ISSN 1330-3651 (Print), ISSN 1848-6339 (Online)
https://doi.org/10.17559/TV-20190730110003
Original scientific paper

Cross-Media Semantic Matching based on Sparse Representation

Gongwen XU, Aidong ZHAI, Jing WANG, Zhijun ZHANG, Xiaomei LI

Abstract: With the rapid growth of multi-modal data, cross-media retrieval has aroused many research interests. In this paper, the cross-media retrieval includes two tasks: query image retrieves relevant text and query text retrieves relevant images. With the development of sparse representation, two independent sparse representation classifiers are used to map the heterogeneous features of images and texts into their common semantic space before implementing similarity comparison. The proposed method makes full use of semantic information, and it is effective in the retrieving task. The performance of this method was evaluated on Wiki dataset, NUS-WIDE dataset, Wiki dataset with CNN features and Pascal dataset with CNN features. The experimental results validate its effectiveness compared with several state-of-the-art algorithms on the Mean Average Precision and other performance indexes.

Keywords: cross-media retrieval; semantic matching; sparse representation

1 INTRODUCTION

With the rapid development of multi-modal data, it is very useful for people to understand and mine information contained in data using the relevant information of multi-modal data [1]. Firstly, through the analysis of pictures and textual comments on the Internet social network, it is easy to effectively understand the people's opinion of the current hot topic or predicting social problems affecting public safety. Secondly, with the development of e-commerce, some online shopping websites such as Taobao and Jingdong, have become an inseparable part of people's lives. Through the analysis of products' style, function and user review information, the e-commerce websites can be adjusted to the marketing strategy. At the same time, the development of the Internet has also changed the way of people's work, learning and entertainment. People begin to use the image to retrieve the similar images or texts, or use the keywords and textual document to retrieve the related images and videos. Through the correlation analysis of the multi-modal data, better service can be provided for Internet users, and improve the efficiency of people's study and work. Therefore, using the semantic correlation of multi-modal media data, analyzing the semantic content of them has become an important research topic in the fields of cross-media retrieval and pattern recognition.

Currently, the correlation modeling among multi-modal data still faces some challenges [2]. On the one hand, the low-level features of different modality data (e.g. an image and a section of text) are heterogeneous. However, the heterogeneous media data can be unified at the semantic level, i.e. semantic consistency of heterogeneous media data. Traditional media technology ignores it, so it is difficult to deal with the heterogeneous data. On the other hand, the correlation modeling of multi-modal media data also needs the semantic information of isomorphic media data (e.g. several images are isomorphic to each other). Although this kind of data is often consistent in feature representation, how to mine correlation information of isomorphic media data using the semantic information is another important problem for cross-media correlation modeling.

In this paper, two independent sparse representation classifiers were used to map the heterogeneous features of images and texts into their common semantic space before implementing similarity comparison. And with their outputs, the common semantic space of images and texts can be obtained further applying cross-media retrieval. This method is named Sparse Representation-Semantic Matching (SRSM) in this paper. Compared with other cross-media retrieval methods, this method considers the semantic information of isomorphic media data as well as semantic consistency of heterogeneous media data. What is more, this method makes full use of semantic information, and it is effective.

The rest of the paper is shown as follows. Related works are introduced in Section 2. The details of SRSM are described in Section 3. The experimental results are shown in Section 4, and the conclusion is made in Section 5.

2 RELATED WORK

2.1 Cross-Media Retrieval based on Subspace Learning

Currently, a significant number of cross-media retrieval works focus on subspace learning method. This kind of method aims to learn a latent subspace of different modalities of media data (shown as Fig. 1). And it can be divided into four parts which are shown as follows:

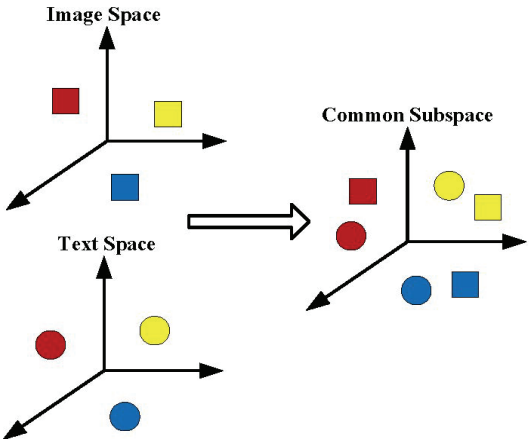

Figure 1 The framework of subspace learning method

Subspace learning based on projection: This kind of subspace learning method uses the feature mapping to extract the latent subspaces shared by different modalities of media data. It can be divided into linear projection methods (e.g. Canonical Correlation Analysis, CCA [3]

and Partial Least Squares, PLS [4]) and nonlinear projection methods (Kernel Canonical Correlation Analysis, KCCA [5] and Deep Canonical Correlation Analysis, DCCA [6]).

