Semi-supervised Cross-media Feature Learning via Efficient $L_{2,q}$ Norm

Zhikai Zong, Aili Han* and Qing Gong
Department of Computer Science and Technology, Shandong University
Weihai 264209, China
[e-mail: hanal@sdu.edu.cn, zzksdu@163.com]
*Corresponding author: Aili Han

Received April 5, 2018; revised August 17, 2018; accepted October 22, 2018; published March 31 2019

Abstract

With the rapid growth of multimedia data, research on cross-media feature learning has significance in many applications, such as multimedia search and recommendation. Existing methods are sensitive to noise and edge information in multimedia data. In this paper, we propose a semi-supervised method for cross-media feature learning by means of $L_{2,q}$ *norm* to improve the performance of cross-media retrieval, which is more robust and efficient than the previous ones. In our method, noise and edge information have less effect on the results of cross-media retrieval and the dynamic patch information of multimedia data is employed to increase the accuracy of cross-media retrieval. Our method can reduce the interference of noise and edge information and achieve fast convergence. Extensive experiments on the XMedia dataset illustrate that our method has better performance than the state-of-the-art methods.

Keywords: Cross-media retrieval, semi-supervised regularization, $L_{2,q}$ norm, sparse regularization

This work is supported by the Shandong Provincial Natural Science Foundation of China under Grant No. ZR2016FM20. The authors appreciate that the Team of Prof. Peng in Peking University has provided the source codes of the previous methods JRL and S2UPG and the XMedia dataset features.

http://doi.org/10.3837/tiis.2019.03.016 **ISSN : 1976-7277**

1. Introduction

With the rapid growth of multimedia data, the efficient cross-media retrieval techniques are needed in many applications. Some content-based retrieval methods for a single type of media data arose in the last decade [1-4], such as text retrieval [5], image retrieval [6], audio retrieval [7], and video retrieval [8]. These retrieval methods for a single type of media data cannot make full use of the information on multiple types of media data when used in cross-media retrieval, so they cannot achieve high performance for cross-media retrieval. Most of the existing methods for cross-media retrieval are from the methods for single-media retrieval, which do not take account of the correlation among multimedia data. Some methods add text annotations into images to achieve the retrieval between image and text, but they cannot effectively use the information on multiple types of media objects [9, 10].

Cross-media retrieval accepts one type of media data as a query and outputs the retrieval results with multiple types of media objects. If a Triumphal Arch picture is as a query, the cross-media retrieval results may be the related images, texts, audios, videos, and 3D models. The existing methods for cross-media retrieval either model semantic information of different types of media data [11], or only use the labeled information [12]. Rasiwasia et al. [13] explore the semantic information between texts and images at the semantic level by canonical correlation analysis (CCA), which models the pairwise correlation and semantic information separately. Zhai et al. [14] propose a heterogeneous matrix learning with joint graph regularization (JGRHML), which can calculate the similarity among different types of media data but relies on the labeled information. Zhai et al. [15] also propose a joint representation learning (JRL) method to exploit the semantic information and pairwise correlation in a unified domain, which models different types of media data into different graphs separately. Peng et al. [16] propose a semi-supervised learning model with unified graph normalization, which employs a joint graph to model five types of media data, which is sensitive to the edge and noise information in multimedia data. The main challenges in this field are that the cross-media retrieval is with low accuracy, and the retrieval results are susceptible to edge and noise information in multimedia data.

Based on the above methods [12-16], we propose an improved semi-supervised cross-media feature learning method (SCFL) in this paper. We integrate different types of media objects into a unified space by learning a semi-supervised robust joint graph and utilize the patches of labeled data to increase the diversity of training samples. The SCFL method uses the $L_{2,q}$ *norm* to improve the performance of robustness against the edge and noise information in multimedia data and achieves better performance of retrieval. Extensive experiments on the widely used XMedia dataset [15] demonstrate the effectiveness of our SCFL method against the state-of-the-art methods.

2. Related Work

Early works on feature learning include single feature learning [17,18] and multimodal feature learning (image and text [19,20], image and audio [21]). In the aspect of feature learning, Jian et al. [28] exploit Quaternionic Distance Based Weber Descriptor for detecting outliers in color image and apply hierarchical scheme for object detection. In order to obtain an accurate facial feature, Jian et al. [29] combine the quaternion number system and object cues for facial-feature detection. Li et al. [32] propose a weighted L_p norm spatial function for

class-specific feature spatial distribution. In recent years, the works on cross-media feature learning have developed rapidly. The cross-media feature learning method aims to explore the feature representation for different types of media data and the correlation between multimedia data.

Some cross-media methods are based on single-media feature learning (e.g., image feature learning). Hu et al. [17] propose a bag-of-visual-phrases model to establish the connection between image and visual words. Jian et al. [27] apply local features to image retrieval. Battiato et al. [18] find the spatial coherence between image representation and codeword, which enhance the effect of image feature representation. Other cross-media methods are based on multi-modal feature learning. Znaidia et al. [22] propose a bag-of-multimodal-words model which combines the heterogeneous information of text and pixel in a multimedia document. Liu et al. [21] propose a bag of audio words (BOA) model which proves the complementarity of video and audio information in video tampering detection.

Some existing methods for cross-media retrieval explore the feature representation for different types of media data. Rasiwasia et al. [13] use CCA to learn a primary space that deals with the correlation among heterogeneous data. Li et al. [23] introduce a cross-modal factor analysis (CFA) which learns a transformation matrix that calculates the correlation among the features from different types of media data. The CCF and CFA only explore the pairwise correlation between two types of media data and do not explore the semantic information among different types of media data. Huang et al. [24] propose a distance-preserving entity projection (DPER) to exploit the fine-grained correlation and generate a unified representation of multimedia content. The cross-media correlation is frail at the entity level. Yuan et al. [25] propose a recursive pyramid network with joint attention (RPJA) to learn the significance of the fine region. In order to get rich local information, Jian et al. [30] use an improved wavelet-based salient-patch detector to detect patches.

3. Cross-media Feature Learning via Efficient $L_{2,q}$ *norm*

Our cross-media feature learning method via efficient $L_{2,q}$ norm is an improvement on the former feature learning method [16]. In this section, we first introduce the former method [16] and then present our SCFL method, including the objective function, the optimization of the objective function, and the SCFL algorithm. Compared with [16], we adopt the $L_{2,q}$ norm instead of Frobenius norm in loss function to reduce the influence of noise and edge information.

