section V concludes the paper.

II. METHODS

This section focuses on the research methodology. Section I shows the construction of multi-modal digital library retrieval system. Section II shows the function design for cross-modal hash retrieval model through semantic segmentation network. Section III shows the extraction of secondary features and the design of optimization techniques for the loss function.

A. Design of Multi-Modal Digital Library Retrieval System

Traditional libraries mainly rely on view resources such as paper books. These resources not only take up a lot of space, but also are difficult to manage and not convenient enough to use. With the arrival of the Internet era, people's reading habits have shifted and they are more willing to read on online platforms [17]. Therefore, digital library came into being. The study designs a three-tier digital library graphic retrieval system based on B/S architecture. The technical route for the realization of this three-tier architecture digital library system is shown in Fig. 1.

In Fig. 1, the system includes a representation layer (RL), a business layer (BL), and a data layer (DL). Among them, the RL is used to interact directly with the user, and it is mainly used for data entry, display, and query. The interface of this layer is realized by using the Web page, by encapsulating the user's request as an HTTP request and sending it to the BL. The BL receives the request from the RL through the API interface and performs operations such as syntactic dependency analysis, segmentation, and entity recognition. After processing the request from the RL, the data is transferred to the DL. The interface of the DL is realized using the JDBC interface, and the main functions of this layer are database connection and SQL statement execution. This layer is able to process and return requests from the BL. The overall function of the designed system is shown in Fig. 2.

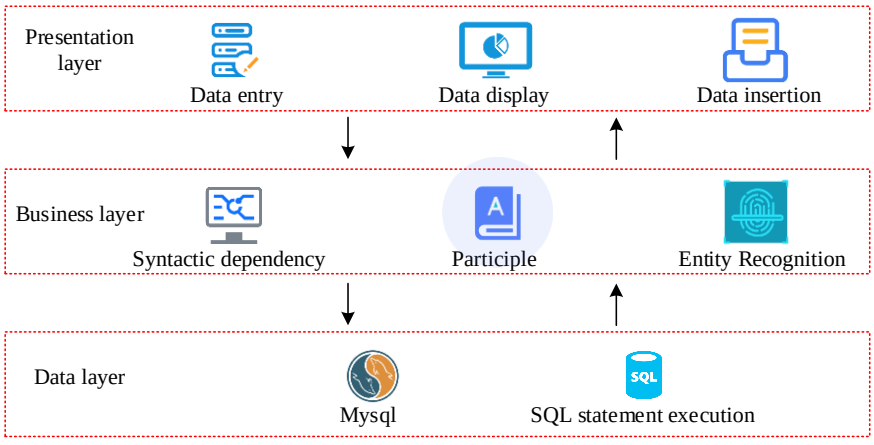


Fig. 1. Technical roadmap of three-tier architecture digital library system.

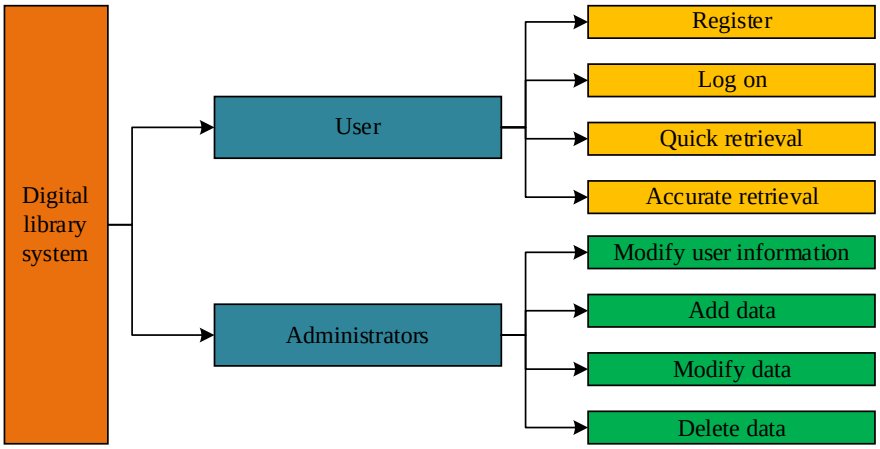


Fig. 2. Overall functionality of the designed system.

In Fig. 2, the designed digital library system mainly consists of two functional modules: user and administrator. Among them, the user module is capable of performing operations such as user registration, login, and search. When a user logs in for the first time, he/she must register an account. After that, the user must enter the correct user name and password each time he/she logs in to enter the digital library system. When the user conducts a search, each user has different needs, so the study designed two search methods, one for quick search and the other for accurate search. When a user performs a quick search, the system directly performs the search operation, while when a user performs a precise search, the system calls the algorithm module to perform the search. The administrator module can perform administrator login, user information modification, data processing and other operations. The study distinguishes between the user module and the administrator module through a User table and an Administrator table. Each table has its primary key and HTML5 role attribute is set to describe the role of different elements. When the role attribute is assigned a value of 0, it represents a user. When the role attribute is assigned to 1, it means administrator. In order to store the images and texts in the digital book, Text table and Picture table with unique primary key are created. Meanwhile, in order to store the final hash representation obtained, Text_feature table and Pic_feature table with unique primary key are created. Moreover, a foreign key is set for image feature (IF) and text feature respectively for joint query.