Subspace learning based on matrix factorization: This kind of subspace learning method uses the matrix factorization to extract the basis vectors of latent subspaces shared by different modalities of media data. It can be divided into nonnegative factorization methods (e.g. Joint Shared Nonnegative Matrix Factorization, JSNMF [7]) and eigen decomposition-based methods (e.g. Multi-Output Regularized Feature Projection, MORFP [8]).

Subspace learning based on task: This kind of subspace learning method learns multiple related tasks at the same time so that it can improve the overall generalization performance of each task. It can be divided into Multi-task learning methods (e.g. Alternating Structure Optimization, ASO [9] and Convex Multi-Task Feature Learning, CMTFL [10]), Multi-label learning methods (e.g. Shared-Subspace Learning for Multi-Label Classification, SSLMC [11]) and Multi-class learning methods (e.g. Shared Structures in Multi-Class Classification, SSMCC [12]).

Subspace learning based on measurement: This kind of subspace learning method aims to learn the great measurement of different modalities of media data so that it can achieve the measurement difference among the data. It can be divided into Euclidean distance measurement methods (e.g. Multi-Modal Distance Metric Learning, MMDML [13]) and Mahalanobis distance measurement methods (e.g. Shared Subspace for Multiple Metric Learning, SSMML [14]).

2.2 Sparse Representation

Researches of neurophysiology show that sparse coding exists in primary visual cortex of humans. In 2000, Vinje and Gallant published a paper in Science [15]. By recording the response characteristics of the macaque's neurons under conditions of open natural scenes and simulated natural scenes, they discovered that the response of neurons in visual cortex meets sparse distribution. Then in 2001, a paper published by Nirengerg et al. in Nature showed similar results [16].

Sparse model is widely applied to domains of signal and image processing. Each signal can be represented by a linear combination of small number of elements in a dictionary. The development of sparse representation on image is roughly as follows: In 1993, Mallat proposed sparse representation for overcomplete dictionary. He used an overcomplete Gabor dictionary to represent an image and proposed Matching Pursuit (MP) algorithm [17]. In 1996, Olshausen et al. revealed the directionality of Human Vision [18]. Furthermore, many other methods have also been proposed [19-22]. For example, in [19], a signal was sparsely coded over a set of redundant bases and classified based on its coding vector. In [20], Wright et al. introduce sparse representation to robust face recognition. This boosts the research of sparse representation classification. And Gao et al. [21] proposed kernel sparse representation in face recognition.

3 SPARSE REPRESENTATION-SEMANTIC MATCHING CROSS-MEDIA RETRIEVAL

In this section, the details of SRSM are introduced. The framework of the model is shown in Fig. 2. Two independent sparse representation classifiers will be used to map the heterogeneous features of images and texts into their common semantic space before implementing similarity comparison. And with the output of the two independent sparse representation classifiers, the common semantic space of images and texts can be obtained and then be applied to cross-media retrieval. Two independent sparse representation classifiers unify the isomorphic features of images and texts to the common semantic level respectively. And then, these models unify the heterogeneous media data to the semantic level.