Cross-media retrieval consists of two phases: cross-media feature learning and similarity among the features data in different media types. We do research on cross-media feature learning to learn robust projection matrix in this paper. We adopt the $L_{2,q}$ *norm* instead of Frobenius norm in the objective function to improve the performance for cross-media feature learning. The exponential q in $L_{2,q}$ *norm* is used to reduce the interference of the edge and noise information on retrieval results.

The set of labeled media data with multiple types is represented as $D=\{\mathrm{D}^{(1)},\cdots,\mathrm{D}^{(t)}\}$, where t is the number of media types, the set of i^{th} media is $x_p^{(i)}$ denotes the p^{th} data in the set of media data with i^{th} type, and $y_p^{(i)}$ is the label of $x_p^{(i)}$. The set of unlabeled media data is represented as $D^*=\{\mathrm{D}^{(1)*},\cdots,\mathrm{D}^{(s)*}\}$, where $D^{(i)*}=\{x_p^{(i)}\}_{p=n^{(i)}+1}^{n^{(i)}+m^{(i)}}$, $m^{(i)}$ is the number of the i^{th} unlabeled media data.

Each type of media object is divided into several patches in [16]. The image is divided into 3*3 patches, the text is divided into sentences, audio and video are divided into fragments in

the same size, 3D model is divided into 47 parts according to different perspectives in [26]. Patches separated from media objects contribute to highlight local information of media objects, and improve the accuracy of cross-media retrieval.

3.1 The former cross-media feature learning

Peng et al. [16] propose a cross-media feature learning with unified patch graph regularization to integrate different types of media data into a unified graph, in which the objective function is represented as follows.

$$\arg\min_{P^{(1)}\cdots P^{(s)}} \; loss(P^{(1)},\cdots,P^{(s)}) + \sum_{r=1}^{s}\lambda\left\|P^{(r)}\right\|_{2,1} + \alpha\Phi(O) \tag{1}$$

where $P^{(i)}$ is the projection matrix for ith media data, $\Phi(O)$ is the hypergraph regularization term. The matrix O is defined as:

$$O=\left(P^{(1)^T}X_a^{(1)}, P^{(1)^T}W_1^{(1)},\cdots,P^{(1)^T}W_{j^{(1)}}^{(1)},\cdots,P^{(s)^T}X_a^{(s)}, P^{(s)^T}W_1^{(s)},\cdots,P^{(s)^T}W_{j^{(s)}}^{(s)}\right) \tag{2}$$

where $X_a^{(i)}$ is the ith media data, $W_{j^{(m)}}^{(n)}(m,n\in[1,\cdots,s])$ is the $j^{(m)}$ th patch in nth media data. the hypergraph regularization term $\Phi(O)$ deals with both global information and local information of media objects. $\Phi(O)$ is defined as:

$$\Phi(O)=\sum_{r=1}^{s}\sum_{k=1}^{s}tr\left(P^{(r)T}\left(X_a^{(r)}+\sum_{t=1}^{j^{(r)}}W_t^{(r)}\right)L_{rk}\times\left(X_a^{(r)}+\sum_{t=1}^{j^{(r)}}W_t^{(r)}\right)^T P^{(k)}\right) \tag{3}$$

The loss function in Peng et al. [16] is defined as follows:

$$loss(P^{(1)},\cdots,P^{(s)})=\sum_{r=1}^{s}\left(\sum_{p=1}^{n^{(r)}}\left\|\frac{P^{(r)^T}}{1+j^{(r)}}\left(x_p^{(r)}+\sum_{q=1}^{j^{(r)}}w_{p,q}^{(r)}\right)+bE_c-y_p^{(r)}\right\|_F^2\right) \tag{4}$$

where $w_{p,q}^{(r)}$ is the qth number of $W_{j^{(m)}}^{(n)}$, the vector E_c is a unit column vector, the label of $x_p^{(r)}$ is $y_p^{(r)}$, and $\|A\|_F$ represents the Frobenius *norm* of matrix A. By setting the derivative of the formula (1) with respect to $p^{(r)}$ to 0, the final projection matrix in Peng et al. [16] is defined as follows:

$$\begin{aligned}P^{(r)}=&\left(Z^{(r)}H_n{H_n}^TZ^{(r)T}+\lambda D^{(r)}+\alpha Z_a^{(r)}L_{rr}Z_a^{(r)T}\right)^{-1}\\&\times\left(Z^{(r)}H_nY^{(r)T}-\alpha\sum_{j=1\&\,j\neq i}^{s}P^{(j)T}Z_a^{(j)}L_{rj}^TZ_a^{(r)T}\right)\end{aligned} \tag{5}$$

where $Z^{(r)}$ is the matrix of the labeled set, the matrix H_n is a centering matrix, and the matrix $D^{(r)}$ is a diagonal matrix for the rth media data.

The cross-media similarity in [16] is defined as follow: the media objects and their patches $x_p^{(r)}$ are projected into a joint feature space $o_p^{(r)}=P^{(r)T}x_p^{(r)}$. It is necessary to determine whether a query and the retrieved results belong to one category in the joint space. Media objects in the set of test samples are divided into several patches which can provide rich local information to improve the performance of cross-media retrieval. Each media object is determined by the

voting of its neighbors, and KNN classifier is applied to find the *k* nearest neighbors of each media object. The similarity of media objects o_p and o_q are defined as follows:

$$sim(o_p, o_q) = \frac{1}{2}\sum_l p(y_p = l|o_p)p(y_q = l|o_q) + \frac{1}{2}\underset{o_r \in S(o_p) \wedge o_s \in S(o_q)}{mean}\sum_l p(y_r = l|o_r)\, p(y_s = l|o_s) \tag{6}$$

where y_x is the label of o_x, $S(o_q)$ is the patch set of o_q, and the probability $p(y_q = l|o_q)$ is the probability of o_q belonging to label *l*, which is represented as follows:

$$p(y_q = l|o_q) = \frac{\sum_{o \in N_k(o_q) \wedge y=l} \sigma\left(-\|o_p - o\|_2\right)}{\sum_{o \in N_k(o_q)} \sigma\left(-\|o_p - o\|_2\right)} \tag{7}$$

where $N_k(o_q)$ denotes the *k* nearest neighbors of o_q in the training dataset. The function $\sigma(x)$ is defined as $\sigma(w) = (1+\exp(-w))^{-1}$. The similarity function $sim(o_p, o_q) \in [0,1]$, where $sim(o_p, o_q)$=1 indicates that all the nearest neighbors belong to the same category. $Sim(o_p, o_q)$=0 means that all the nearest neighbors and query media do not belong to the same category.