B. A Cross-Modal Deep Hash Model Based on Semantic Segmentation Networks

In digital library, with the explosive growth of different types of information, there is heterogeneity between the semantic features of these information. The way to measure the semantic similarity between different types of data has become the main challenge of current cross-modal retrieval in digital library [18]. Therefore, it is investigated to retrieve image and text features in digital library system through DCMH, and introduce Word2vec network with FCN to improve the DCMH algorithm. The algorithm uses Tiny_ViT as the backbone network. The data is categorized into IFs and text features based on semantic features, and the important features are extracted by an FCN IF extraction module and a Word2vec text feature extraction module. An improved deep cross-modal hashing (IDCMH) is proposed by integrating all the extracted features into the hash representation through a fully connected layer (FCL). The technical route of this algorithm implementation is shown in Fig. 3.

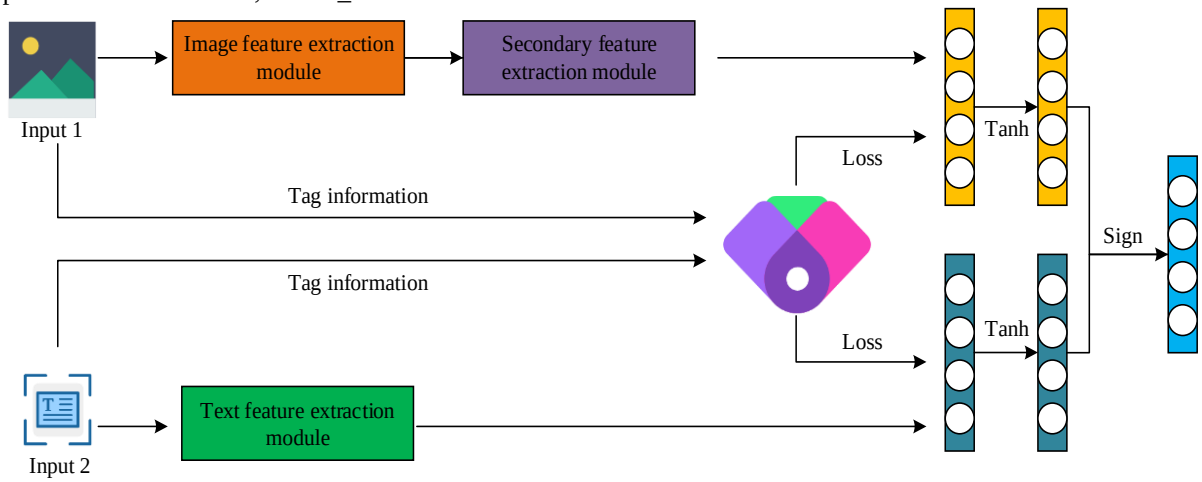


Fig. 3. Technical roadmap of cross-modal deep hashing algorithm.

The Word2vec text feature extraction module, the hash representation module, and the FCN IF extraction module comprise the three components of the IDCMH shown in Fig. 3. FCN is an end-to-end model that is able to learn mapping directly from image pixels to segmentation masks. Word2vec is a word vector model, which is able to transform words in text into real number vectors, and then perform operations such as text categorization and sentiment analysis [19-20]. The input image data and text data are pre-processed. The pre-processed data is input into FCN IF extraction module and Word2vec text feature extraction module. The training sample expressions for IFs and text features are shown in Eq. (1).

$$\begin{cases} X = \{x_i\}_{i=1}^{n} \\ Y = \{y_i\}_{i=1}^{n} \end{cases} \tag{1}$$

In Eq. (1), X represents the training samples of Ifs. Y represents the training samples of text features. i is the i th sample. n is the samples. IF extraction is performed first. Due to the different dimensions of the features in different network layers in the network, these features need to be mapped into a feature space. Therefore, the study inputs all features into a FCL and processes the features by linear transformation operations. The calculation method is shown in Eq. (2).

$$Q_t = W_q Q_m + b_q \tag{2}$$

In Eq. (2), Q_t represents the linearly transformed features. Q_m represents the secondary features. W_q is the weight of the FCL. b_q is the bias of the FCL. Then the processed features are normalized and calculated as shown in Eq. (3).

$$q_n^i = \frac{q_t^i - \mu_i}{\sqrt{\sigma_i^2 + \varepsilon}} \tag{3}$$

In Eq. (3), q_n^i is the normalized feature. μ_i is the mean of the i th feature vector. σ_i^2 is the variance of the i th feature vector. ε represents a real number to avoid a denominator of 0. The next step is to access another connectivity layer to fuse the different local features and utilize the ReLU function as an activation function. The calculation is shown in Eq. (4).

$$Q_b = \mathrm{ReLU}(W_a Q_n + b_a) \tag{4}$$

In Eq. (4), Q_b represents the fused features. ReLU(·) represents the activation function. W_a represents weights. b_a represents the bias. Fig. 4 depicts the model's flow for producing features.

