Multi-modal Subspace Learning with Joint Graph Regularization for Cross-modal Retrieval

Kaiye Wang, Wei Wang, Ran He, Liang Wang, Tieniu Tan
National Lab of Pattern Recognition, Center for Research on Intelligent Perception and Computing
Institute of Automation, Chinese Academy of Sciences
Beijing, China
{*kaiye.wang, wangwei, rhe, wangliang, tnt*}*@nlpr.ia.ac.cn*

***Abstract*—This paper investigates the problem of cross-modal retrieval, where users can search results across various modalities by submitting any modality of query. Since the query and its retrieved results can be of different modalities, how to measure the content similarity between different modalities of data remains a challenge. To address this problem, we propose a joint graph regularized multi-modal subspace learning (JGRMSL) algorithm, which integrates inter-modality similarities and intra-modality similarities into a joint graph regularization to better explore the cross-modal correlation and the local manifold structure in each modality of data. To obtain good class separation, the idea of Linear Discriminant Analysis (LDA) is incorporated into the proposed method by maximizing the between-class covariance of all projected data and minimizing the within-class covariance of all projected data. Experimental results on two public cross-modal datasets demonstrate the effectiveness of our algorithm.**

***Keywords*-cross-modal retrieval; subspace learning; joint graph regularization;**

I. Introduction

Over the last decade, multimedia content such as text, image and video has been increasing rapidly on the Internet. Accordingly, there is an increasing need for efficiently and effectively searching such multi-modal data. Furthermore, users may demand the cross-modal retrieval to search results across various modalities by submitting any modality of query. Since the search results of the cross-modal retrieval are rich in multiple modalities, they are more comprehensive than the results of traditional retrieval methods. And it is very convenient for users to take any modality of content at hand as a query. The main difficulty of the cross-modal retrieval is how to measure the content similarity between different modalities of data. In this paper, we aim to learn a discriminative common subspace in which the similarity between heterogeneous data can be measured.

Several recent approaches for establishing relationships between data from different modalities generally rely on Canonical Correlation Analysis (CCA) [1]. Hardoon et al. [1] and Rasiwasia et al. [2] apply CCA to project text and images to a common latent subspace for the cross-modal multimedia retrieval. Hwang and Grauman [3] use kernel CCA to learn the connections between human-provided tags and visual features, accounting for the relative importance of words. CCA is also used in other domains, such as cross-lingual retrieval [4] and half-face verification [5]. Two other popular approaches for learning a latent subspace are Partial Least Squares (PLS) [6] and Bilinear Model (BLM) [7]. Recently, Sharma and Jacobs [6] use PLS to linearly map images in different modalities to a common subspace for multi-modal face recognition. Then, Sharma et al. [8] apply PLS to cross-media retrieval. Chen et al. [9] use PLS to switch data from one modality to another modal space for cross-modal document retrieval. In [7], Tannenbaum and Freeman [7] propose a bilinear model (BLM) to derive a common space for cross-modal face recognition, and Sharma et al. [8] apply BLM to text-image retrieval tasks.

However, as we know, CCA, PLS and BLM only use pairwise information. They do not make use of the structure information of different spaces and label information. Recently, Sharma et al. [8] propose a supervised multi-view feature extraction approach to extend LDA and Marginal Fisher Analysis (MFA) to their multiview counterpart, i.e., Generalized Multiview LDA (GMLDA) and Generalized Multiview MFA (GMMFA). In GMLDA and GMMFA, the discriminability is obtained within each view and the cross-view correlation is obtained only from pairwise information.

Motivated by [8], we propose a joint graph regularized multi-modal subspace learning (JGRMSL) algorithm for the cross-modal retrieval. The proposed approach integrates inter-modality similarities and intra-modality similarities into a joint graph regularization to better explore the cross-modality correlation among all of data from different modalities and the local manifold structure information. To learn a discriminative subspace, we adopt the idea of LDA. In the learnt space, the between-class covariance of all projected data is maximized, meanwhile the within-class covariance of all projected data is minimized. Here, it should be noted that the discriminability is obtained across all modalities of data, which is very different from that of GMLDA and GMMFA. In our implementation, the joint graph regularizer, the between-class covariance and the within-class covariance are elegantly combined to form a unified formulation, so they can be optimized simultaneously. Finally, we solve the

978-1-4799-2190-4/13 $31.00 © 2013 IEEE
DOI 10.1109/ACPR.2013.44

unified formulation to obtain the project vectors by using the generalized eigenvalue decomposition. Experimental results on two public cross-modal datasets show the promise of the proposed approach.