3.2 The objective function based on $L_{2,q}$ *norm*

We define the objective function to obtain more robust projection matrix. We use $L_{2,q}$ *norm* instead of Frobenius *norm* [16] in the objection function. The objective function of our method is as follows.

$$\underset{P^{(1)} \cdots P^{(s)}}{\arg\min}\ loss(P^{(1)}, \cdots, P^{(s)}) + \sum_{r=1}^{s} \lambda \left\|P^{(r)}\right\|_{2,q}^{q} + \alpha\Phi(O) \tag{8}$$

where $P^{(r)}$ is the projection matrix for the r^{th} media type, $1 \le r \le t$ and $loss(P^{(1)}, \cdots, P^{(s)})$ represents the loss function of the projection matrix. The parameter λ means a trade-off parameter, which is used to balance the relationship between the objective function and projection matrix. Compare with the objective function in [16], our method uses $L_{2,q}$ *norm* instead of Frobenius *norm* in loss function and $L_{2,q}$ *norm* instead of $L_{2,1}$ *norm* in the second part of the objective function.

Let $p^{(r)} \in \mathbb{R}^{d(r) \times c}$, $L_{r,q}$ *norm* of $P^{(r)}$ is defined as follows.

$$\left\|P^{(r)}\right\|_{r,q} = \left(\sum_{i=1}^{d(r)}\left(\sum_{j=1}^{c}\left|p_{ij}^{(r)}\right|^{r}\right)^{\frac{q}{r}}\right)^{\frac{1}{q}} \tag{9}$$

when *r*=2, *q*=1, $L_{r,q}$ *norm* can be instantiated as $L_{2,1}$ *norm*; when 0<*q*<1, $L_{2,q}$ *norm* not only makes $P^{(r)}$ sparser, but also makes loss function more robust to edge and noise media data.

The training dataset has constructed a graph in the following. The data in the training dataset is represented vertices in the graph, and the intensity of these edges is represented similarity between the data in the training set. The set on the edges is expressed as affinity matrix $\boldsymbol{W}$.

The energy function $E(f)$ of the mapping function *f* by means of $\boldsymbol{W}$ is defined as:

$$E(f)=\sum_{i,j}W_{ij}(f(x_i)-f(x_j))^2=\boldsymbol{f}^T\boldsymbol{L}\boldsymbol{f} \tag{10}$$

where $\boldsymbol{L}$ is the Laplacian matrix, $\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{W}$, $D=\mathrm{diag}(d_1,\cdots d_n)$, $d_i=\sum_{j=1}^{n}W_{ij}$.

Our object is to predict the unlabeled vertices in the graph by the labeled ones. The labeled data, the unlabeled data, and their patches are applied to construct a hypergraph, $G=\{\upsilon,e,H,\omega\}$, where υ represents the set of vertex, e is the set of the hyperedge, H is the incidence matrix between vertices and hyperedges, and ω represents the set of hyperedge weights. The Laplacian matrix $\boldsymbol{L}$ in [16] is defined as follows.

$$L=I-D_v^{-\frac{1}{2}}H\omega D_e^{-1}H^TD_v^{-\frac{1}{2}} \tag{11}$$

where I is the unit matrix, D_v is the diagonal matrix of vertex-degrees, in which the element is $d(c_i)=\sum_{e_j\in\varepsilon}w(e_j)H_{ij}$, D_e denotes the matrix of hyperedge degrees.

The loss function in our method is by $L_{2,q}$ *norm* instead of Frobenius *norm* to reduce the interference of edge and noise data. The loss function is as follows.

$$\begin{aligned}loss(P^{(1)},\cdots,P^{(s)})&=\sum_{r=1}^{s}\left(\sum_{p=1}^{n^{(r)}}\frac{1}{1+j^{(r)}}\left\|P^{(r)^T}x_p^{(r)}+b1_c-y_p^{(r)}+\sum_{q=1}^{j^{(r)}}\left(P^{(r)^T}w_{p,q}^{(r)}+b1_c-y_p^{(r)}\right)\right\|_{2,p}^{p}\right)\\&=\sum_{r=1}^{s}\left(\sum_{p=1}^{n^{(r)}}\left\|\frac{P^{(r)^T}}{1+j^{(r)}}\left(x_p^{(r)}+\sum_{q=1}^{j^{(r)}}w_{p,q}^{(r)}\right)+b1_c-y_p^{(r)}\right\|_{2,p}^{p}\right)\end{aligned} \tag{12}$$

where $x_p^{(r)}$ is the p^{th} media object in the set of r^{th} media data, $y_p^{(r)}$ is the label of $x_p^{(r)}$, $w_{p,q}^{(r)}$ is the qth patch of the p^{th} media object in the set of i^{th} media type. b is a dynamic constant, and 1_c is a column vector in which each element is 1.

If the result is not same as the labeled category, the exponential q in $L_{2,q}$ *norm* can make the loss function very small to reduce the interference of noise information on the loss function in the training dataset, make the object function converge fast.

The objective function is defined as follows.

$$\begin{aligned}\arg\min_{P^{(1)},\cdots,P^{(s)}}&\sum_{r=1}^{s}\left(\sum_{p=1}^{n^{(r)}}\left\|\frac{P^{(r)^T}}{1+j^{(r)}}\left(x_p^{(r)}+\sum_{q=1}^{j^{(r)}}w_{p,q}^{(r)}\right)+b1_c-y_p^{(r)}\right\|_{2,p}^{p}\right)\\&+\lambda\sum_{r=1}^{s}\|P^{(R)}\|_{2,q}^{q}+\alpha\sum_{r=1}^{s}\sum_{k=1}^{s}tr\left(P^{(r)T}\left(X_a^{(r)}+\sum_{t=1}^{j^{(r)}}W_t^{(r)}\right)L_{rk}\times\left(X_a^{(r)}+\sum_{t=1}^{j^{(r)}}W_t^{(r)}\right)^TP^{(k)}\right).\end{aligned} \tag{13}$$

where we use $L_{2,q}$ *norm* instead of Frobenius *norm* in loss function and use $L_{2,q}$ *norm* instead of $L_{2,1}$ *norm* in the second part of the objective function.