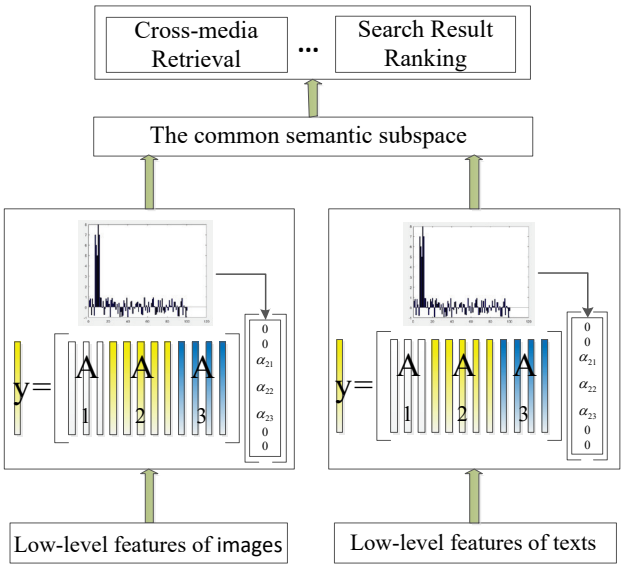

Figure 2 The framework of the proposed SRSM

3.1 Sparse Representation Classifier

With the developing of compressed sensing, sparse representation represents a sample (a test sample) e.g. an image or a text using an overcomplete dictionary (the training samples), and the representation is linear and naturally sparse [23-26]. The total training set is defined as the overcomplete dictionary A of k classes:

$$A = \left[A_1, A_2, ..., A_k\right] = \left[v_{1,1}, v_{1,2}, ..., v_{k,n_k}\right] \tag{1}$$

where the i^{th} class is represented as:

$$A_i = \left[v_{i,1}, v_{i,2}, ..., v_{i,n_i}\right] \in R^{m \times n_i} \tag{2}$$

where, m is the samples dimension (p for image and q for text), and n_i is the number of training samples of i^{th} class.

Then, for a test sample y, it can be represented as a linear representation of the total training samples as:

$$y = Ax \in R^m \tag{3}$$

where $x = \left[\hat{0},..., \hat{0}, \alpha_{i,1}, \alpha_{i,2},..., \hat{0},..., \hat{0}\right]^{\mathrm{T}} \in R^n$ is a coefficient matrix whose elements are close to zero except those related to i^{th} class (as shown in Fig. 3). So a test sample y is effectively represented only using the training samples of the same class.

Recent development in compressed sensing and sparse representation shows that the linear representation $y = Ax$ can be solved by the following l_2-minimization problem:

$$\hat{x}_2 = \arg\min \|x\|_2 \text{ s.t. } Ax = y \tag{4}$$

where $\|\cdot\|_2$ represents l_2-norm.

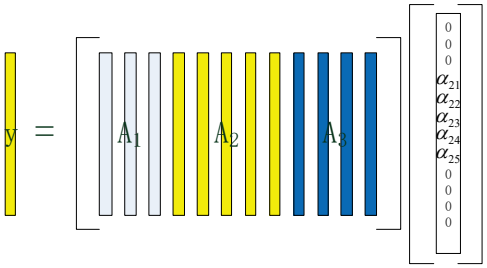

Figure 3 A test sample y is represented as a linear representation of the total training samples

However, the solution of l_2-minimization problem is not sparse enough with a large number of nonzero entries corresponding to the multiple classes. The testing sample should be produced effectively represented only using the training samples of the same class. The representation is sparse naturally. And this problem can be solved by the sparsest solution l_0-minimization problem:

$$\hat{x}_0 = \arg\min \|x\|_0 \text{ s.t. } Ax = y \tag{5}$$

where $\|\cdot\|_0$ represents l_0-norm. However, this problem is NP-hard. And the researches have validated that the solution of l_0-minimization problem is equal to the l_1-minimization problem if x is sparse enough, which is shown as follows:

$$\hat{x}_1 = \arg\min \|x\|_1 \text{ s.t. } Ax = y \tag{6}$$

Furthermore, in the real application, the data is noisy. The testing sample cannot be represented exactly as a sparse linear representation of the training samples. Therefore, the linear representation $y = Ax$ can be rewritten as follows with the possible noise:

$$y = Ax + z \tag{7}$$

where z is a noise term with boundary $\|z\|_2 < \varepsilon$. And the l_1-minimization problem can be changed to:

$$\hat{x}_1 = \arg\min \|x\|_1 \text{ s.t. } \| Ax - y \|_2 \le \varepsilon \tag{8}$$

Now when given a test sample y, its sparse representation $\hat{x}_1$ is computed firstly. And the testing sample y can be effectively represented only using the training samples of the same class. However, there may be a few small nonzero elements of multiple classes in $\hat{x}_1$ because of modeling error and noise. Consequently, for the ith class, let function $\delta_i : R^n \to R^n$ select coefficients related with it, and for $x \in R^n$, $\delta_i(x) \in R^n$ is a vector whose elements are zero except those related with ith class. So the linear representation can be approximately represented as $\dot{y}_i = A\delta_i(\hat{x}_1)$. At last, the test sample y can be classified based on the minimization of residual between $\dot{y}_i$ and y:

$$\min_i r_i(y) = \left\| y - A\delta_i(\hat{x}_1) \right\|_2 \tag{9}$$

And the algorithm of Sparse Representation Classifier is shown as Algorithm 1.

