IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 12, 2019, accepted April 1, 2019, date of publication April 11, 2019, date of current version April 25, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2910198

Modal-Dependent Retrieval Based on Mid-Level Semantic Enhancement Space

SHUNXIN ZHENG, HUAXIANG ZHANG, YUDAN QI, AND BIN ZHANG
School of Information Science and Engineering, Shandong Normal University, Jinan 250014, China

Corresponding author: Huaxiang Zhang (huaxzhang@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61572298, Grant 61772322, and Grant U1836216, and in part by the Key Research and Development Foundation of Shandong Province under Grant 2017GGX10117 and Grant 2017CXGC0703.

ABSTRACT Cross-media retrieval refers to submitting any kind of media type and obtaining similar results in the form of different media types. The existing cross-media retrieval algorithms typically learn a pair of linear (or non-linear) mappings that project image and textual features from their original heterogeneous feature space into a common homogeneous space, all these traditional methods are simply about utilizing the correlation between the original features and the high-level semantic space. At the same time, due to the "semantic gap" incurred by the extraction of the visual feature, the feature of text mode is often more distinctive than the visual features in terms of their distribution in the semantic space. Based on the above-mentioned problems, we have established a medium semantic enhancement space (MMSES) to enhance the discrimination capability of textual features. First of all, projecting the original features into the medium meaning semantic enhancement space by utilizing linear discriminant analysis, then projecting the features into a high meaning semantic space by utilizing fixed-distance projection, and finally, studying a different mapping matrix in response to different cross-media retrieval tasks during the retrieval process. In this paper, we employ traditional Euclidean distance to measure the similarity between different modalities in a common space. The effectiveness of MMSES was validated through extensive experiments on three datasets, Wikipedia, Pascal Sentence, and INRIA-Websearch.

INDEX TERMS Cross-media retrieval, modal-dependent, subspace learning, fixed-distance projection.

I. INTRODUCTION

With the rapid development of the Internet, the different forms of expression of information haves become more and more diverse, and multimedia data have gradually expanded from the original single text data to various forms of media data such as pictures, voice, video, and dynamic graphics. Different types of media data express the same information from different perspectives. For example, a picture or video is usually displayed on a webpage together with a text and used to describe the same object or news event. With the Internet penetrating into people's lives, people have become accustomed to sharing personal images and videos over the Internet, as well as seeking interesting images and texts. Although these media data have different modal forms, their corresponding high-level semantics are often strongly related. In order to effectively retrieve the information from vast amounts of different types of media data, different types of media data need to be queried and retrieved according to semantic information. For example, if we want to obtain the description of a bird, we only need to submit the bird's image as a query. The cross-media retrieval method can return relevant text descriptions, bird's voice and video information, and help us get much more information. Therefore, how to effectively understand the semantic relevance and semantic content between cross-media data have become an important research topic in the field of cross-media and pattern recognition. However, different types of multimedia data have heterogeneity among their underlying feature representations due to different feature dimensions and attributes, and there exist inconsistencies between the underlying characteristics of different types of multimedia data and their high-level semantics, the so-called cross-media heterogeneity gap and cross-media semantic gap.

Research works in the last few decades have mainly been focused on content-based multimedia retrieval [1]. In this retrieval phase, many approaches focus on single media retrieval, such as text retrieval [2]–[4], image search [5], [6],

The associate editor coordinating the review of this manuscript and approving it for publication was Quan Zou.

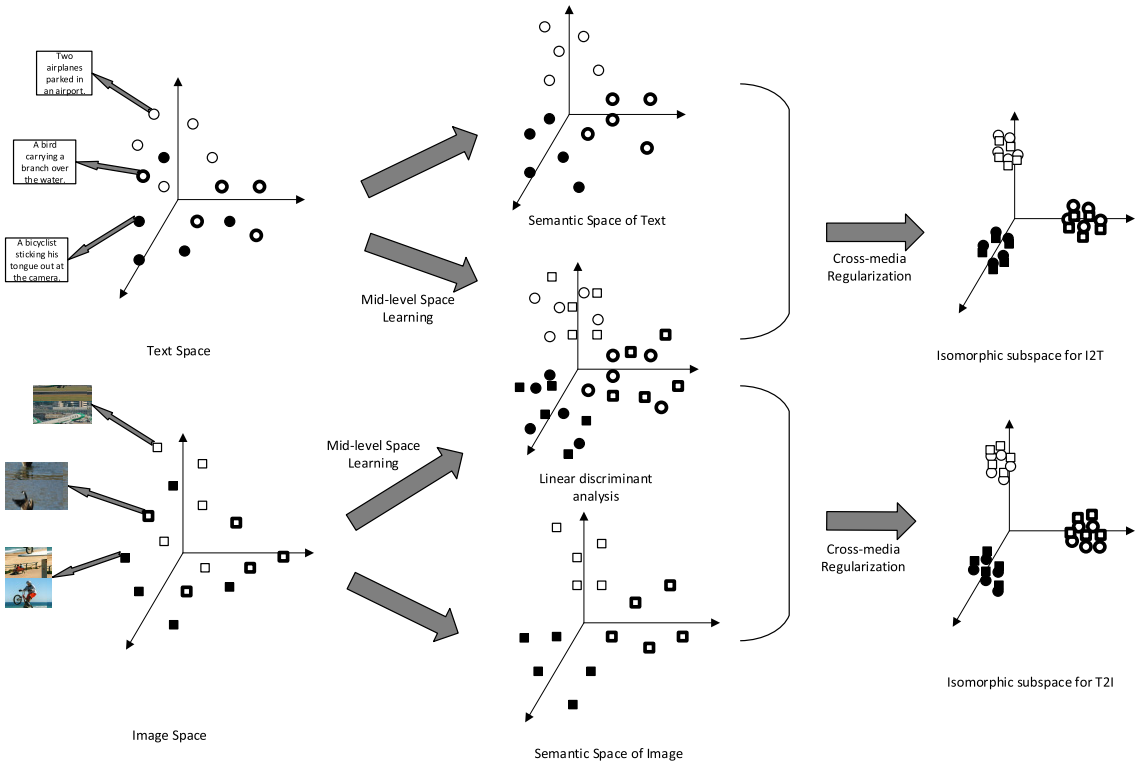


FIGURE 1. The cross-media retrieval framework proposed in this paper.

audio search [7], [8], video search [9]–[12] and so on. However, traditional single media-related technology ignores the semantic commonness of heterogeneous cross-media data, so it is difficult to effectively deal with complex data in which multiple heterogeneous cross-media data coexist. In order to solve this problem, many researches use multi-variable correlation method such as Canonical Correlation Analysis(CCA) [13] and Cross-media Factor Analysis(CFA) [14]. By maximizing the correlation between two groups of features [15], [16], the two groups of features have the highest correlation in subspace. The data in the shared subspace achieve a consistent representation. However, none of the characteristic dimension has a specific semantic theme in the primary feature space, so there will be semantic ambiguity if the correlation between the primary low-level feature and the high-order concept is simply utilized. At the same time, because of the semantic gap in the extraction of visual features, the characteristics of text modality are often more discriminative than the visual features in the semantic spatial distribution. The traditional methods do not make good use of the characteristics of textual features that have strong discriminative power in their feature space distribution to improve the distribution characteristics of their corresponding visual features in the feature space. At the same time, most methods only learn a pair of mappings for different retrieval tasks. Based on this mapping mechanism, a high performance in two retrieval tasks is usually obtained, but obtain the best performance in the respective tasks is difficult.

At present, cross-media retrieval still faces many challenges. On the one hand, the underlying feature dimensions and attributes of different modalities of cross-media data vary greatly, and it is difficult to directly measure the similarity between data on the underlying features, that is, the cross-media heterogeneous gap. On the other hand, the semantic information is abstract, and the semantic association of different modalities of media data is abstract. Therefore, different media data cannot express the underlying features of the same semantic that is, cross-media semantic gap.