3.3 The optimization of the objective function

We optimize the objective function to improve the performance of multimedia retrieval. In order to obtain t transformation matrix $P^{(r)}$ efficiently, we represent the first part in the formula (13) as $Z_{\cdot,p}^{(r)}$,

$$Z_{\cdot,p}^{(r)}=\frac{1}{1+j^{(r)}}\left(x_p^{(r)}+\sum_{q=1}^{j^{(r)}}w_{p,q}^{(r)}\right) \tag{14}$$

We denote $X_a^{(r)}+\sum_{t=1}^{j^{(r)}}W_t^{(r)}$ in the formula (13) as $Z_a^{(r)}$,

$$Z_a^{(r)} = X_a^{(r)} + \sum_{t=1}^{j^{(r)}} W_t^{(r)} \tag{15}$$

where $Z_{.,p}^{(r)}$ is the p^{th} column vector of $Z_a^{(r)}$.

Substituting (14) and (15) into (13), the objective function is improved as follows.

$$\begin{aligned}\arg\min_{P^{(1)},\cdots,P^{(s)}} \sum_{r=1}^{s}\Big(&\left\|P^{(r)T}Z^{(r)}+b1_c-Y^{(r)}\right\|_{2,q}^{q}+\lambda tr(P^{(r)T}D_1^{(r)}P^{(r)})\\&+\alpha\sum_{j=1}^{s}tr(P^{(r)T}Z_a^{(r)}L_{rj}Z_a^{(j)T}P^{(j)})\Big)\end{aligned} \tag{16}$$

By setting the derivative of the formula (16) with respect *b* to 0 in (13), the dynamic constant is obtained as follows.

$$b=\frac{1}{n}(Y^{(r)-}P^{(r)T}Z^{(R)})1_n \tag{17}$$

We observed from (17) that *b* is the difference between the result labels and the set of ground truth.

Substituting (17) into (16), we obtain the objective function as follows.

$$\arg\min_{P^{(1)},\cdots,P^{(s)}} \sum_{r=1}^{s}\left(\left\|P^{(r)T}Z^{(r)}H_n-Y^{(r)}H_n\right\|_{2,q}^{q}+\lambda tr(P^{(r)T}D_1^{(r)}P^{(r)})+\alpha\sum_{j=1}^{s}tr(P^{(r)T}Z_a^{(r)}L_{rj}Z_a^{(j)T}P^{(j)})\right) \tag{18}$$

where H_n is a centering matrix, $H_n=I-1/(n^{(r)})1_n1_n^T$. By setting the derivative of the formula (18) with respect to $P^{(r)}$ to 0, we obtain the final matrix as follows.

$$\begin{aligned}P^{(r)}=&\left(D_2Z^{(r)}H_nH_n{}^TZ^{(r)T}+\lambda D_1+\alpha Z_a^{(r)}L_{rr}Z_a^{(r)T}\right)^{-1}\\&\times\left(D_2Z^{(r)}H_nH_n{}^TY^{(r)T}-\alpha\sum_{j=1\&j\neq i}^{s}P^{(j)T}Z_a^{(j)}L_{rj}^TZ_a^{(r)T}\right)\end{aligned} \tag{19}$$

where $D_1^{(r)}$ is a diagonal matrix in which the diagonal elements are $d1_{jj}^{(r)}=1/\left(2\left\|P_i^{(r)}\right\|\right)^q$, $D_2^{(r)}$ is a diagonal matrix in which the diagonal elements are $d2_{jj}^{(r)}=1/\left(2\left\|\left(P^{(r)T}Z^{(r)}H_n-Y^{(r)}H_n\right)_i\right\|\right)^q$.

Comparing with Peng's method [16], our final projection matrix adds the sparse matrix D_1 and D_2. The exponential *q* in $L_{2,q}$ *norm* used in our method makes the projection matrix more sparse and converges faster.

3.4 The SCFL algorithm

We proposed a cross-media feature learning method, marked as SCFL algorithm which adapts the $L_{2,q}$ *norm* instead of Frobenius *norm* in the objective function. The construction of hypergraph is in the same method as that in [16]. We first use a set of random numbers to initialize $P_0^{(r)}$, and update $P_{t+1}^{(r)}$ according to the formula (19). The SCFL algorithm is as follows.

Algorithm 1. Semi-supervised cross-media feature learning

Input: the set of labeled data $Z^{(r)}$;

the set of labelled and unlabelled data $Z_a^{(r)}$;

the label $Y^{(r)}$ of $Z^{(r)}$;

Parameter λ, α;

1) Construct a hypergraph $G=\{\upsilon,e,H,\omega\}$ by the method in [16];

2) Initialize the projection matrix $P_0^{(r)}$ by a set of random values, and set the iteration times t=0;

3) While not converge do

2.1) Compute the Laplacian matrix $\boldsymbol{L}$ of the hypergraph $G=\{\upsilon,e,H,\omega\}$ according to $L=I-D_v^{-\frac{1}{2}}H\omega D_e^{-1}H^TD_v^{-\frac{1}{2}}$:

$I \leftarrow sparse(diag(sparse(ones(1,size(Row,1)))))$

for i = 0 : 1 : the patch numbers

H = *sparse*(*zeros*(*m*,*n*))

for j =1 : *size*(H)

ind=*indx*(1 , 2 : j+1);

D_v(*ind*) = D_v(*ind*) + 1;

H(*ind*, j) = 1;

end

end

2.2) Calculate the elements of the diagonal matrixs $D_1^{(r)}$ and $D_2^{(r)}$ according to $d1_{jj}^{(r)}=1/\left(2\left\|P_i^{(r)}\right\|\right)^q$ and $d2_{jj}^{(r)}=1/\left(2\left\|\left(P^{(r)T}Z^{(r)}H_n-Y^{(r)}H_n\right)_i\right\|\right)^q$, in which q is the exponential of $L_{2,q}$ *norm*.

for i = 1:*dimension*

$A1 \leftarrow P^{(r)T}Z^{(r)}H_n-Y^{(r)}H_n$

D_1(*i*, *i*) = 1/(2*norm(*A*1(*i*, :)));

D_2 (i, i) = 1/(2*norm(*P*(*i*, :)));

end

D_1=(1/q)*D_1.^q;

D_2=(1/q)*D_2.^q;

2.3) Update $P_{t+1}^{(r)}$ ….according to (19).