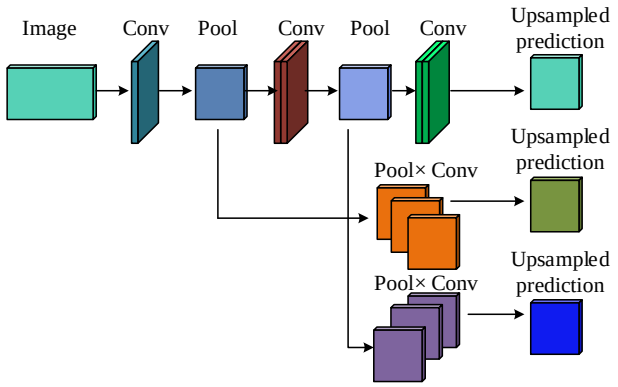


Fig. 4. The process of generating features from the model.

For text feature extraction, the continuous BWM of word2vec is used, which is able to characterize the semantic information of the words in terms of word vectors by learning the text. The continuous BWM consists of three layers: input layer, projection layer and output layer. The features to be extracted are fed into the input layer and converted into word vectors. The contextual representation of word vector is shown in Eq. (5).

$$Z = (z^{c-m}, \ldots, z^{c-1}, z^{c+1}, z^{c+m}) \tag{5}$$

In Eq. (5), Z represents the contextual representation of the word vector. m represents the number of windows. c represents the center word.

C. Improved Cross-Modal Deep Hash Model Considering Secondary Features

Since the information obtained from the digital library does not only contain important features, other secondary features that are easily ignored can also affect the performance of model extraction. Therefore, to guarantee that the extracted features are complete, a secondary feature extraction module is built. To produce the overall IF representation, the features extracted from the IF extraction module are fused with the IFs that are overlooked, which are first extracted using the secondary feature extraction module. The feature fusion process is shown in Fig. 5.

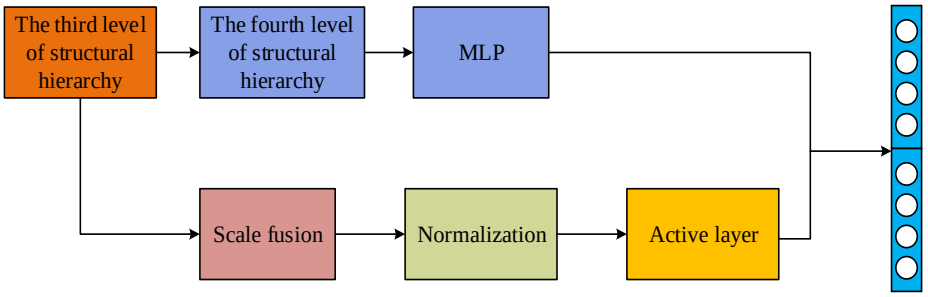


Fig. 5. Feature fusion process.

In Fig. 5, secondary features are mainly extracted by three steps including scale fusion, normalization, and feature activation. The word vector matrix is obtained by multiplying all the input word matrices with the effective coding vectors of each word, and all the word vector matrices are fused and vector averaged to represent the center word vector. The calculation is shown in Eq. (6).

$$\hat{v} = \frac{v_{c-m} + ... + v_{c+m}}{2m} \tag{6}$$

In Eq. (6), $\hat{v}$ represents the center word vector. Then it is processed using Softmax activation function and converted into probability values. The calculation is shown in Eq. (7).

$$\hat{y} = \text{Softmax}(z) \tag{7}$$

In Eq. (7), g_t represents the generated probability value. The text word vector is obtained by multiplying the obtained probability value with the input word matrix. Finally, the Ifs are hashed with the text features through the symbol function. The calculation method is shown in Eq. (8).

$$\begin{cases} F = Sign(Q_b) \\ G = Sign(\hat{y}) \end{cases} \tag{8}$$

In Eq. (8), F represents the hash code of Ifs. G represents the hash code of text features. $Sign(\cdot)$ represents the symbol function. To further deal with the heterogeneity between different features, the study is conducted by optimizing the loss function. Firstly, the smooth L1 loss function is utilized to eliminate the differences between different labels and hash representations. Eq. (9) displays the computation.