The rest of this paper is organized as follows. In Section II, we introduce our joint graph regularized multi-modal subspace learning (JGRMSL) algorithm for cross-modal retrieval. The experimental results on two public datasets are presented in Section III. Finally, we conclude the paper in Section IV.

II. Joint Graph Regularized Multi-modal Subspace Learning

In this section, we present a novel subspace learning algorithm for cross-modal retrieval. Without loss of generality, we introduce our method in the two-modality case. First, we formulate the inter-modality similarities and intra-modality similarities to a joint graph regularization term. Second, we integrate the joint graph regularizer, the between-class covariance and the within-class covariance of all projected data into a unified formulation, which can be solved efficiently as a generalized eigenvalue problem. Finally, we extend the proposed algorithm to the multi-modal case (i.e., more than two modalities).

A. The Joint Graph Regularization Term

Suppose that we have a collection of data from two different modalities, i.e., $X_1 = [\mathbf{x}_1^{(1)}, \mathbf{x}_2^{(1)}, ..., \mathbf{x}_n^{(1)}] \in \mathbb{R}^{d_1 \times n}$ and $X_2 = [\mathbf{x}_1^{(2)}, \mathbf{x}_2^{(2)}, ..., \mathbf{x}_n^{(2)}] \in \mathbb{R}^{d_2 \times n}$, where n is the number of the samples, d_1 and d_2 are the dimensions of the two modalities of data, respectively.

We aim to learn two projections to map data from different modalities into a common latent subspace, where the similar samples should be as close as possible. And the local manifold structure should be preserved in the common latent subspace, which can prevent overfitting and make the solution smoother. To learn such two projections, we minimize the following function:

$$\begin{aligned} &J(\mathbf{u}_1, \mathbf{u}_2) = \\ &\sum_{i,j=1}^{n} z_{ij}(\mathbf{u}_1^T \mathbf{x}_i^{(1)} - \mathbf{u}_2^T \mathbf{x}_j^{(2)})^2 + \frac{\lambda_1}{2} \sum_{i,j=1}^{n} s_{ij}^{(1)}(\mathbf{u}_1^T \mathbf{x}_i^{(1)} - \mathbf{u}_1^T \mathbf{x}_j^{(1)})^2 \\ &+ \frac{\lambda_2}{2} \sum_{i,j=1}^{n} s_{ij}^{(2)}(\mathbf{u}_2^T \mathbf{x}_i^{(2)} - \mathbf{u}_2^T \mathbf{x}_j^{(2)})^2 \end{aligned} \tag{1}$$

where $\mathbf{u}_1$ and $\mathbf{u}_2$ are the two projections, λ_1 and λ_2 are trade-off parameters, z_{ij} and $s_{ij}^{(v)}(v = 1, 2)$ are defined as follows respectively:

$$z_{ij} = \begin{cases} 1 & if\ \mathbf{x}_i^{(1)}\ is\ similar\ to\ \mathbf{x}_j^{(2)} \\ 0 & otherwise \end{cases} \tag{2}$$

$$s_{ij}^{(v)} = \begin{cases} 1 & if\ \mathbf{x}_i^{(v)} \in N_k(\mathbf{x}_j^{(v)})\ or\ \mathbf{x}_j^{(v)} \in N_k(\mathbf{x}_i^{(v)}) \\ 0 & otherwise \end{cases} \tag{3}$$

where $\mathbf{x}_i^{(1)}$ is similar to $\mathbf{x}_j^{(2)}$ if they belong to the same class, $N_k(\mathbf{x}_i^{(v)})$ denotes the set of k nearest neighbors of $\mathbf{x}_i^{(v)}$. The first term of the above function utilizes the inter-modality similarities to enforce the similar samples as close as possible after the projection. The second and third terms try to preserve the local manifold structure of each modality of data in the common latent space.