2.4) t=t+1;

4) Until Convergence

Output: The final projection matrix $P^{(r)}$.

In the SCFL algorithm, the termination condition is set as: the ratio change between two iterations is less than 1.5% or the maximum iteration number is 3. Experimental results demonstrate that our algorithm converges faster than that in [16]. The mapping matrix we obtained is more sparse than that obtained by [16]. **Table 1** and **Table 2** show an 8*8 block in the mapping matrix of text -> audio by the method in [16] and by our method, respectively. Form **Table 1** and **Table 2**, it can be known that our matrix is closer to zero, which means it is more sparse than that obtained by [16].

Table 1. The 8*8 block in the mapping matrix of [16]

-0.0899	-0.0385	-0.0045	0.0024	0.0347	-0.0181	0.0163	0.0364
-0.0259	-0.0199	0.0120	0.0220	-0.0096	0.0292	-0.0373	-0.0026
0.0013	-0.0278	5.376e-05	-1.652e-05	0.0318	-0.0091	-0.0282	0.0286
0.0251	0.0077	0.0053	-0.0053	-0.0037	-0.0016	-00053	0.0204
-0.0026	0.0084	0.0063	-0.0203	0.0256	0.0035	0.0205	0.0073
0.0053	0.0311	0.0051	0.0034	-0.0049	0.0029	-0.0375	0.0099
0.0016	-0.0061	-0.0010	-0.0032	0.0088	-8.1234e-04	0.0305	-0.0051
-0.0038	0.0191	0.0145	-0.0010	-5.1767-e-04	0.0064	-0.0178	-0.0044

Table 2. The 8*8 block in the mapping matrix of our method

-0.0759	-0.0336	-0.0054	0.0026	0.0256	-0.0136	0.0144	0.0271
-0.0209	-0.0130	0.0112	0.0156	-0.0089	0.0209	-0.0319	0.0021
0.0013	-0.0255	-6.1844e-04	-8.8954e-04	0.0244	-0.0055	-0.0177	0.0176
0.0196	0.0068	0.0037	-0.0036	-0.0068	-8.0560e-04	-0.0056	0.0131
8.594e-04	0.0076	0.0045	-0.0514	0.0199	0.0041	0.0125	0.0026
5.996e-04	0.0233	0.0038	0.0020	-0.0064	0.0021	-0.0310	0.0048
5.436e-04	-0.0069	-0.0025	-1.8177e-04	0.0070	-0.0016	0.0256	-0.0069
-0.0042	0.0157	0.0133	-8.8369e-04	-0.0013	0.0075	-0.0173	-0.0029

4. Experimental Results and Analysis

We conduct experiments on the XMedia dataset, which is a widely used dataset to verify the effectiveness of our SCFL algorithm. Compared with the state-of-the-art methods, the experimental results on the XMedia dataset show the effectiveness of our proposed method.

4.1 XMedia Dataset

XMedia dataset [15]: The dataset consists of 5000 images, 5000 texts, 500 3D models, 500 videos, and 1000 audios. All of the media objects are collected on the internet. The data in the XMedia dataset contains 20 categories, in each category there are 600 media objects. These media objects are randomly divided into the training dataset and the test dataset. The training dataset contains 9600 media objects, and the test dataset contains 2400 ones. Some samples in the XMedia dataset are shown in **Fig. 1**.

In the XMedia dataset [15], image and text representation are generated in the same way. The Bag of Words model and top model are used to represent the image and text objects [13]. Each image is represented by the histogram of 128-codeword SIFT codebook, each text is represented by LDA model with 10-dimensional histogram, each audio is represented by 29-D MFCC features and each video is segmented into shots to extract keyframes, which are represented by the histogram of 128-codeword SIFT codebook. The similarity between video data and other media data is obtained by calculating the average similarity between all keyframes and other media data. We adopt 4700-D vector of LightField descriptors to represent the 3D model [26].

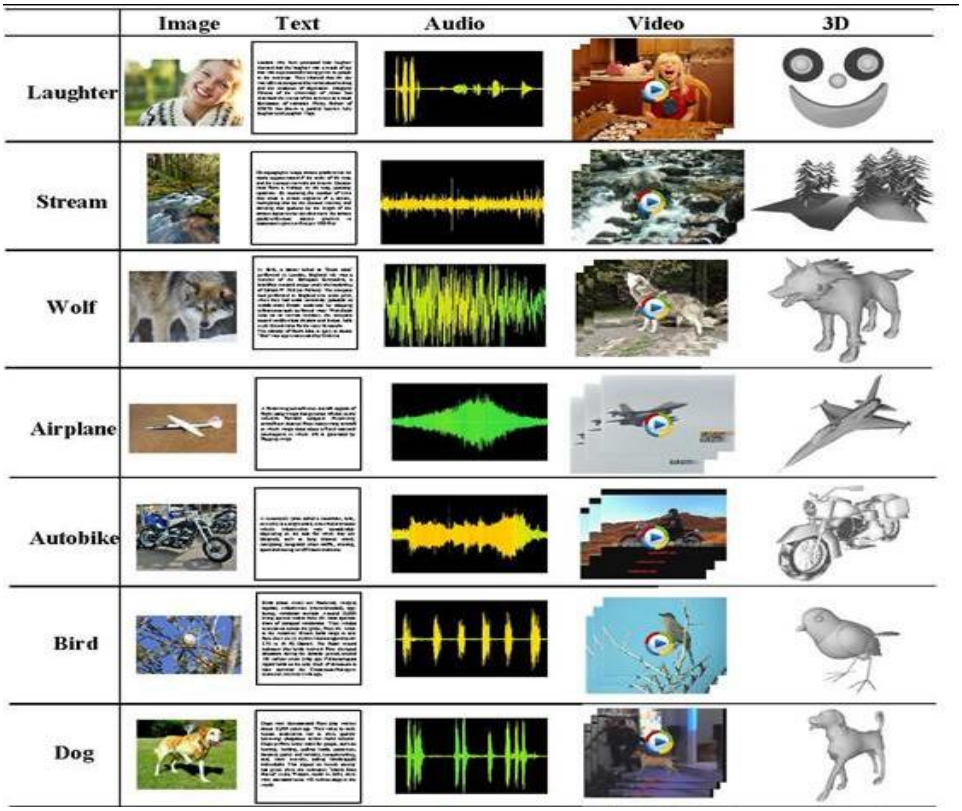


Fig. 1. Examples from seven categories of XMedia dataset, which is from *http://www.icst.pku.edu.cn/mipl/xmedia/*.