$$L(x,y) = \begin{cases} -\dfrac{0.5(x-y)^2}{\gamma}, |x-y| < \gamma \\ |x-y| - 0.5 * \gamma, |x-y| \geq \gamma \end{cases} \tag{9}$$

In Eq. (9), L represents the smooth L1 loss function. γ represents the difference between the label and the hash representation. Then cosine similarity with Jaccard coefficient is utilized to represent the semantic loss from image to text. The calculation is shown in Eq. (10).

$$J_{in} = \frac{1}{n \times n_{batchsize}} \sum_{i=1}^{n} \sum_{j=1}^{n_{batchsize}} \lambda S_{ij}^{vt}, cos\left(Q_b, \hat{y}\right)_L \tag{10}$$

In Eq. (10), J_{in} represents the semantic loss from image to text. Meanwhile, the loss function of Ifs is shown in Eq. (11).

$$J_i^v = \frac{1}{n \times n_{batchsize}} \sum_{i=1}^{n} \sum_{j=1}^{n_{batchsize}} \lambda S_{ij}^{vv}, cos\left(Q_{b_i}, Q_{b_j}\right)_L \tag{11}$$

The loss function for text features is shown in Eq. (12).

$$J_i^t = \frac{1}{n \times n_{batchsize}} \sum_{i=1}^{n} \sum_{j=1}^{n_{batchsize}} \lambda S_{ij}^{tt}, cos\left(\hat{y}_i, \hat{y}_j\right)_L \tag{12}$$

The overall loss can be obtained by IF loss and text feature loss, which is calculated as shown in Eq. (13).

$$J_O = J_i^v + J_i^t \tag{13}$$

The distance of the generated hash representation from different hash codes is minimized to quantize the hash code. The calculation is shown in Eq. (14).

$$J_q = \frac{1}{2nk}\left(\sum_{i=1}^{n}\sum_{j=1}^{k}(h_{ij} - q_{ij}) + \sum_{i=1}^{n}\sum_{j=1}^{k}(b_{ij} - \hat{y}_{ij})\right) \tag{14}$$

In Eq. (14), h_{ij} is the actual hash code. q_{ij} is the j th value of IF q_i. $\hat{y}_{ij}$ is the j th value of text feature $\hat{y}_i$. Finally, the final objective function can be obtained by synthesizing the three. Eq. (15) presents the calculating procedure.

$$\min_{B, W_v, W_t} J = J_{in} + \alpha J_O + \beta J_q \tag{15}$$

In Eq. (15), $\min_{B, W_v, W_t} J$ represents the final objective function. α and β represent the weight parameters. The above improved cross-modal hash retrieval model can quickly accomplish the retrieval of digital library. The flow is shown in Fig. 6.

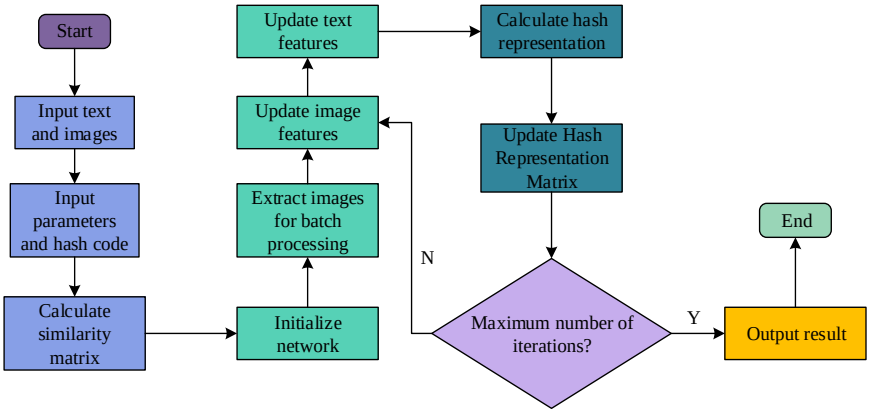


Fig. 6. Improved cross-modal hash retrieval model.