In view of the above problems, this paper proposes a modal based on mid-level semantic enhancement space (MMSES). Figure 1 shows the framework of cross-media retrieval proposed in this paper. In MDCR [17], images and texts are regressed from their respective feature spaces to their semantic space by linear regression. Through the correlation analysis between pictures and texts, MDCR can ensure the relevance of paired texts and images in the shared subspace. Thus, two different sets of mappings are learned, one is applied to Image to Text(I2T) task and the other is applied to T2I task. The framework proposed in this paper differs from MDCR in that:

- In the process of correlation analysis, MDCR only considers the similar position of multimedia data with the same semantic relationship in the shared subspace for each group of mappings. In this paper, the location of multimedia data with different semantic relations in the shared subspace is further removed, and the retrieval efficiency is improved.

- In this paper, a mid-level semantic enhancement space is established, in which LDA is used to process textual features, so that text discrimination is higher, and this discrimination can be migrated to image features through distance-preserving mapping, thus, further improving of retrieval efficiency is achieved.

We have established a MMSES to further enhance the discriminant capability of text modal, which can be transferred to corresponding visual features through shared subspace, thus improving the representational capacity of visual features. Specifically speaking, we use linear discriminant analysis (LDA) to project text information into medium meaning semantic enhancement space to enhance the discriminant capability of text modal, create mapping matrix through fixed-distance projection, and transfer the enhanced discriminant capability into visual features. We also learn different mapping matrix in response to different cross-media retrieval tasks during the retrieval process. The cross-media retrieval method for modal-dependent and discriminant analysis separates the search tasks for different modalities to perform the modal-dependent retrieval [16]. Modal-dependent [17] differs from the previous methods of learning a pair of projections by learning two pairs of mappings that project Image to Text retrieval(I2T) and Text to Image retrieval(T2I) from their original feature space into two public potential subspaces. For that if two tasks are learning at the same time, the common subspace obtained is the optimal subspace common to I2T and T2I, which is usually not optimal for the semantic understanding of retrieval modality. For example, in I2T, it is generally considered that the accurate representation of the query in the image semantic space is more important to be retrieved than the text. If the semantics of the query are misjudged, it is more difficult to retrieve the relevant text. If executed separately, the image can be projected separately into its semantic space when the image is retrieved. At this time, the semantic understanding of the image without text interference is optimal. When the image semantic is understood, the retrieved data is more accurate, thereby improving the accuracy of cross-media retrieval. At the same time, during the process of mapping, the discriminative features of the text are improved by using linear discriminant analysis of LDA [18] with strong discriminative ability of textual features. LDA is a classical algorithm in pattern recognition, which is mainly used to analyze the differences between data belonging to different classes.

The main contributions of this paper are as follows:

1) This paper considers the particularity of different retrieval tasks and learns different mapping matrices for different cross-media retrieval tasks to ensure the accuracy of retrieval modal features in the process of regression in order to achieve optimal retrieval performance.

2) In most cases, because of the 'semantic gap' in the process of extracting image features, textual modality features often have stronger discriminative ability. Therefore, this paper uses LDA to process textual features to further enhance the discriminative ability of textual features.

3) Common projection of visual and textual features into a shared intermediate feature space creates an intermediate representation, and then more efficient image and text high-level semantic features can be computed from the middle feature space instead of the underlying feature space.

The rest of this article is organized as follows. In Section 2, we briefly review the related work of cross-media retrieval. We introduce our algorithm in detail in Section 3 and show the experiments in Section 4. Finally, we conclude our work in Section 5.

II. RELATED WORK

For cross-media issues, many existing retrieval studies on cross-media have learnt the common subspace of multimodal data, such as the classical method CCA [13]. Through the mapping mechanism, the features of multimedia data in the heterogeneous space are mapped into the isomorphic space. Hardoon *et al.* [19] mapped the samples of the original feature space to some other feature space through nonlinear transformation. Then, the nonlinear problem of original feature space is solved implicitly by using CCA in the new feature space. The Depth Canonical Correlation Analysis (DCCA) model was proposed by Andrew *et al.* [20]. The MSL [21] method proposed by Wei *et al.* found regularization projections by exploiting the correlation between visual features and textual features. The DPEP method proposed by Rasiwasia *et al.* [22] are aimed at establishing a fine-grained semantic level, studying semantic projection of fixed-distance, and formulating the uniform presentation of media contents.

In cross-media retrieval related tasks, there is also a way to project the raw data into a 0,1 Hamming space. In the process of projection, the original neighbor relationship and semantic information were kept as much as possible, and the retrieval efficiency was improved by calculating the hash code. Jin *et al.* [23] proposed a semantic neighbor graph cross-modal hashing method (SNGH), and its purpose is to preserve fine-grained similarity measures by jointly pursuing semantic supervision and local community structure on the basis of semantic graphs. Xu *et al.* [24] proposed a novel supervised Discrete Cross-modal Hashing method (DCH) to learn to recognize a binary code by learning a linear classifier. Li *et al.* [25] put forward a novel Label Preserving Multimedia Hashing method (LPMH), and a two-stage discrete hashing framework was used and a general approach was proposed to solve binary code based on class cation-based optimization goals.

In addition to the above-mentioned related works, many researchers proposed the method of using a depth model to solve cross-media-related issues. Socher *et al.* [26] introduced a Dependency Tree Recursive Neural Networks (DT-RNNs) which employs a dependency tree to embed the sentences in a vector space so as to retrieve images described by those sentences. Balaneshin-Kordan and Kotov [27] proposed a gated neural structure to project images and keyword

queries as well as the same low-dimensional embedding space as the multi-modality retrieval unit, and performed semantic matching in this space. Depth models [28], [29] are neural network structures with multiple hidden layers. The training effect of deep neural network is improved by adjusting the connection method and activation function of neurons. The performance of depth models will increase with the increase of data size. In the process of model building, in-depth learning decomposes the complex mapping into a series of nested simple mappings (each described by different layers of the model) to solve the problem. The input is displayed in the visible layer, followed by a series of hidden layers that extract more and more abstract features from the image. Therefore, in this field, the experimental results of depth model are often higher.

With the traditional approach described above, it can be seen that the existing cross-media search algorithms usually learn a pair of linear (or non-linear) mappings. The image and textual features were projected from their original heterogeneous feature space into a common homogenous space, and the similarity between the image and the text was measured by using the traditional metric method (in favor of Euclidean distance). The MMSES method proposed in this paper is different from the traditional cross-media retrieval algorithm that only learns a set of mappings. MESES uses various mapping mechanisms for different cross-media retrieval tasks. By combining the similarity of image, text and modal (text or image) features into the mapping matrix in its corresponding semantic space, one set is for I2T tasks and the other set is for T2I tasks. At the same time, due to the 'semantic gap' in the extraction of visual features, the characteristics of textual modality were often more discriminating than the visual features in the semantic spatial distribution. Therefore, in a specific I2T or T2I search task, the effective use of the strong discriminative power of textual features will effectively enhance the semantic representation of the corresponding visual features. The MESES method put forward in this paper analyzes the linear discriminant terms of phase and textual features by jointly optimizing the correlation between the images and texts. The discriminative ability of textual features has been further enhanced and it has migrated to the corresponding visual feature summary through the sharing of subspace learning, thereby enhancing the visual feature representation ability.

III. THE PROPOSED APPROACH

In this section, we proposed the MMSES approach. We considered the problem of cross-media retrieval between images and text. In the training sample, each pair of pictures and text corresponded to the specific semantic information, namely category labels. Gong *et al.* [30] proposed the use of such semantic information as a third perspective in subspace learning to enhance the similarity of different modal data with the same semantics in shared subspaces. This chapter also makes use of semantic information to improve the consistency of the feature representation of data with the same semantics in the shared subspace. The difference is that the shared subspace dimension in this paper is determined by the number of dataset categories, and this paper uses different mapping methods for different search tasks.