Algorithm 1: Sparse Representation-based Classification (SRC)

(1) Input: a matrix of training samples $A = [A_1, A_2,\ldots, A_k] \in R^{m\times n}$, a test sample $y \in R^m$, (and an optional error tolerance $\varepsilon > 0$)
(2) Normalize the columns of A to have unit l_2
(3) Solve the l_1-minimization optimization problem: $\hat{x}_1 = \arg\min \|x\|_1$ s.t. $Ax = y$
Or alternatively, $\hat{x}_1 = \arg\min \|x\|_1$ s.t. $\| Ax - y \|_2 \le \varepsilon$
(4) Computing the residuals $r_i(y) = \left\| y - A\delta_i(\hat{x}_1) \right\|_2$, $i = 1, 2,..., k$
(5) Output: $r = [r_1, r_2,\ldots, r_i,\ldots, r_k] \in R^{1\times k}$

3.2 Sparse Representation-Semantic Matching

In SRSM, two independent sparse representation classifiers are used to map the heterogeneous features of images and texts into their common semantic space before implementing similarity comparison. Firstly, all training images or texts are used to reconstruct each image or text based on algorithm 1. And then, after obtaining the residuals vectors of testing images and texts, a little change is made to them: transforming them to probability representations and setting the maximum value of each residuals vector to be 1 while other being 0. And the algorithm of SRSM is shown as Algorithm 2.

Algorithm 2: Sparse Representation-Semantic Matching (SRSM)

(1) Input: training samples of images $I = [I_1, I_2,\ldots, I_k] \in R^{p\times n}$ and texts $T = [T_1, T_2,\ldots, T_k] \in R^{q\times n}$, testing samples of images $Iy \in R^{p\times te}$ and texts $Ty \in R^{q\times te}$.
(2) FOR each testing image in Iy
Compute its residuals $Ir_j \in R^{1\times k}$ using algorithm 1.

$$Ir_j = \frac{1}{Ir_j}$$

FOR each Ir_{ji} in Ir_j

$$P_{ji} = \frac{\exp(Ir_{ji})}{\sum_{s=1}^{k} \exp(Ir_{js})}$$

$$Ir_{ji} = P_{ji}$$

END FOR

Set the maximum value of Ir_j to be 1.
END FOR
Obtain $Ir = [Ir_1, Ir_2, \ldots, Ir_j, \ldots, Ir_k]^{\mathrm{T}} \in R^{te \times k}$
(3) FOR each testing text in Ty
Compute its residuals $Tr_j \in R^{1 \times k}$ using Algorithm 1.

$$Tr_j = \frac{1}{Tr_j}$$

FOR each Tr_{ji} in Tr_j

$$P_{ji} = \frac{\exp(Tr_{ji})}{\sum_{s=1}^{k} \exp(Tr_{js})}$$

$$Tr_{ji} = P_{ji}$$

END FOR
Set the maximum value of Ty_j to be 1.
ENDFOR
Obtain $Tr = [Tr_1, Tr_2, \ldots, Tr_j, \ldots, Tr_k]^{\mathrm{T}} \in R^{te \times k}$
(4) Output: Ir and Tr

Now the low-level features of images and texts are mapped into their common semantic subspace in which the feature dimension of images and texts is the same, which is shown as follows:

$$M_I : I^{p \times n} \rightarrow Ir^{k \times n} \tag{10}$$

$$M_T : T^{q \times n} \rightarrow Tr^{k \times n} \tag{11}$$

4 EXPERIMENTS

In this section, the experimental results of SRSM are shown and compared with some other cross-media retrieval methods on four datasets: Wikipedia dataset [3], NUS-WIDE dataset [27], Wikipedia dataset with CNN features, and Pascal dataset with CNN features [28-29]. And the experimental results validate the effectiveness of this method.

4.1 Dataset

Wikipedia dataset [3]: It contains 2866 image-text pairs from Wikipedia's articles and the related images. All of them are classified to 10 categories. In this dataset, the low-level features of texts are 10-dimensional Latent Dirichlet Allocation (LDA) features [30] while images are 128-dimensional Scale Invariant Feature Transformation (SIFT) [31].