III. RESULTS

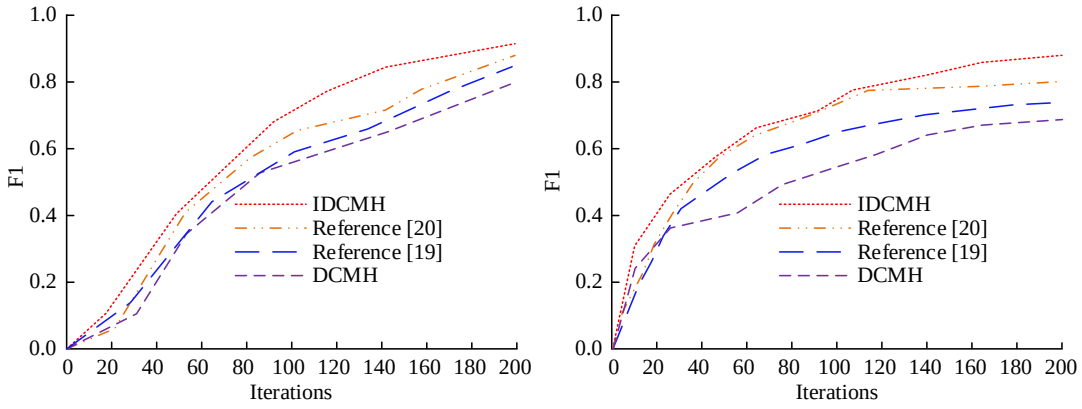
This section deals with the experimental results of the designed IDCMH retrieval model for digital library. Section I shows the performance analysis of the IDCMH retrieval model. The usefulness of the developed IDCMH retrieval model in real-world situations is examined in Section II.

A. IDCMH Model Performance Analysis

To verify the performance of the designed IDCMH model, the initial learning rate of the model is set to 0.0001, the batch size is set to 32, and the maximum iterations is 200. Firstly, the F1 metrics, the MIRFlickr dataset and the NUS-WIDE dataset are introduced, and the F1 values of the model on the different datasets are calculated respectively. Moreover, the F1 values are compared with those of the DCMH model, the model in study [19] and the model in study [20]. The MIRFlickr dataset is selected by Flickr for multimedia retrieval and contains one million images, each with text information, aesthetic quality annotations, and EXIF metadata. The NUS-WIDE dataset is an image dataset with network label annotations created by the Media Search Laboratory of the National University of Singapore, which includes 269648 images from websites with 5018 different labels. Fig. 7 presents the findings.

In Fig. 7(a), on the MIRFlickr dataset, the F1 values of all four models increase with the number of iterations. When the maximum iterations is reached, the F1 value of the DCMH model, study [19]' method, study [20]' method and IDCMH model is 0.803, 0.841, 0.887, and 0.936. In Fig. 7(b), the trend of the F1 values of the models on the NUS-WIDE dataset is the same as the trend of the F1 values on the MIRFlickr dataset with the same trend. When the maximum iterations is reached, the F1 values of the four models, DCMH, study [19], study [20] and IDCMH, are 0.682, 0.761, 0.804 and 0.877, respectively. In summary, the F1 values of the proposed IDCMH model are greater than the other three models in different datasets, which proves its better performance. It also proves its generalization ability and robustness. For validating the retrieval accuracy of the proposed IDCMH model, the study calculates the accuracy of the IDCMH model on the MIRFlickr dataset and the NUS-WIDE dataset when the hash code is 16bits, 32bits, and 64bits, respectively. The accuracy is also compared with that of DCMH, study [19], and study [20]. Table I presents the findings.

In Table I, in the MIRFlickr dataset, when the hash code is 16bits, the accuracy of the IDCMH model is 0.948. When the hash code is 32bits, the accuracy of the IDCMH model is 0.952. The IDCMH model's accuracy when the hash code is 64 bits is 0.923, which is a substantial improvement above the accuracy of other models. In the NUS-WIDE dataset, when the hash code is 16bits, 32bits, and 64bits, the accuracy of IDCMH model is 0.932, 0.936, and 0.934, respectively. It suggests that the IDCMH model offers notable performance advantages and is more efficient at executing graphic retrieval from digital libraries. Precision-recall (PR) curves for the MIRFlickr and NUS-WIDE datasets are computed for the IDCMH model, respectively, and compared to the PR curves of the other models in order to confirm the model's overall performance. Fig. 8 presents the findings.



(a) F1 values of different models on the MIRFlickr dataset

(b) F1 values of different models on the NUS-WIDE dataset

Fig. 7. F1 values of different models on the MIRFlickr and NUS-WIDE datasets.

TABLE I. PRECISION OF DIFFERENT MODELS ON MIRFLICKR DATASET AND NUS-WIDE DATASET

Model		Data set	Hash code		
			16bits	*32bits*	*64bits*
Precision	DCMH	MIRFlickr	0.702	0.618	0.622
		NUS-WIDE	0.687	0.624	0.643
	Reference [19]	MIRFlickr	0.849	0.875	0.769
		NUS-WIDE	0.837	0.831	0.853
	Reference [20]	MIRFlickr	0.871	0.883	0.887
		NUS-WIDE	0.840	0.876	0.868
	IDCMH	MIRFlickr	0.948	0.952	0.923
		NUS-WIDE	0.932	0.936	0.934