Assuming that we have three datasets, we give a dataset $\Omega = \{(i_i, t_i)\}_{i=1}^{n}$ with n data samples, the $i_i \in R^p$ and $t_i \in R^q$ are defined to separately represent the original low-level features of the text and image. Regarding the dataset I of the image modality and the data set T concerning the text modality, two data matrices $I = [i_1, i_2, \ldots i_n] \in R^{n\times p}$ and $T = [t_1, t_2, \ldots t_n] \in R^{n\times q}$ are defined. If there are k categories in Ω, a semantic matrix $Y = [y_1, y_2, \ldots y_n] \in R^{n\times k}$ will be set to be consistent with the image and text. In addition, two projection matrices $V \in R^{k\times p}$ and $W \in R^{k\times q}$ are set.

A. MID-LEVEL SEMANTIC ENHANCEMENT SPACE

1) LINEAR DISCRIMINANT ANALYSIS BASED ON TEXT

The LDA [17] algorithm is mainly used to analyze the differences between various kinds of data, and this algorithm is mainly used to classify samples. The core idea of classification is to project high-dimensional sample data into the best-classified vector space so as to ensure that there is a greater inter-class distance and smaller Intra-class distance [31] in the new subspace. Letting $\overline{m}_j$ be the mean value of the j-th textual features, $\overline{m}$ mean the textual features of all classes, the $yi = [1, 2, \ldots n]$ is defined to be the semantic matrix with *ith* row being the semantic vector each pair of i_i and t_i , the intra-class divergence matrix may be expressed as $U_w = \sum_{j=1}^{n}\sum_{yi=j} {}^{1}\!/_{n}(t_i - \overline{m}_j)(t_i - \overline{m}_j)^T$, and the global divergence matrix may be expressed as $U_t = \sum_{i=1}^{n} {}^{1}\!/_{n}(t_i - \overline{m})(t_i - \overline{m})^T$. In summary, linear discriminant analysis LDA is:

$$\min_{UU^T=I_K} \frac{tr(WU_wW^T)}{tr(WU_tW^T)} \tag{1}$$

where $tr(\bullet)$ represents the trace of the matrix, $W \in R^{k\times q}$ belongs to the projection matrix of the text, it consists of K basis vectors, and I_K is the identity matrix.

2) FIXED-DISTANCE PROJECTION

To utilize the cross-media relevance of different media objects, to study fixe-distance projection, to minimize the distance between media content to create a positive correlation, and to maximize the distance between media content to create a negative correlation. In consideration of effective and convenient solutions, we employed Euclidean distance to measure the wastage and employed linear projection to obtain the unified representation of the physical plane. The contiguous item of cross-media $f(V, W)$is defined as follows:

$$f(V, W) = \sum_{i=1}^{n}\sum_{j=1}^{n} s_{ij}||t_iW - i_jV||_2^2 \tag{2}$$

The s_{ij} is defined as the similarity matrix of text t_i and image i_j, y_i and y_j are defined as the label Matrix-Vector. as

Algorithm 1 The Main I2T Steps of MMSES are Displayed in Algorithm 1. Optimization for MMSES Method Over I2T

Input: The feature matrix I of the image, the feature matrix T of the text, and the semantics matrix Y of the image and the text consistent.
1: Initialize the projection matrix V_1^n, W_1^m, $n \leftarrow 0$ and $m \leftarrow 0$.set parameters as α, β, λ, μ and ε, where μ is the step length in the iterative update process, and ε is the convergence condition.
2: compute the intra-class divergence matrix U_w of text data.
compute the global divergence matrix U_t of text data
compute the matrix L form Eq.(5)
3: Repeat
4: Set $value1 = \Omega(V_1^n, W_1^m)$;
5: Update $V_1^{n+1} = V_1^n - \mu\Delta_{V_1^n}\Omega(V_1^n, W_1^m)$;
6: Set $value2 = \Omega(V_1^{n+1}, W_1^m)$, $n \leftarrow n+1$;
7: Until $value1 - value2 \leq \varepsilon$;
8: Repeat
9: Set $value1 = \Omega(V_1^n, W_1^m)$;
9: Update $W_1^{m+1} = W_1^m - \mu\Delta_{W_1^m}\Omega(V_1^n, W_1^m)$;
10: Set $value2 = \Omega(V_1^n, W_1^{m+1})$, $m \leftarrow m+1$;
11: Until $value1 - value2 \leq \varepsilon$;
Output: projection matrices V_1, W_1.

shown in Eq(3).

$$s_{ij} = \begin{cases} 1, & \text{when } y_i = y_j \\ -1 & \text{when } y_i \neq y_j \end{cases} \tag{3}$$

The single-media similarity matrices are defined as S_{TT} and S_{II}, S_{TT} and S_{IT} are defined as the cross-media similarity matrices. And$S = \{s_{ij}\}_{2m\times 2m}$ is defined as the whole similarity matrix, depicted in Eq. (4).

$$S = \begin{pmatrix} S_{TT} & S_{TI} \\ S_{TT} & S_{II} \end{pmatrix} \tag{4}$$

In Eq. (4), $S_{IT} = S_{TI} = S_{TI}^T$, $S_{II} = S_{TT}$, $S = S^T$. In $f(V, W)$, S_{TT} and S_{II} are set as zero matrices mainly by utilizing the cross-media correlation. In order to balance the positive and negative correlations, the S is normalized so that the sum of each row in S is equal to zero and transform S into a symmetric matrix.

$D = Diag\left(d_{1,1}, d_{2,2}, \ldots, d_{2n,2n}\right)$ is defined as a diagonal matrix, where $d_{i,i} = \sum_{i=1}^{n} s_{ii}$. Then, Eq. (2) can be rewritten in a matrix form, where$L_{IT} = L_{TI} = L_{TI}^T$ and $L_{II} = L_{TT}$.

$$L = D - S = \begin{pmatrix} L_{TT} & L_{TI} \\ L_{IT} & L_{II} \end{pmatrix} \tag{5}$$

$$\begin{aligned} f(V, W) &= tr\left(\begin{pmatrix} TW^T \\ IV^T \end{pmatrix}^T L \begin{pmatrix} TW^T \\ IV^T \end{pmatrix}\right) \\ &= tr\left(\left(TW^T\right)^T L_{TT} TW^T + 2\left(TW^T\right)^T L_{TI} IV^T \right. \\ &\quad \left. + \left(IV^T\right)^T L_{II} IV^T\right) \end{aligned} \tag{6}$$

B. UNIFORMING REPRESENTATION LEARNING OF MODALITY-DEPENDENT

MMSES provides a shared characteristic space $R^{n\times q}$ for image and text, where the visual features and corresponding textual features are distributed in a similar manner. In this process, MMSES studies two groups of mapping matrices according to the bottom textual features and visual features, which can be used to represent the optimization problem of V and W as below,

$$\min_{V,W} f(V, W) + \alpha g(V, W) + \beta r(V, W) + \lambda l(W, U_w, U_t) \tag{7}$$

$\min_{V,W} f(V, W)$ is defined, the cross-media correlation term is aimed at studying fixed-distance projection, $\alpha g(V, W)$ is defined as a linear regression term, it is learnt that the projection matrix, the original features of the image are mapped into a high-level semantic space, $\beta r(V, W)$ is defined as a regularization term, the role is to control the complexity of the two projection matrices, to avoid overfitting, and to denote. $l(W, U_w, U_t)$ is a discriminant analysis item used to improve text feature clustering obtained. α, β and λ are the parameters for balancing the weights of different terms. Next, we will introduce the two algorithms applied to I2T and T2I.