For simplicity, we rewrite the above function as a joint graph embedding formulation. Let X be a $(d_1 + d_2) \times 2n$ matrix representing the multi-modality data, $\mathbf{u}$ be a projection vector of length $d_1 + d_2$ and W be a $2n \times 2n$ matrix in the following form:

$$X = \begin{bmatrix} X_1 & \mathbf{0} \\ \mathbf{0} & X_2 \end{bmatrix}; \mathbf{u} = \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{bmatrix}; W = \begin{bmatrix} \lambda_1 S_1 & Z \\ Z^T & \lambda_2 S_2 \end{bmatrix} \tag{4}$$

Then, Eq. (1) can be reformulated as:

$$\begin{aligned} J(\mathbf{u}) &= \frac{1}{2} \sum_{i,j=1}^{2n} W_{ij}(\mathbf{u}^T X_{(i)} - \mathbf{u}^T X_{(j)})^2 \\ &= \frac{1}{2} \mathbf{u}^T X(D - W) X^T \mathbf{u} \\ &= \frac{1}{2} \mathbf{u}^T X L X^T \mathbf{u} \end{aligned} \tag{5}$$

where D is a diagonal matrix, its entries are column sum of W, $D_{ii} = \sum_j W_{ij}$. $L = D - W$ is the Laplacian matrix. S_1 and S_2 indicate the intra-modality similarity, and Z indicates the inter-modality similarity. If we set S_1 and S_2 as zero matrices (i.e., ignore the intra-modality similarity), and set Z as an identity matrix (i.e., only use pairwise information), Eq. (5) is equivalent to CCA [10].

We integrate inter-modality similarities and intra-modality similarities into the joint graph formulation, which better explores the cross-modality correlation and the local manifold structure information. We further would like different-class samples to be mapped far apart while the same-class samples lie as close as possible. To learn such discriminative common space, the idea of LDA is incorporated into our algorithm, which will be detailed in the next subsection.

B. The Objective Function

To learn a discriminative common space, the between-class covariance of all projected data across both modalities is maximized, meanwhile the within-class covariance is minimized. The objective function takes the form:

$$\begin{aligned} &\arg\max_{\mathbf{u}} \frac{S_B}{S_W + \alpha J(\mathbf{u})} \\ \Rightarrow &\arg\max_{\mathbf{u}} \frac{S_B}{S_W + \alpha \mathbf{u}^T X L X^T \mathbf{u}} \end{aligned} \tag{6}$$

where α is a trade-off parameter, S_B is the between-class covariance and S_W is the within-class covariance. Let $X_v = \left\{ \mathbf{x}_{ik}^{(v)} | v = 1, 2; i = 1, ..., c; k = 1, ...n_i^{(v)} \right\}$ be the samples from the vth modality, where $\mathbf{x}_{ik}^{(v)}$ is the

kth sample of the ith class from the vth modality. $Y_v = \left\{ \mathbf{y}_{ik}^{(v)} = \mathbf{u}_v^T \mathbf{x}_{ik}^{(v)} | v = 1, 2; i = 1, ..., c; k = 1, ..., n_i^{(v)} \right\}$ denotes the projected data in the common latent space. S_W is given by

$$S_W = \sum_{i=1}^{c} \sum_{v=1}^{2} \sum_{k=1}^{n_i^{(v)}} (\mathbf{y}_{ik}^{(v)} - \mu_i)(\mathbf{y}_{ik}^{(v)} - \mu_i)^T \tag{7}$$

where $\mu_i = \frac{1}{n_i} \sum_{v=1}^{2} \sum_{k=1}^{n_i^{(v)}} \mathbf{y}_{ik}^{(v)}$ is the mean of the projected data across both modalities from the ith class and n_i is the number of samples in the ith class. The within-class covariance can be reformulated in the following form [11]:

$$S_W = [\mathbf{u}_1^T \ \mathbf{u}_2^T] \begin{pmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{pmatrix} \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{bmatrix} = \mathbf{u}^T R \mathbf{u} \tag{8}$$

where $R_{vv'}$ is defined as follows:

$$R_{vv'} = \begin{cases} \sum_{i=1}^{c} \left(\sum_{k=1}^{n_i^{(v)}} \mathbf{x}_{ik}^{(v)} \mathbf{x}_{ik}^{(v)T} - \frac{n_i^{(v)} n_i^{(v)}}{n_i} \mathbf{m}_i^{(v)} \mathbf{m}_i^{(v)^T} \right), v = v' \\ -\sum_{i=1}^{c} \frac{n_i^{(v)} n_i^{(v')}}{n_i} \mathbf{m}_i^{(v)} \mathbf{m}_i^{(v')^T}, \quad otherwise \end{cases} \tag{9}$$