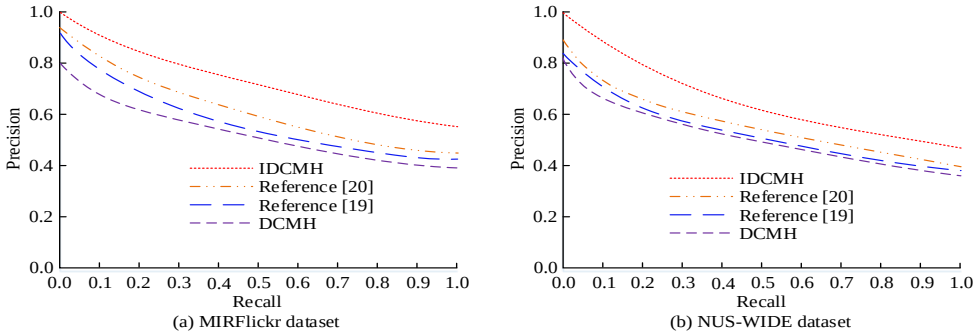


Fig. 8. PR curves of different models on MIRFlickr and NUS-WIDE datasets.

In Fig. 8(a), in the MIRFlickr dataset, the precision of all four models tends to decrease with the increase of recall. When the recall rate (RR) is 0.8, the precision of IDCMH model, study [20]' method, study [19]' method and DCMH model is 0.61, 0.47, 0.43 and 0.41. In Fig. 8(b), in the NUS-WIDE dataset, the trends of the PR curves of the four models are consistent with the trends of the PR curves in the MIRFlickr dataset. When the RR is 0.8, the RRs of the four models are 0.46, 0.40, 0.39, and 0.38, respectively. In summary, in different datasets, when the same RR is taken, the precision rate of IDCMH model is greater than that of other models. Meanwhile, it can be found that the PR curves of IDCMH model are always above the PR curves of other models, which proves that the comprehensive performance is better. In addition, it indicates that the IDCMH model has strong ability to handle large-scale datasets, proving its scalability.

B. *Practical Application Effect Analysis*

To validate the effectiveness of the designed IDCMH model in practical applications, the study is conducted in an operating environment equipped with Intel Core i9-9900k central processor, 128G running memory, 1TB hard disk, and Windows 10, and the simulation analysis is performed using Python 3.7 software. Firstly, ROUGE scores are introduced to calculate the ROUGE1 scores and ROUGE2 scores of the DCMH model in the MIRFlickr dataset, respectively. The results are also compared with the ROUGE1 score and ROUGE2 score of DCMH, study [19]' method and study [20]' method. Fig. 9 presents the findings.

In Fig. 9(a), the ROUGE1 scores of the different models increase with the number of iterations. When the maximum iterations are reached, the ROUGE1 score of IDCMH, study [20]' method, study [19]' method and DCMH is 0.85, 0.82, 0.78 and 0.76. In Fig. 9(b), the ROUGE2 scores of the models with the iteration number of the trend is consistent with the trend of the ROUGE1 score of each model. When the maximum iterations is reached, the ROUGE2 scores of the four models are 0.96, 0.90, 0.86 and 0.82, respectively. It can be found that the ROUGE1 score and ROUGE2 score of the IDCMH model are significantly larger than those of other models, which proves that it performs better in measuring the similarity of graphical features and repetition. To further validate the performance of IDCMH model, the study introduces four metrics, namely, CPU utilization, memory utilization, number of concurrent operations, and operation success rate, and calculates the metric values of different models respectively. Table II presents the findings.

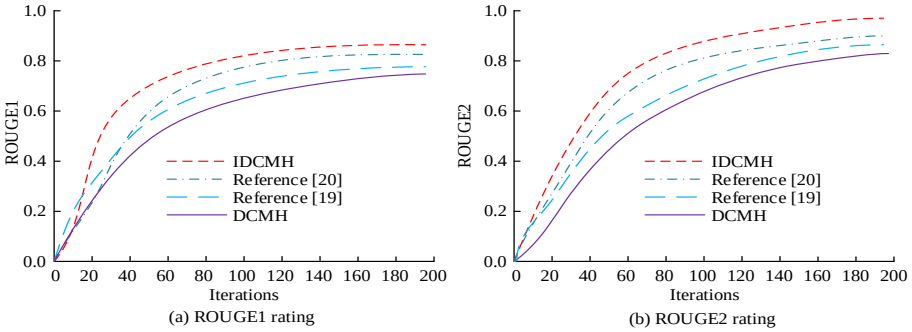


Fig. 9. ROUGE1 and ROUGE2 ratings for different models.