1) IMAGE RETRIEVAL TEXT

First of all, the task of learning image retrieve text is introduced, because this article uses modal-dependent, so based on the two different projection matrices $V_1 \in R^{k\times p}$ and $W_1 \in R^{k\times q}$ as well as the above analysis, the image retrieve text optimization framework is as follows:

$$\begin{aligned} \Omega(V_1, W_1) = &\, tr((TW_1^T)^T L_{TT} TW_1^T + 2(TW_1^T)^T L_{TI} IV_1^T \\ &+ (IV_1^T) Y^T L_{II} IV_1^T) \\ &+ \alpha \left\| IV_1^T - Y \right\|_F^2 + \beta \left\| V_1 \right\|_F^2 + \beta \left\| W_1 \right\|_F^2 \\ &+ tr(W_1 U_w W_1^T) - \lambda tr(W_1 U_t W_1^T) \end{aligned} \tag{8}$$

Algorithm 2 The Main T2I Steps of MMSES are Displayed in Algorithm 2. Optimization for MMSES Method Over T2I

Input: the feature matrices I of the image, the feature matrix T of the text, and the semantics matrix Y of the image and the text consistent.

1: Initialize the projection matrix V_2^n, W_2^m, $n \leftarrow 0$ and $m \leftarrow 0$.set parameters as α, β, λ, μ and ε, where μ is the step length in the iterative update process, and ε is the convergence condition.

2: compute the within-class scatter matrix U_w of text data.
compute the total scatter matrix U_w of text data
compute the matrix L form Eq.(5)

3: Repeat

4: Set $value1 = \Omega(V_2^n, W_2^m)$;

5: Update $W_2^{m+1} = W_2^m - \mu \Delta_{W_2^m} \Omega(V_2^n, W_2^m)$;

6: Set $value2 = \Omega(V_2^n, W_2^{m+1})$, $m \leftarrow m + 1$;

7: Until $value1 - value2 \leq \varepsilon$;

8: Repeat

9: Set $value1 = \Omega(V_2^n, W_2^m)$;

9: Update$V_2^{n+1} = V_2^n - \mu \Delta_{V_2^n} \Omega(V_2^n, W_2^m)$;

10: Set $value2 = \Omega(V_1^{n+1}, W_1^m)$, $n \leftarrow n + 1$;

11: Until $value1 - value2 \leq \varepsilon$;

Output: projection matrices V_2, W_2.

TABLE 1. The mAP scores on wikipedia (numbers in boldface are the best.

Method	mAP		
	I2T	T2I	Average
PLS	0.359	0.351	0.355
CCA	0.331	0.316	0.355
SM	0.368	0.386	0.377
SCM	0.374	0.392	0.383
GMMFA	0.284	0.248	0.266
GMLDA	0.300	0.280	0.290
JFSSL	0.392	0.381	0.386
MDCR	0.410	0.377	0.394
JLSLR	0.393	0.369	0.381
GSSSL	0.413	0.376	0.395
CMOLRS	0.424	0.382	0.403
MMSES	**0.438**	**0.395**	**0.417**

TABLE 2. The mAP scores on pascal sentence (numbers in boldface are the best.

Method	mAP		
	I2T	T2I	Average
PLS	0.365	0.376	0.370
CCA	0.379	0.372	0.375
SM	0.449	0.433	0.441
SCM	0.407	0.393	0.400
GMMFA	0.373	0.347	0.360
GMLDA	0.408	0.387	0.397
JFSSL	0.406	0.401	0.404
MDCR	0.432	0.462	0.447
JLSLR	0.454	0.455	0.454
GSSSL	0.411	0.425	0.418
CMOLRS	0.358	0.374	0.366
MMSES	**0.485**	**0.490**	**0.488**

Among them, α, β and λ are the balance parameters, which is between 0-1. $tr((TW_1^T)^T L_{TT} TW_1^T + 2(TW_1^T)^T L_{TI} IV_1^T + (IV_1^T)Y^T L_{II} IV_1^T)$are the cross-media correlation term aimed at studying fixed-distance projection, $\alpha \parallel IV_1^T - Y \parallel_F^2$is a linear regression term, learning the projection matrix, the original features of the image are mapped into a high-level semantic space. $\beta \parallel V_1 \parallel_F^2 + \beta \parallel W_1 \parallel_F^2$ is a regularization term, its role is to control the complexity of the two projection matrices, to avoid overfitting. $tr(W_1 U_w W_1^T) - \lambda tr(W_1 U_t W_1^T)$ is a discriminant analysis item to improve text feature clustering obtained.

2) TEXT RETRIEVAL IMAGE

In the task of text retrieve images, we learn two projection matrices $V_2 \in R^{k \times p}$ and $W_2 \in R^{k \times q}$ that are different from the image retrieve texts. The regression terms are changed to the regression of the vector quantity of the corresponding semantic characteristics in the process of text retrieve image, which is different from the objective function of I2T.

Based on the above analysis, the optimized framework for obtaining text retrieve images is as follows:

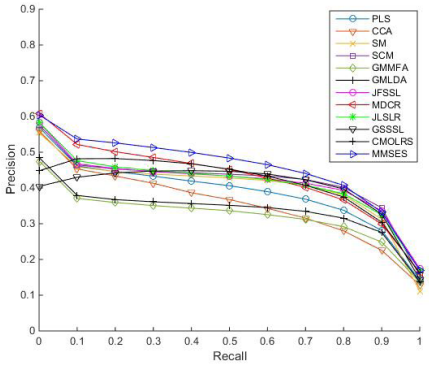
$$\begin{aligned}\Omega(V_2, W_2) = {} & tr((TW_2^T)^T L_{TT} TW_2^T + 2(TW_2^T)^T L_{TI} IV_2^T \\ & + (IV_2^T)Y^T L_{II} IV_2^T) \\ & + \alpha \left\| TW_2^T - Y \right\|_F^2 + \beta \parallel V_2 \parallel_F^2 + \beta \parallel W_2 \parallel_F^2 \\ & + tr(W_2 U_w W_2^T) - \lambda tr(W_2 U_t W_2^T) \end{aligned} \quad (9)$$

C. OPTIMIZATION

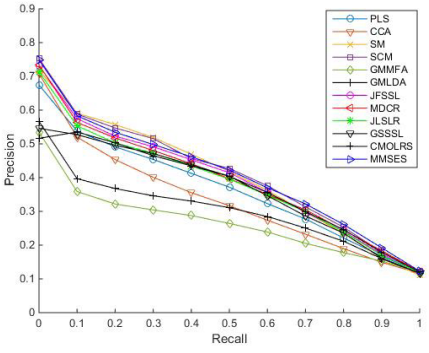
This paper adopts an iteratively updated strategy to optimize the function and to obtain the final result by finding partial derivatives.

TABLE 3. The mAP scores on INRIA-Websearch(numbers in boldface are the best).

Method	mAP		
	I2T	T2I	Average
PLS	0.193	0.260	0.227
CCA	0.260	0.279	0.269
SM	0.378	0.353	0.365
SCM	0.354	0.308	0.331
GMMFA	0.280	0.303	0.292
GMLDA	0.475	0.540	0.508
JFSSL	0.532	0.562	0.547
MDCR	0.470	0.459	0.465
JLSLR	0.525	0.544	0.534
GSSSL	0.371	0.365	0.368
CMOLRS	0.415	0.423	0.419
MMSES	**0.536**	**0.568**	**0.552**



(a)



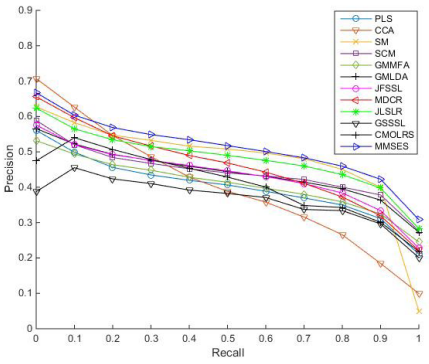
(b)

FIGURE 2. Comparison of recall rate curves for wikipedia datasets. (a) I2T. (b) T2I.