NUS-WIDE dataset [27]: It contains 26,9648 image-text pairs. There are 81 semantic categories in all. In the experiments, 10 categories with maximum number of samples are selected (i.e. sky, lake, grass, plants, window, water, animal, buildings, clouds and person) to construct the dataset. In this dataset, the low-level features of texts are 1000-dimensional tag feature vectors while images are 500-dimensional SIFT features [31].

Wikipedia-CNN dataset [28-29]: It extracts CNN features from original images and textual features from original texts respectively. The low-level features of images are represented as 4096-dimensional CNN features while texts are 100-dimensional LDA features.

Pascal-CNN dataset [28]: It contains 1000 image-text pairs. There are 20 categories totally. In the experiment, 600 pairs are selected for training and 400 for testing. In this dataset, the low-level features of images are represented as 4096-dimensional CNN features while texts are 100-dimensional LDA features.

4.2 Evaluation Metric and Distance Functions

In experiment, Mean Average Precision (MAP) and Precision-Recall (PR) [3, 28, 29] are used to evaluate the performance of this method and compared ones. MAP and PR are widely used in performance evaluation of cross-media retrieval algorithms. In this paper, the cross-media retrieval includes two tasks: query image retrieves relevant text and query text retrieves relevant images.

The last step of general cross-media retrieval methods is computing the distances of each image and text samples. The distance functions includes L_1 distance, Normalized Correlation (*NC*), L_2 distance, Kullback-Leibler Divergence (*KL*), and Centered Correlation (*CC*), which are shown as follows:

For convenience, $i^{k \times 1}$ are used to represent a sample of $Ir^{k \times n}$ and $t^{k \times 1}$ to represent a sample of $Tr^{k \times n}$.

L_1 distance:

$$dis_{L_1} = \sum_{m=1}^{k} | i_m - t_m | \tag{12}$$

L_2 distance:

$$dis_{L_2} = \sqrt{\sum_{m=1}^{k} (i_m - t_m)^2} \tag{13}$$

KL distance:

$$dis_{KL} = \sum_{m=1}^{k} i_m \log \frac{i_m}{t_m} \tag{14}$$

NC distance:

$$dis_{NC} = \frac{\sum_{m=1}^{k} i_m \times t_m}{\sqrt{\sum_{m=1}^{k} (i_m)^2} \sqrt{\sum_{m=1}^{k} (t_m)^2}} \tag{15}$$

CC distance:

$$dis_{CC} = -\frac{(i - \bar{i})^{\mathrm{T}} (t - \bar{t})}{k} \tag{16}$$

Table 1 MAP obtained by different distance metrics used SRSM on Wiki-CNN dataset

Methods	Distance Metric	Image Query	Text Query	Average
SRSM	L_1	0.355	0.372	0.364
	L_2	0.367	0.384	0.376
	KL	0.370	0.382	0.376
	NC	0.425	0.385	0.405
	CC	**0.467**	**0.395**	**0.431**

Then the performance of SRSM is evaluated using the 5 different distance functions on Wikipedia-CNN dataset in order to find which one is the most suitable. The experimental results are shown in Tab. 1. It can be found that CC distance obtains the best performance. So CC distance is used in the whole experiments.

4.3 Experimental Results

In these experiments, the performance of SRSM is compared with some other cross-media retrieval methods on Wikipedia dataset which is designed for cross-media retrieval and NUS-WIDE dataset which is much larger than Wikipedia dataset. The MAP scores on both datasets are shown in Tab. 2. For the MAP scores obtained by compared methods, the results in [30] are cited. The results verify the effectiveness of the proposed SRSM method.