4.2 State-of-the-art Methods

In order to evaluate the proposed SCFL method, we compare our method with four start-of-art methods, which are summarized as:

[1] CCA+SMN [13]: It considers the correlation analysis and abstract representation between text and image data. The correlation analysis between image and text is obtained by CCA. Image and text are represented by abstract representation in general semantics.

[2] Joint Graph Regularized Heterogeneous Metric Learning (JGRHML) [14]: It learns a regularized heterogeneous matrix metric method of a joint graph, which is to compute the similarity of different types of the media object.

[3] Joint Representation Learning (JRL) [15]: It learns a sparse matrix, which maps different media data into a joint space. This method can explore the correlation and semantic information of different media objects.

[4] Semi-Supervised Cross-media Feature Learning Algorithm with Unified Patch Graph Regularization (S2UPG) [16]: It can map media objects and their patches into a hypergraph simultaneously, and patch information can emphasize important parts.

In order to evaluate our proposed SCFL method, two retrieval tasks (cross-media retrieval and signal-media retrieval) are considered. 1) cross-media retrieval. We can retrieve all the media data by submitting any types of media data. The retrieval results conclude all types of the media object(text, image, audio 3D model and video). 2) single-media retrieval. We apply the cross-media method to single media retrieval in order to show the advantage of our proposed SCFL method. Text retrieval, Image retrieval, audio retrieval, 3D model retrieval and video retrieval are evaluated in this task. The query and the retrieval results are the same types.

We adapt Mean Average Precision (MAP) to evaluate the retrieval result. The MAP of a set of queries is the average of the average precision (AP) for each query, and the average precision (AP) is defined as:

$$AP = \frac{1}{R}\sum_{k=1}^{m}\frac{R_k}{k}\times rel_k \tag{21}$$

where m is the sample number of the test dataset, R is the number of relevant items, and R_k is the number of the top k related items. When $rel_k=1$, the item ranking at the kth position is related and $rel_k=0$ otherwise.

4.3 Experimental Result comparison

We compare the proposed SCFL algorithm with state-of-the-art algorithms. To be fair, our experimental platform is Matlab 2015b, Intel Corel I7-4770 CPU 3.40Ghz, 8GB memory. **Table 3** shows the MAP scores of SCFL algorithm and three other state-of-the-art algorithms. Comparing with the latest cross-media algorithm, our proposed SCFL improves the average MAP from 29.2% to 32.4%. The query object can be any media type, and the retrieval results contain all the media types in the test dataset. JRL performs better than CCA+SMN, because JRL can explore the semantic information of media and correlation information in joint space. S2UPG makes full use of media objects and patches to improve the precision of retrieval. Retrieval performance of SCFL is better than other three current state-of-the-art methods, for the reason that SCFL applies $L_{2,q}$ *norm* to reduce the influence of noise and edge data on projection matrix. **Fig. 4** shows the PR curves of JRL, S2UPG and our proposed method on the XMedia dataset. We only show the two figures(text queries all and image queries all) due to the limitation of paper. It can be seen that the method can get higher precision at most recall levels. **Fig. 3** shows the cross-validation of parameters λ and α for the cross-media retrieval results. From **Fig. 3**, we can conclude that the cross-media retrieval results are not sensitive to parameters.

In single media retrieval, we use joint representation to compare the performance of SCFL and other algorithms. During the training stage, we learn the transformed matrix $P^{(i)}_{d^{(r)}\times c}$ for the original features of all media types. While during the testing stage, take text->text as an example, we project original feature $x_p^{(r)}$ into joint space by way of $o_p^{(r)} = P^{(r)T}x_p^{(r)}$, then single media retrieval is conducted in joint feature space. **Table 4** shows single media retrieval results on SCFL and other methods. The retrieval results show that SCFL performs better than any other methods mentioned on single media retrieval. This is because SCFL not only uses media object and media patch information but also reduces the influence of the noise data and edge data in the dataset. It is a pity that the SCFL is not applicable to single media retrieval of 3D objects.

Table 3. Cross-media retrieval on five types of media data

Dataset	Task	CCA+SMN	JRL	S^2UPG	SCFL(ours)
XMedia Dataset	Image->All	0.143	0.252	0.27	0.293
	Text->All	0.141	0.199	0.238	0.258
	Audio->All	0.125	0.197	0.202	0.215
	Video->All	0.116	0.152	0.207	0.225
	3D->All	0.082	0.181	0.25	0.303
Average		0.121	0.196	0.292	0.324

Table 4. Cross-media retrieval on five types of media data

Dataset	Task	CCA+SMN	JRL	S^2UPG	SCFL(ours)
XMedia Dataset	Image->Image	0.295	0.317	0.352	0.368
	Text->Text	0.197	0.212	0.277	0.329
	Audio->Audio	0.368	0.366	0.363	0.369
	Video->Video	0.385	0.299	0.334	0.331
Average		0.311	0.299	0.332	0.349

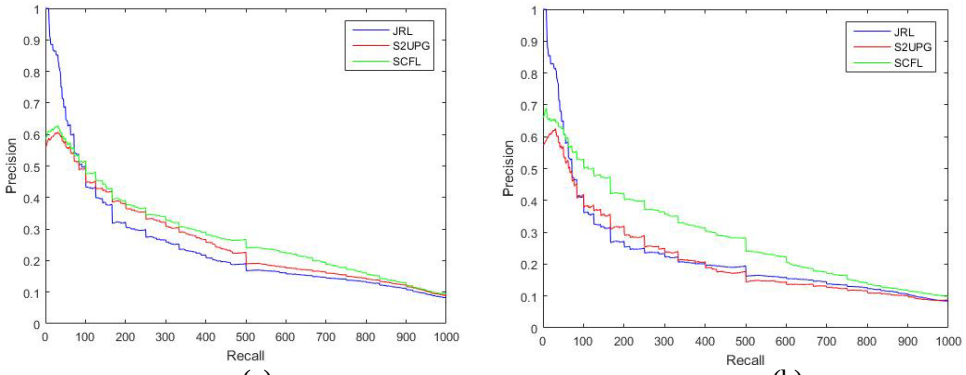


Fig. 2. Precision-Recall curves on the XMedia dataset. (a) Image->All. (b) Text->All.