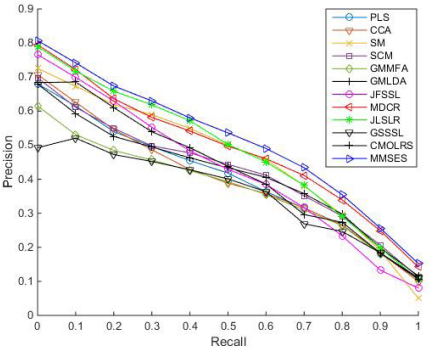
find the partial derivative of V_1 and W_1 according to Eq.(10)and Eq.(11):

$$\Delta_{V_1}\Omega(V_1, W_1) = 2I^T L_{II} V^T + 2I^T L_{TI} T W^T + 2\alpha I^T (IV^T - Y) + 2\beta V \quad (10)$$

$$\Delta_{W_1}\Omega(V_1, W_1) = 2T^T L_{II} T W^T + 2T^T L_{II} I V^T + 2\alpha W^T + W_1 U_w^T + W_1 U_w - \lambda(W_1 U_t^T + W_1 U_t) \quad (11)$$



(a)



(b)

FIGURE 3. Comparison of recall rate curves on the pascal sentence dataset. (a) I2T. (b) T2I.

find the partial derivative of V_2 and W_2 according to Eq.(12)and Eq.(13):

$$\Delta_{V_2}\Omega(V_2, W_2) = 2I^T L_{II} I V^T + 2I^T L_{TI} T W^T + 2\beta V \quad (12)$$

$$\Delta_{W_2}\Omega(V_2, W_2) = 2T^T L_{II} T W^T + 2T^T L_{II} I V^T + 2\alpha T^T \times (TW^T - Y) + \cdots + 2\beta W^T + W_2 U_w^T + W_2 U_w - \lambda(W_2 U_t^T + W_2 U_t) \quad (13)$$

IV. EXPERIMENTS

The method in this article is also applicable for the retrieval of other modalities. Here we use text and image as examples for experiments. In this section, we will compare our method with other advanced methods on the three open cross-media datasets wikipedia [22], pascal sentence [32] and inria-websearch [33].

A. DATASETS DESCRIPTION

Wikipedia [22]: In this paper, the most advanced algorithms based on the visual features of the 4096-dimensional convolutional neural network (CNN) and the 100-dimensional LDA textual features are also compared. For 100-dimensional textual features, first based on the probability distributions of 500 grouped histogram features in 100 implicit topics, this paper uses them as textual features.

Pascal Sentence [32]: it contains 1000 textual image pairs, with a total of 20 categories (each containing 50 pairs). This chapter randomly selects 30 pairs from each class for training

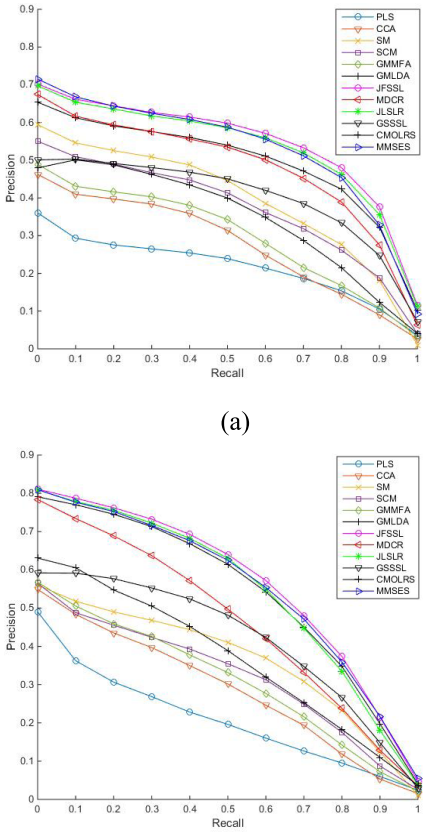


FIGURE 4. Comparison of recall rate curves on the inria-websearch dataset. (a) I2T. (b) T2I.

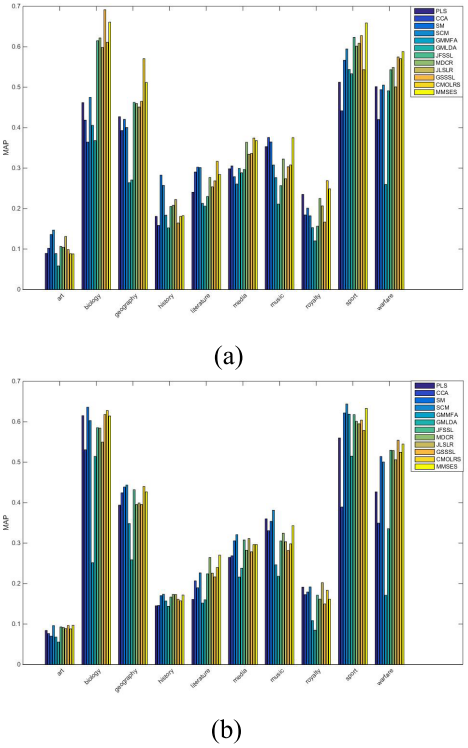


FIGURE 5. Performance comparison of each class of wikipedia dataset. (a) I2T. (b) T2I.

and the rest for testing. For images, this paper uses 4096-dimensional CNN visual features. The probability distribution of each text in 100 hidden topics is calculated by LDA, and it is expressed as a text feature.

INRIA-Websearch [33]: Its original dataset contains a total of 71478 image text pairs from 353 categories. This chapter selects the top 100 classes with the largest number of samples and then obtains a subset containing 14698 pairs of samples. Among these, 70% are randomly selected for training in each category, and the rest is used for testing. For images, this chapter uses 4096-dimensional CNN visual features. For text, the probability distribution of each text in 1000 hidden topics is calculated by LDA, and it is expressed as a feature of the text.

B. COMPARISON METHODS AND EVALUATION INDICATORS

In this paper, Euclidean distance is used to measure the similarity between the text and the image feature isomorphic space, and the I2T and Text2Img tasks are processed separately. Mean AP and Precision-Recall curve (PR) are used to evaluate the performance of cross-media retrieval.

$$AP = \frac{\sum_{j=1}^{R} P(j)rel(j)}{\sum_{i=1}^{R} rel(j)} \tag{14}$$

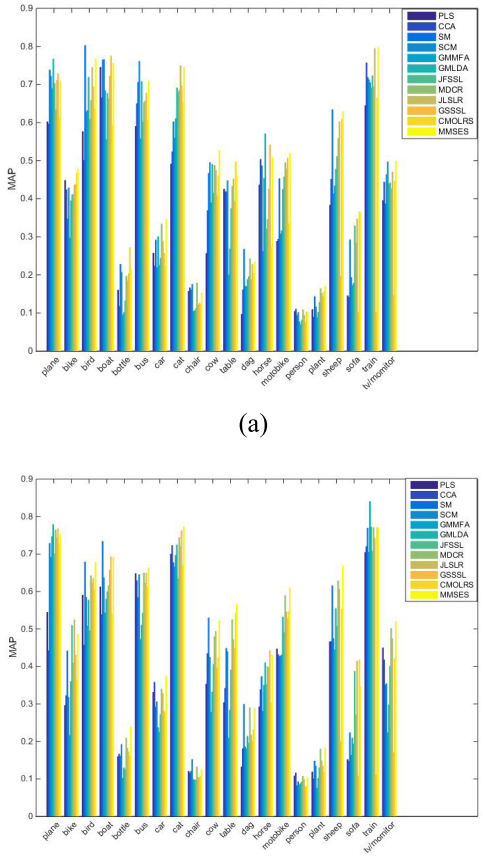
Among which R will refer to the number of searching results, if the sample of No j is consistent with the semantics of the item being searched, otherwise. $P(j)$ will refer to the accuracy rate of j results, and the final mAP can be calculated out by calculating the average value of all searched items AP.