TABLE II. CPU UTILIZATION, MEMORY UTILIZATION, CONCURRENT OPERATIONS, AND OPERATION SUCCESS RATE OF DIFFERENT MODELS

Test indicators	Model			
	DCMH	***Reference [19]***	***Reference [20]***	***IDCMH***
Cpu utilization (%)	73.6	68.6	66.4	45.4
Memory utilization (%)	80.2	77.1	74.3	36.3
Number of concurrent operations	22	46	57	92
Operation success rate (%)	86.5	88.3	90.2	96.7

In Table II, the CPU utilization of IDCMH model is 45.4%, the memory utilization is 36.3%, the number of concurrent operations is 92, and the operation success rate is 96.7%. The analysis indicates that the CPU utilization and memory utilization of the IDCMH model are much lower than the values of the two indexes of the other models. However, the number of concurrent operations and operation success rate are significantly higher than those of the other three models. The aforementioned findings demonstrated the stability and dependability of the IDCMH model and showed that its resource utilization efficiency is excellent. Lastly, the variation of the suggested IDCMH model's response time with data amount is calculated and compared with other models' results to confirm the model's speed of operation. The results are shown in Fig. 10.

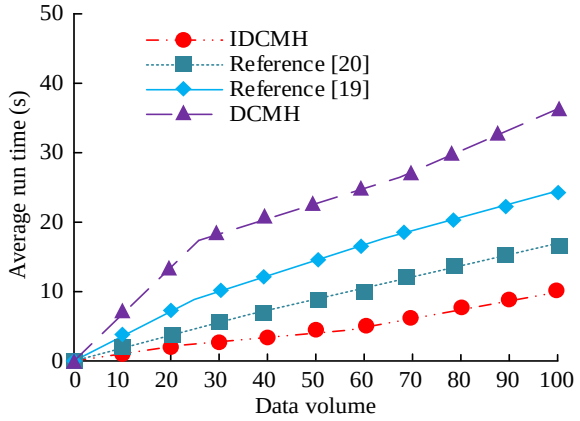


Fig. 10. The variation of response time with data volume, etc.

In Fig. 10, as the amount of processed data increases, the average response time of the proposed IDCMH model is 10.2s when the amount of processed data is 100. The average response time of the method in study [20] is 17.4s, that of the method in study [19] is 24.6s, and that of the DCMH model is 36.1s. It can be found that compared to the DCMH model, study [19] and study [20], the average response time of IDCMH model has decreased by 71.7%, 58.5% and 41.3%, respectively. It is proved that the IDCMH model runs more efficiently.

IV. Discussion

The research aims to explore a resource retrieval method for digital libraries, in order to overcome the disadvantages of slow retrieval speed and low retrieval accuracy of traditional digital library resource retrieval techniques. This research presents the development of a three-tier digital library image and text retrieval system based on the B/S architecture, utilizing Tiny_ViT as the underlying network and integrating FCN and word2vec models to enhance DCMH, culminating in the proposal of an IDCMH model. The results showed that on the MIRFlickr dataset, the F1 value of the IDCMH model reached 0.936. Moreover, on the NUS-WIDE dataset, the F1 value of the IDCMH model was 0.877, which was significantly higher than other models. It demonstrated its strong generalization ability and robustness. This was similar to the conclusion drawn by Zeng Z et al. [5]. This study proposed a visual search model based on bag of words model and multiple semantic associations that could extract local and global features of images using scale invariant feature changes and hue, saturation, and brightness, with strong generalization ability. However, compared to it, the IDCMH model was significantly better. This was because the model mainly relied on manual feature extraction and lacked understanding of deep semantics, while the IDCMH model could automatically extract deep features of images and text through FCN and Word2vec. At the same time, the IDCMH model had higher accuracy and recall than other models on different datasets, proving its good comprehensive performance and scalability. This conclusion was similar to the findings of Ahmadi A et al. [11]. The image semantic retrieval method proposed by Ahmadi A et al., which combined ontology and composite models, effectively improved retrieval accuracy. However, the construction process of the ontology was too complex, which increased the complexity of the system and leads to low scalability. The IDCMH model proposed by the research simplified the model structure and greatly enhanced scalability through optimized feature extraction and loss functions. Therefore, the IDCMH model could automatically extract deep features of images and text, and performed well in cross-modal retrieval of digital libraries, with important application value in a wider range of practical scenarios.

V. Conclusion

The rapid development of the digital era has played a good role in promoting the construction and management of digital resources in libraries. The advanced information technology can digitize the numerous digital resources, so as to maximize the contribution and utilization of information resources. Meanwhile, digital library users have increasingly high requirements for the efficiency and precision of graphic retrieval. Therefore, the research optimized the DCMH through the graphic semantic features of digital books, introduced FCN and word2vec model for semantic feature segmentation, and designed an IDCMH retrieval model. The findings showed that, on the MIRFlickr and NUS-WIDE datasets, the IDCMH retrieval model's F1 values were 0.936 and 0.877, respectively, which were significantly higher than the F1 values of the other models on the two datasets. It demonstrated the improved performance, resilience, and capacity for generalization of the model. The accuracy of IDCMH model could reach up to 0.952 on both datasets, which proved its higher accuracy. The results indicated that the IDCMH model outperformed the other models in quantifying the similarity and repetition of graphic aspects. Its ROUGE1 and ROUGE2 scores were 0.85 and 0.96, respectively. In conclusion, the proposed IDCMH model is more effective. However, the dataset used in the study contains a small amount of target data, and the calculated results may be affected to some extent. In the future, larger sources, quantities, and types of datasets will be selected to expand the variety of datasets and the richness of annotated data, providing more comprehensive data support for applying cross-modal retrieval to digital libraries and further improving the accuracy of results.