Table 2 MAP scores on Wikipedia dataset and NUS-WIDE dataset

Methods	Wikipedia			NUS-WIDE		
	I2T	T2I	Avg	I2T	T2I	Avg
PLS	0.2402	0.1633	0.2032	0.2752	0.2661	0.2706
BLM	0.2562	0.2023	0.2293	0.2976	0.2809	0.2892
CCA	0.2409	0.1906	0.2157	0.2298	0.2298	0.2298
LSCMR	0.2021	0.2229	0.2125	0.1424	0.2491	0.1958
Bi-LSCMR	0.2123	0.2528	0.2326	0.1453	0.238	0.1917
CDFE	0.2655	0.2059	0.2357	0.2595	0.2869	0.2732
GMMFA	0.275	0.2139	0.2445	0.2983	0.2939	0.2961
GMLDA	0.2751	0.2098	0.2425	0.3243	0.3076	0.3159
CCA-3V	0.2752	0.2242	0.2497	0.3513	0.326	0.3386
SliM^2	0.2548	0.2021	0.2285	0.3154	0.2924	0.3039
M^{3}R	0.2298	**0.2677**	0.2488	0.2445	0.3044	0.2742
LCFS	0.2798	0.2141	0.247	0.383	0.346	0.3645
JFSSL	0.3036	0.2275	0.2669	0.4035	0.3747	0.3891
SRSM	**0.3374**	0.2409	**0.2892**	**0.4156**	**0.3775**	**0.3966**

Recently, it has been proved in many domains that CNN features enjoy more powerful performance for image representation. Consequently, the methods of cross-media retrieval with CNN features have a better performance. Therefore, the SRSM method is also compared with compared methods on Wikipedia-CNN dataset and Pascal-CNN dataset which extract CNN visual features from original images. For compared methods, some classical algorithms are selected, which include MDCR, GMMFA [31], GMMLDA [32], CCA-3V [33], SCM, and CCA. The MAP scores are shown on these two datasets in Tab. 3.

Table 3 MAP on Wikipedia-CNN dataset and Pascal-CNN dataset

Methods	Wikipedia-CNN			Pascal-CNN		
	I2T	T2I	Avg	I2T	T2I	Avg
CCA	0.226	0.246	0.236	0.261	0.356	0.309
SCM	0.351	0.324	0.337	0.369	0.375	0.372
CCA-3V	0.31	0.316	0.313	0.337	0.439	0.388
GMLDA	0.372	0.322	0.347	0.456	0.448	0.452
GMMFA	0.371	0.322	0.346	0.455	0.447	0.451
MDCR	0.435	0.394	0.415	0.455	**0.471**	0.463
SRSM	**0.467**	**0.395**	**0.431**	**0.478**	0.455	**0.467**

And the PR curves for image query and text query on Wikipedia-CNN dataset and Pascal-CNN dataset. Furthermore, the MAP scores per class for image query, text query and the average performance on Wikipedia-CNN dataset and Pascal-CNN dataset are also shown. The results verify the superior performance of SRSM compared with other methods on the two datasets with CNN features.

And lastly, the left column is the query image (text), and the top 5 results are listed on the right columns. The upper part is using image to retrieve texts. The lower part is using text to retrieve images. For both parts, the first row is the success case while the second row is the failure case. For convenience, the images corresponding to the texts are used to represent both the query texts and the retrieved texts.

5 CONCLUSION

In this paper, two independent sparse representation classifiers are used to map the heterogeneous features of images and texts into their common semantic space before implementing similarity comparison. And with the output of the two independent sparse representation classifiers, the common semantic space of images and texts is obtained and further applied into cross-media retrieval. This method is named as Sparse Representation-Semantic Matching (SRSM) in this paper. The cross-media retrieval in this paper includes two tasks: query image retrieves relevant text and query text retrieves relevant images. In the experiments, the semantic information is made full of use. Through the analysis of the results, this method is effective obviously. The performance of this method on Wiki dataset, NUS-WIDE dataset, Wiki dataset with CNN features and Pascal dataset with CNN features is shown. The experimental results validate its effectiveness compared with several state-of-the-art algorithms. With this method, the images or texts can be retrieved effectively.

Acknowledgements

This work was supported in part by the Key Research and Development Foundation of Shandong Province (2016GGX101035, 2019GGX101004), the Shandong Education Department Teaching Reform Project (Z2016M014, Z2016M016, Z2016Z013, M2018x197) and China Scholarship Council under Grant CSC (201809995006).