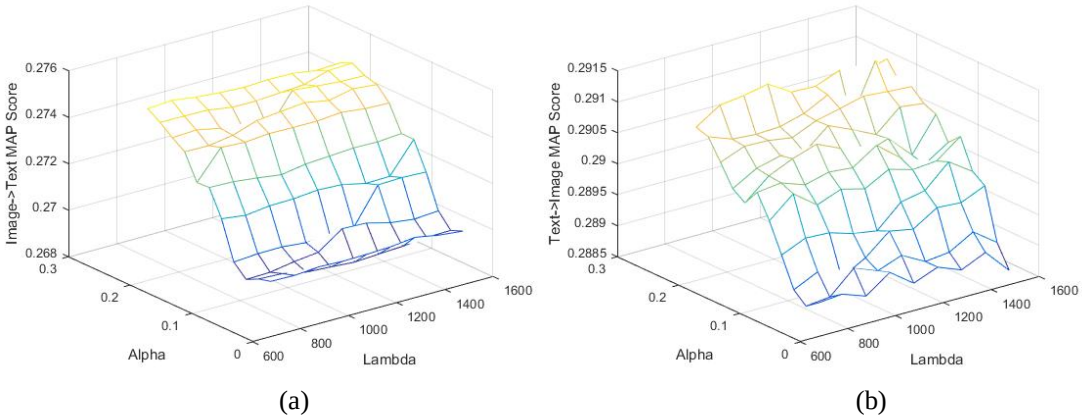


Fig. 3. Cross-media retrieval by cross-validation on the XMedia dataset for SCFL algorithm. (a) Image->Text. (b) Text->Image.

5. Conclusion

We propose a semi-supervised cross-media feature learning (SCFL) method which can reduce the influence of edge and noise information on cross-media retrieval result. The $L_{2,q}$ *norm* is used to make the SCFL method converge quickly. The SCFL method integrates sparse regularization to project all media into a joint feature space. Extensive experiments on XMedia dataset have illustrated the effectiveness of the SCFL method. Our future work will focus on how to choose more effective patch information and highlight the key parts of multimedia objects. The SCFL method can be applied to other applications.

References

[1] M. Lew, N. Sebe, C. Djeraba, and R. Jain, "Content-based multimedia information retrieval: State-of-the-art and challenges," *ACM Trans. Multimedia Comput. Commun., Applicat.*, vol. 2, no. 1, pp. 1–19, Feb. 2006. Article(CrossRefLink)

[2] H. Greenspan, J. Goldberger, and A. Mayer, "Probabilistic space-time video modeling via piecewise GMM," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 3, pp. 384–396, Mar. 2004. Article(CrossRefLink)

[3] J. Jeon, V. Lavrenko, and R. Manmatha, "Automatic image annotation and retrieval using cross-media relevance models," in *Proc. of 26th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr*, pp. 119–126, 2003. Article(CrossRefLink)

[4] Pan, J., Yang, H., Faloutsos, C., Duygulu, P, "Automatic multimedia cross-modal correlation discovery," in *Proc. of ACM International Conference on Knowledge Discovery and Data Mining*, SIGKDD, 2004. Article(CrossRefLink)

[5] A. Moffat and J. Zobel, "Self-indexing inverted files for fast text retrieval," *ACM Trans. Inf. Syst.*, vol. 14, no. 4, pp. 349–379, 1996. Article(CrossRefLink)

[6] J. Yu and Q. Tian, "Semantic subspace projection and its applications in image retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 4, pp. 544–548, Apr. 2008. Article(CrossRefLink)

[7] R. Typke, F. Wiring, and R. C. Veltkamp, "A survey of music information retrieval systems," in *Proc. of ISMIR*, pp. 153–160, 2005. Article(CrossRefLink)

[8] Y. G. Jiang. C. W. Ngo and J. Yang, "Towards optimal bag-of-features for object categorization and semantic video retrieval," in *Proc. of the 6th ACM international conference on Image and video retrieval*, pp. 494–501, Jul. 2007. Article(CrossRefLink)

[9] S. Moran and V. Lavrenko, "A sparse kernel relevance model for automatic image annotation," *International Journal of Multimedia Information Retrieval*, Vol. 3, no. 4, pp. 209–229, Nov 2014. Article(CrossRefLink)

[10] D. Grangier and S. Bengio, "A discriminative kernel-based model to rank images from text queries," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 8, pp. 1371–1384, Aug. 2008. Article(CrossRefLink)

[11] Y. T. Zhuang, Y. F. Wang, F. Wu, Y. Zhang, and W. M. Lu, "Supervised coupled dictionary learning with group structures for multi-modal retrieval," in *Proc. of 27th AAAI Conf. Artif. Intell.*, pp. 1070–1076, 2013. Article(CrossRefLink)

[12] X. Zhai, Y. Peng, and J. Xiao, "Cross-Modality Correlation Propagation for Cross-Media Retrieval," in *Proc. of Int. Conf. Comput. Vision*, pp. 2407-2414, Nov, 2011. Article(CrossRefLink)

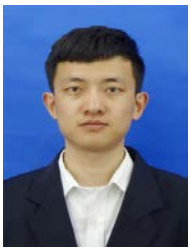
[13] N. Rasiwasia, J. Pereira, E. Coviello, G. Doyle, G. Lanckriet, R. Levy, et al., "A new approach to cross-modal multimedia retrieval," in *Proc. of ACM Int. Conf. Multimedia*, pp. 251–260, 2010. Article(CrossRefLink)