C. EXPERIMENTAL RESULTS AND ANALYSIS

The MMSES method proposed in this paper is mainly compared according to the following six popular cross-media retrieval algorithms:

Overview and Recent Advances in Partial Least Squares (PLS) [34], Canonical Correlation Analysis(CCA) [13], Semantic Matching(SM) [23], Semantic Correlation Matching(SCM) [35], Generalized Multi-view Marginal Fisher Analysis(GMMFA) [35],Generalized Multi-view Linear Discriminant Analysis(GMLDA), Joint Feature Selection and Subspace Learning for Cross-Modal Retrieval(JFSSL) [36], Modality-Dependent Cross-Media Retrieval(MDCR) [17], Latent Subspace Learning and Regression for Cross-Modal Retrieval(JLSLR) [37], Generalized Semi-supervised and Structured Subspace Learning for Cross-modal Retrieval [38], Online low-rank similarity function learning with adaptive relative margin for cross-modal retrieval [39]. During the experiment, all methods were compared by using the same dataset. Table 1 gives the comparison results of this method with other cross-media retrieval methods.

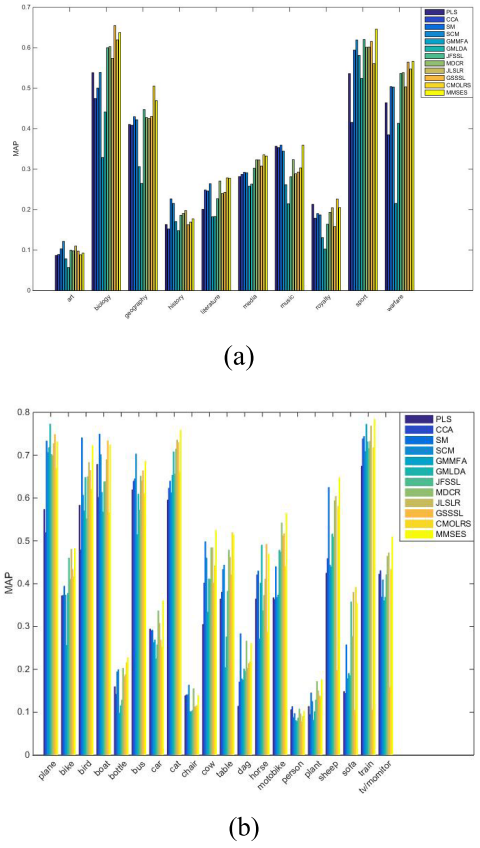
For the Wikipedia dataset, after testing the different parameter settings, for the image retrieve text, this article sets the parameters: $\alpha = 0.1$, $\beta = 0.9$, $\lambda = 0.7$, $\mu = 0.002$



(a)

(b)

FIGURE 6. Performance comparison of each type of pascal sentence dataset. (a) I2T. (b) T2I.



(a)

(b)

FIGURE 7. Comparison of the average performance of each class of wikipedia and pascal sentence datasets. (a) Wikipedia dataset. (b) Pascal sentence.

and $\varepsilon = 10^{-4}$. For text retrieve images, the parameters are set as $\beta = 0.1$, $\lambda = 0.6$, $\mu = 0.002$ and $\varepsilon = 10^{-2}$, Figure 2 shows the dataset recall rate curve performance comparison, we can see the overall method is more efficient than these advanced algorithms.

For Pascal Sentence, the different parameter settings are tested, for the image retrieve text, parameters$\alpha = 0.1$, $\beta = 0.3$, $\lambda = 0.1$, $\mu = 0.002$ and $\varepsilon = 10^{-3}$ are set. For text retrieve images, the parameters $\alpha = 0.5$, $\beta = 0.5$, $\lambda = 0.2$, $\mu = 0.002$ and $\varepsilon = 10^{-4}$ are set. Figure 3 compares the performance of recall curves on this data set. It can be seen that the overall method has been significantly improved over these advanced algorithms.

For the INRIA-Websearch dataset, after testing the different parameter settings, for the image retrieve text, the parameters $\alpha = 0.6$, $\beta = 0.5$, $\lambda = 0.1$, $\mu = 0.0002$ and $\varepsilon = 10^{-4}$ are set. For text retrieve images, the parameters $\alpha = 0.7$, $\beta = 0.5$, $\lambda = 0.1$, $\mu = 0.0002$ and $\varepsilon = 10^{-4}$ are set. Figure 4 compares the performance of recall curves on this data set. It can be seen that the overall method is more efficient than the advanced algorithms.

Figure 5 and Figure 6 show the cross-media locked mAP comparisons for each method on the Wikipedia dataset and the Pascal Sentence for different methods. Figure 7 is the

TABLE 4. mAP comparison between MMSES and its variants.

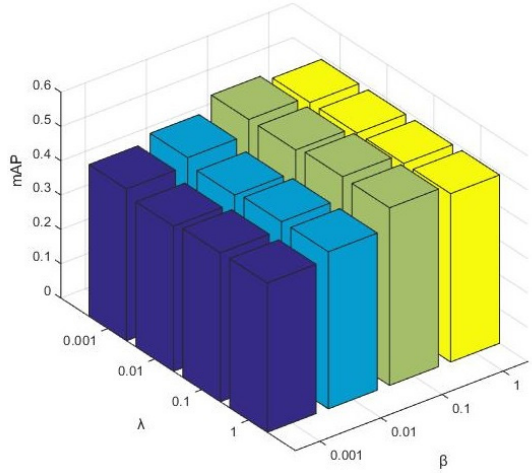
	I2T	T2I	Average
MMSES	0.536	0.568	0.552
MMSES-I	0.452	0.470	0.461
MMSES-II	0.389	0.442	0.416

Wikipedia dataset and the Pascal Sentence for image retrieve text and text retrieve image, and the average of the performance comparisons in each class. These comparative analyses further prove the effectiveness of the proposed algorithm.

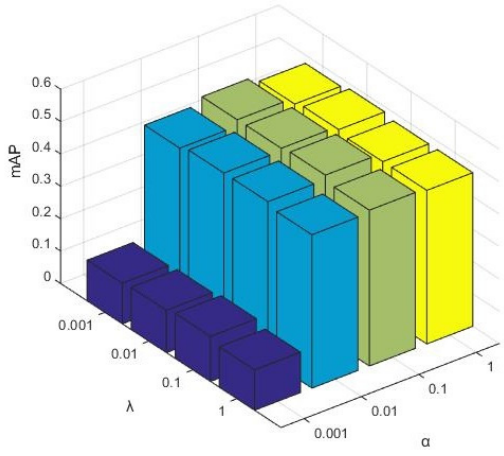
D. EFFECTS OF MODAL-DEPENDENT AND MID-LEVEL SEMANTIC ENHANCEMENT SPACE

In this section, ablation studies are tested to verify their validity based on Pascal sentence. In the experiment, only the objective function is different. The main results are shown in Table 4. MMSES-I represents a variant of our method, which learns the same projection matrix for different cross-media retrieval tasks. In implementation, the objective function of MMSES-I is