6 REFERENCES

[1] Xu, G. W., Sang, Z. Q., & Zhang, Z. J. (2018). Cross-media retrieval based on pseudo-label learning and semantic consistency algorithm. *International Journal of Performability Engineering*, 14(9), 2219-2229. https://doi.org/10.23940/ijpe.18.09.p31.22192229

[2] Peng, Y. X., Zhu, W. W., Zhao, Y., Xu, C. S., Huang, Q. M., Lu, H. Q., Zheng, Q. H., Huang, T. J., & Gao, W. (2017). Cross-media analysis and reasoning: advances and directions. *Frontiers of Information Technology & Electronic Engineering*, 18(1), 44-57. https://doi.org/10.1631/FITEE.1601787

[3] Rasiwasia, N., Pereira, J. C.,Coviello, E.,Doyle, G., Lanckriet, G., Levy, R., & Vasconcelos, N. (2010). A new approach to cross-modal multimedia retrieval. *International Conference on Multimedia,* ACM, Firenze, Italy, 251-260. https://doi.org/10.1145/1873951.1873987

[4] Andersen, A. H., Rayens, W. S., Liu, Y. S., & Smith, C. D. (2012).Partial least squares for discrimination. *Journal of Chemometrics*, 30(3), 446-452. https://doi.org/10.1016/j.mri.2011.11.001

[5] Hardoon, D. R., Szedmak, S. R., & Shawe-Taylor, J. R. (2004). Canonical correlation analysis: An overview with