[14] X. Zhai, Y. Peng, and J. Xiao, "Heterogeneous metric learning with joint graph regularization for cross-media retrieval," in *Proc. of 27th AAAI Conf. Artif. Intell.*, pp. 1198–1204, 2013. Article(CrossRefLink)

[15] X. Zhai, Y. Peng, and J. Xiao, "Learning cross-media joint representation with sparse and semi-supervised regularization," *IEEE Trans. Circuits Syst. Video Technol*., vol. 24, no. 6, pp. 965–978, Jun. 2014. Article(CrossRefLink)
[16] Y. X. Peng, X. H. Zhai, Y. Z. Zhao, and X. Huang, "Semi-Supervised Cross-Media Feature Learning With Unified Patch Graph Regularization," *IEEE transactions on circuits and systems for video technology*, Vol. 26, no. 3, Mar 2016. Article(CrossRefLink)
[17] Y. Hu, X. Cheng, L.-T. Chia, X. Xie, D. Rajan, and A.-H. Tan, "Coherent phrase model for efficient image near-duplicate retrieval," *IEEE Trans. Multimedia*, vol. 11, no. 8, pp. 1434–1445, Dec. 2009. Article(CrossRefLink)
[18] S. Battiato et al., "Bags of phrases with codebooks alignment for near duplicate image detection," in *Proc. of 2nd ACM Workshop Multimedia Forensics, Secur., Intell*., pp. 65–70, 2010. Article(CrossRefLink)
[19] Y. X. Peng, X. H, and J. W. Qi, "Cross-Media Shared Representation by Hierarchical learning with Multiple Depp Networks," in *Proc. of the Tweety-Fifth International Joint Conference on Artificial Intelligence*, pp. 3846-3853. 2016. Article(CrossRefLink)
[20] S. Battiato, G. M. Farinella, G. Giuffrida, C. Sismeiro, and G. Tribulato, "Using visual and text features for direct marketing on multimedia messaging service," *Multimedia Tool &Application*, Vol. 42, no. 1, pp. 5-30, Mar. 2009. Article(CrossRefLink)
[21] Y. Liu, W.-L. Zhao, C.-W. Ngo, C.-S. Xu, and H.-Q. Lu, "Coherent bagof audio words model for efficient large-scale video copy detection," in *Proc. of ACM Int. Conf. Image Video Retr*., pp. 89–96, 2010. Article(CrossRefLink)
[22] A. Znaidia, A. Shabou, H. Le Borgne, C. Hudelot, and N. Paragios, "Bag-of-multimedia-words for image classification," in *Proc. of 21st Int. Conf. Pattern Recognit. (ICPR)*, pp. 1509–1512, Nov. 2012. Article(CrossRefLink)
[23] D. Li, N. Dimitrova, M. Li, and I. K. Sethi, "Multimedia content processing through cross-modal association," in *Proc. of 11th ACM Int. Conf. Multimedia (ACM-MM)*, pp. 604–611, 2003. Article(CrossRefLink)
[24] Lei Huang and Yuxin Peng, "Cross-media retrieval by exploiting fine-grained correlation at entity level," *Neurocomputing*, Vol. 236, pp. 123-133, May, 2017. Article(CrossRefLink)
[25] Yuxin Yuan and Yuxin Peng, "Recursive pyramid network with joint attention for cross-media retrieval," in *Proc. of 24th International Conference on Multimedia Modeling (MMM)*, pp. 405−416, Bangkok, Thailand, Feb. 5-7, 2018. Article(CrossRefLink)
[26] D.-Y. Chen, X.-P. Tian, Y.-T. Shen, and M. Ouhyoung, "On visual similarity based 3D model retrieval," *Comput. Graph. Forum*, vol. 22,no. 3, pp. 223–232, 2003. Article(CrossRefLink)
[27] M. W. Jian, Y. L. Yin, J. Y. Dong and K. M. Lam, "Content-based image retrieval via a hierarchical-local-feature extraction scheme," *Multimedia Tools and Applications*, 53 (1), May 2018. Article(CrossRefLink)
[28] M. W. Jian, K. M. Lam, J. Y. Dong and L. L. Shen, "Visual-patch-attention-aware Saliency Detection," *IEEE Transactions on Cybernetics*, Vol. 45, No. 8, pp. 1575-1586, 2015. Article(CrossRefLink)
[29] M. W. Jian, K. M. Lam and J. Y. Dong, "Facial-Feature Detection and Localization Based on a Hierarchical Scheme," *Information Sciences*, vol. 262, pp. 1-14, 2014. Article(CrossRefLink)
[30] M. W. Jian, Q. Qi, J. Y. Dong, Y. L. Yin and K. M. Lam, "Integrating QDWD with Pattern Distinctness and Local Contrast for Underwater Saliency Detection," *Journal of Visual Communication and Image Representation*, vol. 53, pp. 31–41, 2018. Article(CrossRefLink)
[31] M. W. Jian, Q. Qi, J. Y. Dong, X. Sun, Y. J. Sun and K. M. Lam, "Saliency Detection Using Quaternionic Distance Based Weber Local Descriptor and Level Priors," *Multimedia Tools and Applications*, 77 (11), pp. 14343–14360, 2018. Article(CrossRefLink)
[32] T. Li, Z. J. Meng, B. B. Ni, J. B. Shen and M. Wang, "Robust geometric źp-norm feature pooling for image classification and action recognition," *Image Vision Comput*., pp. 64-76. 2016. Article(CrossRefLink)

Zhikai Zong, he was born in Shandong province, China, in 1993. He received his B.E. degree in electronic information engineering from Shandong University of Technology, China, in 2016. He is currently pursuing his M.E. degree at Shandong University, Weihai, China.
His current research interests include digital image processing and computer vision.

Aili Han is a Professor in Shandong University, China. She is a member of the Technical Committee on Computer Vision, CCF, and a member of the Special Interest Group on Computer Science Education China Council, ACM. She received her Ph.D. degree in Computer Software from Shandong University, China. Her research interests include intelligent computing, image processing, and multimedia retrieval. She has published over 30 papers in refereed international journals and conference proceedings.



Qing Gong, he was born in Shanxi province, China, in 1993. He received his B.E. degree in Measurement and controlling technology and instrumentation from Shandong University, China, in 2016. He is currently pursuing his M.E. degree at Shandong University, China.
His current research interests include digital image processing and computer vision.