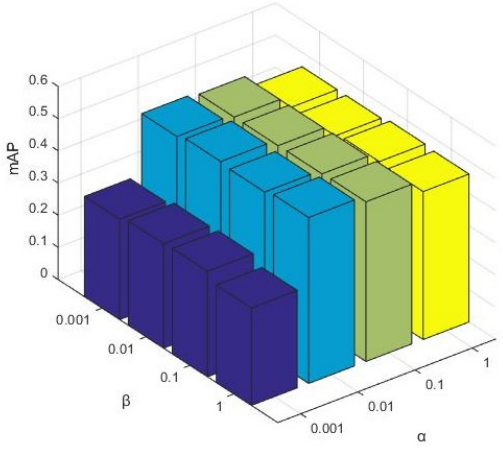
$$\begin{aligned}\Omega(V_1, W_1) = & tr((TW_1^T)^T L_{TT} TW_1^T + 2(TW_1^T)^T L_{TI} IV_1^T \\ & + (IV_1^T)Y^T L_{II} IV_1^T) + \alpha \left\| IV_1^T - Y \right\|_F^2 \\ & + \alpha \left\| TW_2^T - Y \right\|_F^2 + \beta \left\| V_1 \right\|_F^2 + \beta \left\| W_1 \right\|_F^2 \\ & + tr(W_1 U_w W_1^T) - \lambda tr(W_1 U_t W_1^T) \end{aligned} \quad (15)$$



(a)



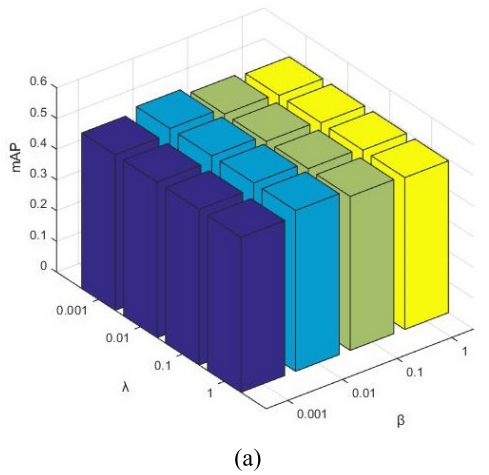
(b)



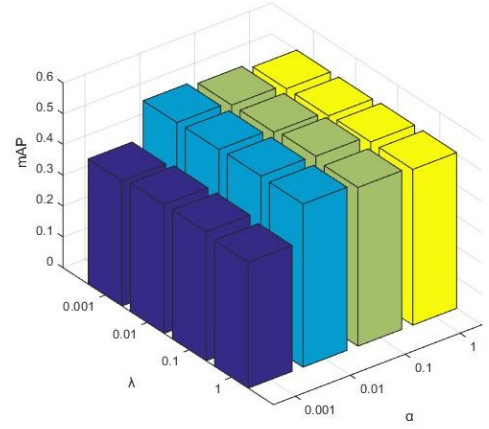
(c)

FIGURE 8. Parameter sensitivity and I2T experiment on INRIA-Websearch. (a) $\alpha = 0.6$. (b) $\beta = 0.5$. (c) $\lambda = 0.1$.

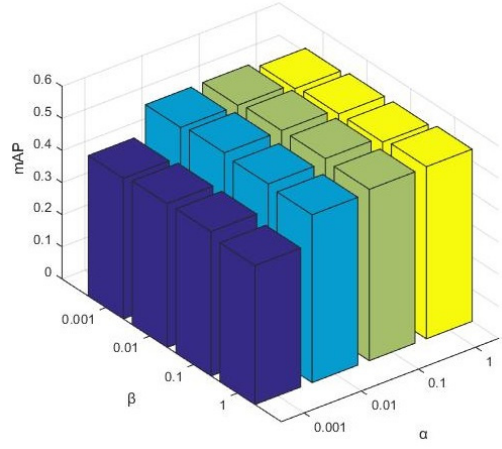
MMSES-I represents another variant of our method, which removes the intermediate semantic enhancement space in MMSES. In order to implement MMSES-II, LDA constraints are removed.



(a)



(b)



(c)

FIGURE 9. Parameter sensitivity and T2I experiment on INRIA-Websearch. (a) $\alpha = 0.6$. (b) $\beta = 0.5$. (c) $\lambda = 0.1$.

$$\begin{aligned}\Omega(V_1, W_1) = tr((TW_1^T)^T L_{TT} TW_1^T + 2(TW_1^T)^T L_{II} IV_1^T \\ +(IV_1^T)Y^T L_{II} IV_1^T) + \alpha \left\| IV_1^T - Y \right\|_F^2 \\ +\beta \left\| V_1 \right\|_F^2 + \beta \left\| W_1 \right\|_F^2 \end{aligned} \quad (16)$$

$$\Omega(V_1, W_1) = tr((TW_1^T)^T L_{TT} TW_1^T + 2(TW_1^T)^T L_{TI} IV_1^T + (IV_1^T)Y^T L_{II} IV_1^T) + \alpha \left\| TW_2^T - Y \right\|_F^2 + \beta \|V_1\|_F^2 + \beta \|W_1\|_F^2 \quad (17)$$

which is the objective function of the proposed approach. As shown in Table 4, the performance of different retrieval tasks is greatly reduced, especially for image queries. The reason for the poor efficiency of the retrieval of these two variants is that different retrieval tasks learn the same projection matrix and cannot effectively improve the discriminability of textual features.

E. PARAMETER SENSITIVITY

In this section, we conduct experiments to evaluate the robustness of the proposed method. The method has three parameters: α, β and λ. α is a weighted parameter of linear regression term. β is the equilibrium parameter of the regular term. λ is the equilibrium parameter of LDA. In the experiment, on different datasets, α, β and λ are set to different values according to different retrieval tasks to optimize retrieval performance. In this section, parameter sensitivity experiments are carried out on INRIA-Websearch dataset. The values of the three parameters are adjusted in 0.001, 0.01, 0.1 and 1. We determine one parameter to observe the performance changes of the other two parameters. The experimental results of I2T task are shown in Fig. 8 (a) - (c), and those of T2I task are shown in Fig. 9 (a) - (c). It can be seen that the performance of this method is relatively stable for parameters β and λ, and stable for α in the range of 0.01, 0.1, 1.

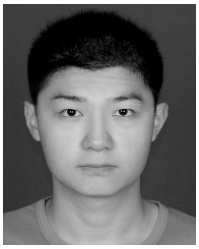
V. CONCLUSION

In this paper a medium meaning semantic enhancement space (MMSES), established based on modality-dependent, the semantic features of the text are further enhanced, then the enhanced discrimination capability of textual features is transferred to the corresponding visual features, which studies different mapping matrices in response to different cross-media retrieval tasks during the retrieval process. At last, the traditional Euclidean distance is employed to measure the similarity of the characteristics of different modal in homogeneous space. The effectiveness of MMSES was validated through extensive experiments on three datasets, Wikipedia, Pascal Sentence and INRIA-Websearch.