where $\mathbf{m}_i^{(v)} = \frac{1}{n_i^{(v)}} \sum_{k=1}^{n_i^{(v)}} \mathbf{x}_{ik}^{(v)}$ is the mean of the samples from the ith class of the vth modality. And S_B is given by

$$S_B = \sum_{i=1}^{c} n_i (\mu_i - \mu)(\mu_i - \mu)^T \tag{10}$$

where μ is the mean of all projected data across both modalities. Similarly, the between-class covariance can be reformulated in the following form [11]:

$$S_B = [\mathbf{u}_1^T \ \mathbf{u}_2^T] \begin{pmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{pmatrix} \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{bmatrix} = \mathbf{u}^T Q \mathbf{u} \tag{11}$$

where $Q_{vv'}$ is defined as follows:

$$\begin{aligned} Q_{vv'} = & \left(\sum_{i=1}^{c} \frac{n_i^{(v)} n_i^{(v')}}{n_i} \mathbf{m}_i^{(v)} \mathbf{m}_i^{(v')^T} \right) \\ & - \frac{1}{n} \left(\sum_{i=1}^{c} n_i^{(v)} \mathbf{m}_i^{(v)} \right) \left(\sum_{i=1}^{c} n_i^{(v')} \mathbf{m}_i^{(v')} \right)^T \end{aligned} \tag{12}$$

Substituting (8) and (11) into (6), the objective function can be rewritten as

$$\begin{aligned} & \arg\max_{\mathbf{u}} \frac{S_B}{S_W + \alpha \mathbf{u}^T X L X^T \mathbf{u}} \\ \Rightarrow & \arg\max_{\mathbf{u}} \frac{\mathbf{u}^T Q \mathbf{u}}{\mathbf{u}^T R \mathbf{u} + \alpha \mathbf{u}^T X L X^T \mathbf{u}} \\ \Rightarrow & \arg\max_{\mathbf{u}} \frac{\mathbf{u}^T Q \mathbf{u}}{\mathbf{u}^T (R + \alpha X L X^T) \mathbf{u}} \end{aligned} \tag{13}$$

The projection vector $\mathbf{u}$ that maximizes the above objective function is given by the maximum eigenvalue solution to the following generalized eigenvalue problem:

$$Q\mathbf{u} = \lambda (R + \alpha X L X^T) \mathbf{u} \tag{14}$$

The proposed method takes the inter-modality similarities and the intra-modality similarities into consideration through the joint graph regularization term. Furthermore, it also obtains good class separation by maximizing the between-class covariance of all projected data and minimizing the within-class covariance of all projected data. Using the learnt projections, we can project data from different modalities into a common latent space, in which the content similarity between different modal data can be measured. The proposed method can be easily extended to the case of more than two modalities.

III. Experimental Results

We conduct a series of experiments in the two-modality case here, due to the lack of public datasets containing more than two modalities of data in the recent literature. We test the proposed JGRMSL method on two publicly available datasets - Pascal VOC 2007 [12] and Wiki image-text dataset [2]. For the cross-modal retrieval problem, we learn two projections on the training set using the proposed method. Then, we project the data from all modalities into the learnt common subspace, in which we can measure the content similarity between different modalities of data using normal distance functions. For the test set, we take data from one modality as the query set, and data from another modality as the database set.

A. Experimental Settings

We compare the proposed JGRMSL approach with PLS [6], BLM [8], CCA [2], GMMFA and GMLDA [8] in terms of common cross-modal retrieval tasks: (1) Image query vs. Text database, (2) Text query vs. Image database. Specifically, we use an image as a query to retrieve relevant text (or tags) from the text database and use a text (or tags) as a query to retrieve relevant images from the image database. And the cosine distance is used to measure the similarity of features.

We evaluate the overall performance of the algorithms with the mean average precision (MAP) [2]. To compute MAP, we first evaluate the average precision (AP) of a set of N retrieved documents by $\mathrm{AP} = \frac{1}{L} \sum_{r=1}^{N} P(r) \cdot rel(r)$, where L is the number of relevant documents in the retrieved set, $P(r)$ denotes the precision of the top r retrieved documents, and $rel(r) = 1$ if the rth retrieved document is relevant (where 'relevant' means belonging to the class of the query) and $rel(r) = 0$ otherwise. The MAP is then computed by averaging the AP values over all queries in the query set. The larger the MAP, the better the performance.