References

[1] Liu S, Nie W, Wang C, Lu J, Qiao Z, Liu L, Anandkumar A. Multi-modal molecule structure-text model for text-based retrieval and editing. Nature Machine Intelligence, 2023, 5(12): 1447-1457.

[2] Candela G, Sáez M D, Escobar Esteban M P, Marco-Such, M. Reusing digital collections from GLAM institutions. Journal of Information Science, 2022, 48(2): 251-267.

[3] Senthil Kumaran V, Latha R. Towards personal learning environment by enhancing adaptive access to digital library using ontology-supported collaborative filtering. Library Hi Tech, 2023, 41(6): 1658-1675.

[4] Kroll H, Pirklbauer J, Plötzky F, Balke W T. A detailed library perspective on nearly unsupervised information extraction workflows in digital libraries. International Journal on Digital Libraries, 2024, 25(2): 401-425.

[5] Zeng Z, Sun S, Li T, Yin J, Shen Y. Mobile visual search model for Dunhuang murals in the smart library. Library Hi Tech, 2022, 40(6): 1796-1818.

[6] Krishnaraj N, Elhoseny M, Lydia E L, Shankar K, AL Dabbas O. An efficient radix trie-based semantic visual indexing model for large-scale image retrieval in cloud environment. Software: Practice and Experience, 2021, 51(3): 489-502.

[7] Khan U A, Javed A, Ashraf R. An effective hybrid framework for content based image retrieval (CBIR). Multimedia Tools and Applications, 2021, 80(17): 26911-26937.

[8] Yu H, Huang M, Zhang J J. Domain adaptation problem in sketch based image retrieval. ACM Transactions on Multimedia Computing, Communications and Applications, 2023, 19(3): 1-17.

[9] He S, Wang W, Wang Z, Xu X, Yang Y, Wang X, Shen H T. Category alignment adversarial learning for cross-modal retrieval. IEEE Transactions on Knowledge and Data Engineering, 2022, 35(5): 4527-4538.

[10] Wang X, Jia M. Development of a unified digital library system: integration of image processing, big data, and deep learning. International Journal of Information and Communication Technology, 2024, 24(3): 378-391.

[11] Ahmadi A, Khodabin M, Samiei M. Application of Ontologies in Information Retrieval of Digital Collections with Emphasis on Images. Journal of Knowledge Retrieval and Semantic Systems, 2022, 9(31): 189-219.

[12] Rong H, Chen Z, Lu Z, Xu F, Sheng V S. Multization: Multi-Modal Summarization Enhanced by Multi-Contextually Relevant and Irrelevant Attention Alignment. ACM Transactions on Asian and Low-Resource Language Information Processing, 2024, 23(5): 1-29.

[13] Zhen Z. Establishment of an Open Information Platform for the National Sports Center in China. Revista de Psicología del Deporte (Journal of Sport Psychology), 2024, 33(2): 99-107.

[14] Baldrati A, Bertini M, Uricchio T, Del Bimbo A. Composed image retrieval using contrastive learning and task-oriented clip-based features. ACM Transactions on Multimedia Computing, Communications and Applications, 2023, 20(3): 1-24.

[15] Nimrah S, Saifullah S. Context-Free Word Importance Scores for Attacking Neural Networks. Journal of Computational and Cognitive Engineering, 2022, 1(4): 187-192.

[16] Shahzad K, Khan S A. Factors affecting the adoption of integrated semantic digital libraries (SDLs): a systematic review. Library Hi Tech, 2023, 41(2): 386-412.

[17] Malakhov K, Petrenko M, Cohn E. Developing an ontology-based system for semantic processing of scientific digital libraries. South African Computer Journal, 2023, 35(1): 19-36.

[18] Wu Z, Xie J, Shen S, Lin C, Xu G, Chen E. A confusion method for the protection of user topic privacy in Chinese keyword-based book retrieval. ACM transactions on asian and low-resource language information processing, 2023, 22(5): 1-19.

[19] de Oliveira L L, Vargas D S, Alexandre A M A, ordeiro F C, Gomes D D S M, Rodrigue, M D. C, Moreira V P. Evaluating and mitigating the impact of OCR errors on information retrieval. International Journal on Digital Libraries, 2023, 24(1): 45-62.

[20] Song Y, Wei K, Yang S, Shu F, Qiu J. Analysis on the research progress of library and information science since the new century. Library Hi Tech, 2023, 41(4): 1145-1157.