REFERENCES

[1] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain, "Content-based multimedia information retrieval: State of the art and challenges," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 2, no. 1, pp. 1–19, 2006.
[2] S. Haiduc, G. Bavota, A. Marcus, R. Oliveto, A. De Lucia, and T. Menzies, "Automatic query reformulations for text retrieval in software engineering," in *Proc. Int. Conf. Softw. Eng.*, 2013, pp. 842–851.
[3] S. Shehata, F. Karray, and M. S. Kamel, "An efficient concept-based retrieval model for enhancing text retrieval quality," *Knowl. Inf. Syst.*, vol. 35, no. 2, pp. 411–434, 2013.
[4] W. Song, Y. Cui, and Z. Peng, "A full-text retrieval algorithm for encrypted data in cloud storage applications," in *Natural Language Processing and Chinese Computing*. Springer, 2015.
[5] G.-H. Liu and J.-Y. Yang, "Content-based image retrieval using color difference histogram," *Pattern Recognit.*, vol. 46, no. 1, pp. 188–198, 2013.
[6] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 12, pp. 1349–1380, Dec. 2000.
[7] P. Hu, W. Liu, W. Jiang, and Z. Yang, "Latent topic model for audio retrieval," *Pattern Recognit.*, vol. 47, no. 3, pp. 1138–1143, 2014.
[8] C. Pedraza, J. Vitola, J. Sepulveda, and J. I. Martinez, "Fast content-based audio retrieval algorithm," in *Proc. Symp. Signals, Images Artif. Vis.*, Sep. 2013, pp. 1–5.
[9] B. Andre, T. Vercauteren, A. M. Buchner, M. B. Wallace, and N. Ayache, "Learning semantic and visual similarity for endomicroscopy video retrieval," *IEEE Trans. Med. Imag.*, vol. 31, no. 6, pp. 1276–1288, Jun. 2012.
[10] X. Chang, Y.-L. Yu, Y. Yang, and E. P. Xing, "Semantic pooling for complex event analysis in untrimmed videos," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 8, pp. 1617–1632, Aug. 2017.
[11] X. Nie, Y. Yin, J. Sun, J. Liu, and C. Cui, "Comprehensive feature-based robust video fingerprinting using tensor model," *IEEE Trans. Multimedia*, vol. 19, no. 4, pp. 785–796, Apr. 2017.
[12] J. Sun, X. Liu, W. Wan, J. Li, D. Zhao, and H. Zhang, "Video hashing based on appearance and attention features fusion via DBN," *Neurocomputing*, vol. 213, pp. 84–94, Nov. 2016.
[13] F. Wu, H. Zhang, and Y. Zhuang, "Learning semantic correlations for cross-media retrieval," in *Proc. Int. Conf. Image Process.*, Oct. 2006, pp. 1465–1468.
[14] D. Li, N. Dimitrova, M. Li, and I. K. Sethi, "Multimedia content processing through cross-modal association," in *Proc. 11th ACM Int. Conf. Multimedia*, 2003, pp. 604–611.
[15] Y. Yuan and Y. Peng, "Recursive pyramid network with joint attention for cross-media retrieval," in *Proc. Int. Conf. Multimedia Modeling*, 2018, pp. 405–416.
[16] X. Dong *et al.*, "Semi-supervised modality-dependent cross-media retrieval," *Multimedia Tools Appl.*, vol. 77, no. 3, pp. 3579–3595, 2018.
[17] Y. Wei *et al.*, "Modality-dependent cross-media retrieval," *ACM Trans. Intell. Syst. Technol.*, vol. 7, no. 4, 2016, Art. no. 57.
[18] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. Hoboken, NJ, USA: Wiley, 2001.
[19] D. R. Hardoon, S. R. Szedmak, and J. R. Shawe-Taylor, *Canonical Correlation Analysis: An Overview with Application to Learning Methods*. Cambridge, MA, USA: MIT Press, 2004.
[20] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 1247–1255.
[21] Y. Wei, Y. Zhao, Z. Zhu, Y. Xiao, and S. Wei, "Learning a mid-level feature space for cross-media regularization," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2014, pp. 1–6.
[22] N. Rasiwasia *et al.*, "A new approach to cross-modal multimedia retrieval," in *Proc. 18th ACM Int. Conf. Multimedia*, 2010, pp. 251–260.
[23] L. Jin, K. Li, H. Hu, G.-J. Qi, and J. Tang, "Semantic neighbor graph hashing for multimodal retrieval," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1405–1417, Mar. 2018.
[24] X. Xu, F. Shen, Y. Yang, H. T. Shen, and X. Li, "Learning discriminative binary codes for large-scale cross-modal retrieval," *IEEE Trans. Image Process.*, vol. 26, no. 5, pp. 2494–2507, May 2017.
[25] K. Li, G.-J. Qi, and K. A. Hua, "Learning label preserving binary codes for multimedia retrieval: A general approach," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 14, no. 1, 2017, Art. no. 2.
[26] R. Socher, A. Karpathy, Q. V. Le, C. D. Manning, and A. Y. Ng. (2013) *Grounded Compo–Sitional Semantics for Finding and Describing Images With Sentences*. [Online]. Available: https://nlp.stanford.edu/
[27] S. Balaneshin-Kordan and A. Kotov, "Deep neural architecture for multi-modal retrieval based on joint embedding space for text and images," in *Proc. 11th ACM Int. Conf. Web Search Data Mining*, 2018, pp. 28–36.
[28] B. Wang, Y. Yang, X. Xu, A. Hanjalic, and H. T. Shen, "Adversarial cross-modal retrieval," in *Proc. 25th ACM Int. Conf. Multimedia*, 2017, pp. 154–162.
[29] X. Huang and Y. Peng, "Deep cross-media knowledge transfer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 8837–8846.
[30] Y. Gong, Q. Ke, M. Isard, and S. Lazebnik, "A multi-view embedding space for modeling Internet images, tags, and their semantics," *Int. J. Comput. Vis.*, vol. 106, no. 2, pp. 210–233, 2014.

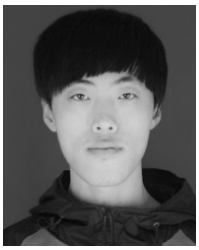
[31] X. Xu, L. He, H. Lu, L. Gao, and Y. Ji, "Deep adversarial metric learning for cross-modal retrieval," *World Wide Web*, vol. 22, no. 2, pp. 657–672, 2019.
[32] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier, "Collecting image annotations using Amazon's mechanical turk," in *Proc. NAACL HLT Workshop Creating Speech Lang. Data Amazon's Mech. Turk*, 2010, pp. 139–147.
[33] J. Krapac, M. Allan, J. Verbeek, and F. Juried, "Improving web image search results using query-relative classifiers," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 1094–1101.
[34] R. Rosipal and N. Krämer, "Overview and recent advances in partial least squares," in *Subspace, Latent Structure and Feature Selection* (Lecture Notes in Computer Science). Berlin, Germany: Springer, 2005.
[35] A. Sharma, A. Kumar, H. Daume, and D. W. Jacobs, "Generalized multiview analysis: A discriminative latent space," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2160–2167.
[36] K. Wang, R. He, L. Wang, W. Wang, and T. Tan, "Joint feature selection and subspace learning for cross-modal retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 10, pp. 2010–2023, Oct. 2016.
[37] J. Wu, Z. Lin, and H. Zha, "Joint latent subspace learning and regression for cross-modal retrieval," in *Proc. 40th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2017, pp. 917–920.
[38] Z. Liang, B. Ma, G. Li, Q. Huang, and Q. Tian, "Generalized semi-supervised and structured subspace learning for cross-modal retrieval," *IEEE Trans. Multimedia*, vol. 20, no. 1, pp. 128–141, Jan. 2018.
[39] Y. Wu, S. Wang, W. Zhang, and Q. Huang, "Online low-rank similarity function learning with adaptive relative margin for cross-modal retrieval," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2017, pp. 823–828.



SHUNXIN ZHENG received the bachelor's degree from the School of Information Science and Engineering, Shandong Normal University, Jinan, in 2017, where he is currently pursuing the master's degree in computer software and theory. His research interests include multimedia, machine learning, and signal processing. He is a Student Member of the CCF.

HUAXIANG ZHANG received the Ph.D. degree from Shanghai Jiao Tong University, in 2004. He was an Associate Professor with the Department of Computer Science, Shandong Normal University, China, from 2004 to 2005. He is currently a Professor with the School of Information Science and Engineering, Institute of Data Science and Technology, Shandong Normal University. He has authored over 170 journal and conference papers. He holds ten invention patents. His current research interests include machine learning, pattern recognition, evolutionary computation, cross-media retrieval, Web information processing, and bioinformatics.

YUDAN QI received the B.S. degree in computer science and technology from Zaozhuang University, China, in 2016. She is currently pursuing the master's degree with the School of Information Science and Engineering, Shandong Normal University. Her research interests include cross-modal retrieval, machine learning, and image processing. She is a Student Member of the CCF.



BIN ZHANG received the bachelor's degree from the School of Information Science and Engineering, Shandong Normal University, Jinan, in 2015, where he is currently pursuing the master's degree in computer software and theory. His research interests include multimedia, machine learning, and signal processing. He is a Student Member of the CCF.