application to learning methods. *Neural Computation*, 16(12), 2639-2664. https://doi.org/10.1162/0899766042321814
[6] Andrew, G., Arora, R., Bilmes, J., & Livescu, K. (2013). Deep canonical correlation analysis. *International Conference on Machine Learning*, Hong Kong, 1247-1255.
[7] Gupta, S. K., Phung, D., Adams, B., Tran, T., & Venkatesh, S. (2010). Nonnegative shared subspace learning and its application to social media retrieval. *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Washington, Dc, USA, 1169-1178. https://doi.org/10.1145/1835804.1835951
[8] Yu, S. P., Yu, K., Tresp, V., & Kriegel, H. P. (2007).Multi-output regularized feature projection. *IEEE Transactions on Knowledge & Data Engineering*, 18(12), 1600-1613. https://doi.org/10.1109/TKDE.2006.194
[9] Ando, R. K. & Zhang, T. (2005).A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research, 6*(3), 1817-1853.
[10] Argyriou, A., Evgeniou, T., & Pontil, M. (2008). Convex multi-task feature learning. *Machine Learning*, 73(3), 243-272. https://doi.org/10.1007/s10994-007-5040-8
[11] Kong, X. N., Ng, M. K., & Zhou, Z. H. (2011). Transductive Multi-Label Learning via Label Set Propagation. *IEEE Transactions on Knowledge & Data Engineering, 99*(3), 704-719. https://doi.org/10.1109/TKDE.2011.141
[12] Amit, Y., Fink, M., Srebro, N., & Ullman, S. (2007). Uncovering shared structures in multiclass classification. *Machine Learning, Twenty-Fourth International Conference*, Corvalis, Oregon, USA, 17-24. https://doi.org/10.1145/1273496.1273499
[13] Xie, P. & Xing, E. P. (2013). Multi-modal distance metric learning. *International Joint Conference on Artificial Intelligence*, Beijing, China, 1806-1812.
[14] Yang, P. P., Huang, K., & Liu, C. L. (2013). A multi-task framework for metric learning with common subspace. *Neural Computing & Applications, 22*(7-8), 1337-1347. https://doi.org/10.1007/s00521-012-0956-8
[15] Vinje, W. E. & Gallant, J. L. (2000). Sparse coding and decorrelation in primary visual cortex during natural vision. *Science*, 287(5456), 1273-1276. https://doi.org/10.1126/science.287.5456.1273
[16] Nirenberg, S., Carcieri, S. M., Jacobs, A. L., & Latham, P. E. (2001). Retinal ganglion cells act largely as independent encoders. *Nature, 411*(6838), 698-701. https://doi.org/10.1038/35079612
[17] Mallat, S. G. & Zhang, Z. (1993). Matching pursuits with time-frequency dictionaries. *IEEE Transactions on signal processing, 41*(12), 3397-3415. https://doi.org/10.1109/78.258082
[18] Olshausen, B. A. & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature, 381*(6583), 607-609. https://doi.org/10.1038/381607a0
[19] Schölkopf, B., Platt, J., & Hofmann, T. (2006). Sparse Representation for Signal Classification. *Advances in Neural Information Processing Systems 19, Proceedings of the Twentieth Conference on Neural Information Processing Systems*, Vancouver, British Columbia, Canada, 609-616. https://doi.org/10.7551/mitpress/7503.001.0001
[20] Wright, J., Yang, A. Y., Ganesh, A., Sastry, S. S., & Ma, Y. (2009). Robust Face Recognition via Sparse Representation. *IEEE Transactions on Pattern Analysis & Machine Intelligence, 31*(2), 210-227. https://doi.org/10.1109/TPAMI.2008.79
[21] Gao, S., Tsang, W. H., & Chia, L. T. (2010). Kernel sparse representation for image classification and face recognition. *The 11th European Conference on Computer Vision*, Heraklion, Crete, Greece, 1-14. https://doi.org/10.1007/978-3-642-15561-1_1
[22] Zhang, Z., Xu, Y., Yang, J., et al, (2015). A survey of sparse representation: Algorithms and applications. *IEEE Access,* (3), 490-530. https://doi.org/10.1109/ACCESS.2015.2430359
[23] Xu, G. W., Xu, L. N., Zhang, M. J., & Li, X. M. (2018). Two-stage semantic matching for cross-media retrieval. *International Journal of Performability Engineering, 14*(4), 795-804. https://doi.org/10.23940/ijpe.18.04.p21.795804
[24] Yang, X., Wu, W., Liu, K., Chen, W. L., & Zhou, Z. L. (2018). Multiple dictionary pairs learning and sparse representation-based infrared image super-resolution with improved fuzzy clustering. *Soft Computing, 22*(5), 1385-1398. https://doi.org/10.1007/s00500-017-2812-3
[25] Li, C., Ma, Y., Mei, X., Liu, C., & Ma, J. (2016). Hyperspectral image classification with robust sparse representation. *IEEE Geoscience & Remote Sensing Letters, 13*(5), 641-645. https://doi.org/10.1109/LGRS.2016.2532380
[26] Chan, C. H. & Kittler, J. (2010). Sparse representation of (Multiscale) histograms for face recognition robust to registration and illumination problems. *IEEE International Conference on Image Processing*, Hong Kong, 2441-2444. https://doi.org/10.1109/ICIP.2010.5651933
[27] Chua, T. S., Tang, J., Hong, R., Li, H., Luo, Z. P., Zheng, Y. T. (2009). NUS-WIDE: A real-world web image database from National University of Singapore. *ACM International Conference on Image and Video Retrieval*, Santorini Island, Greece, 48-56. https://doi.org/10.1145/1646396.1646452
[28] Wei, Y. C., Zhao, Y., Zhu, Z., Wei, S. K., Xiao, Y. H., Feng, J., & Yan, S. C. (2016). Modality-dependent cross-media retrieval. *ACM Transactions on Intelligent Systems & Technology, 7*(4), 1-13. https://doi.org/10.1145/2775109
[29] Wang, K. Y., Yin, Q. Y., Wang, W., Wu, S., & Wang, L. (2016). A comprehensive survey on cross-modal retrieval. *arXiv*: 1607.06215.
[30] Blei, D. M., Ng, A. Y., & Jordan. M. I. (2012). Latent Dirichlet allocation. *Journal of Machine Learning Research*, (3), 993-1022.
[31] Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision, 60*(2), 91-110. https://doi.org/10.1023/B:VISI.0000029664.99615.94
[32] Sharma, A., Kumar, A., Daume, H., & Jacobs, D. W. (2012). Generalized multiview analysis: A discriminative latent space. *IEEE Conference on Computer Vision and Pattern Recognition*, Rhode Island, USA, 2160-2167. https://doi.org/10.1109/CVPR.2012.6247923
[33] Gong, Y. C., Ke, Q. F., Isard, M., & Lazebnik, S. (2014). A multi-view embedding space for modeling internet images, tags, and their semantics. *International Journal of Computer Vision, 106*(2), 210-233. https://doi.org/10.1007/s11263-013-0658-4

Contact information:

Gongwen XU
(Corresponding author)
Business School,
Shandong Jianzhu University,
Jinan 250101, China
E-mail: xugongwen@163.com

Aidong ZHAI
Maternity and Child Health Care Hospital,
Zibo 255029, China
E-mail: 172270580@qq.com

Jing WANG
Business School,
Shandong Jianzhu University,
Jinan 250101, China
E-mail: wangjing@sdjzu.edu.cn

Zhijun ZHANG
Computer Science and Technology School,
Shandong Jianzhu University,
Jinan 250101, China
E-mail: zhangzj@sdjzu.edu.cn

Xiaomei LI
The Second Hospital,
Shandong University,
Jinan 250013, China
E-mail: sdulixiaomei@163.com