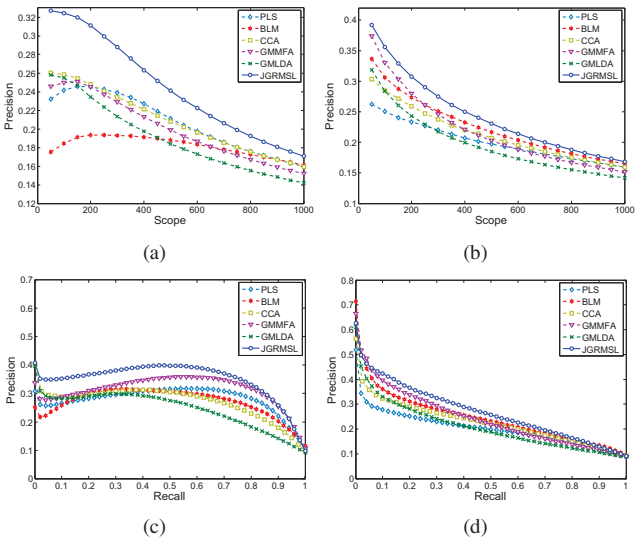

Figure 1. Performance of different methods on the Pascal VOC dataset, based on precision-scope curve (top row) for $K = 50$ to 1000 and precision-recall curve (bottom row). Left column: Image query vs. Text database. Right column: Text query vs. Image database.

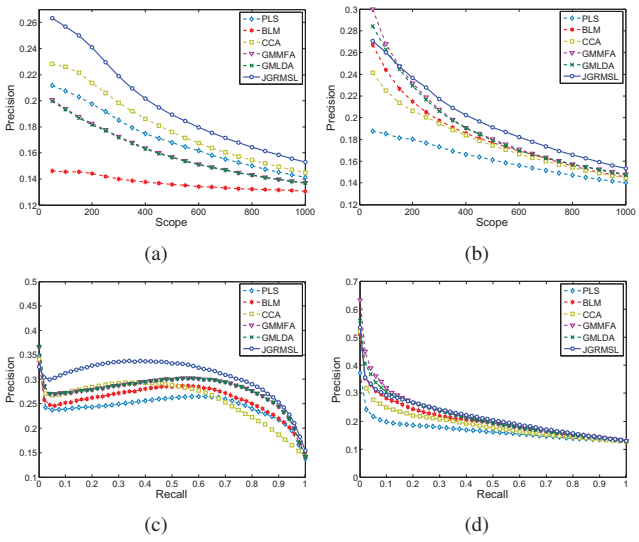

Figure 2. Performance of different methods on the Wiki dataset, based on precision-scope curve (top row) for $K = 50$ to 1000 and precision-recall curve (bottom row). Left column: Image query vs. Text database. Right column: Text query vs. Image database.

In addition, the precision-scope curve [13] and precision-recall curve [2] are also used to evaluate the effectiveness of different approaches. The precision-recall curve is a classical measure of information retrieval performance, but some researchers [13] consider the precision-scope curve more expressive for multimedia retrieval. Here, we report results with both of the two measures.

B. Results on Pascal VOC Dataset

The Pascal VOC 2007 dataset [12] contains a total of 9963 image-tag pairs, which can be categorized into 20 different classes. The dataset is split into a training set of 5011 image-tag pairs and a test set of 4952 image-tag pairs. Some images are multi-labeled, so we select images with only one object, which results in 2808 training and 2841 testing data. Each image is represented by a 512-dimensional Gist feature [12], and each text is represented by a 399-dimensional word frequency feature [12]. And Principal Component Analysis (PCA) is used to reduce the dimensions of the original features here.

Table I shows the MAP scores achieved by PLS, BLM, CCA, GMMFA, GMLDA and the proposed method (JGRMSL) on the Pascal VOC 2007 dataset. It can be observed that the proposed method outperforms its several counterparts for both forms of cross-modal retrieval tasks. This may be because the proposed method better explores the inter-modality similarities among all of data from different modalities through the joint graph regularization, meanwhile the local manifold structure is preserved to make the solution smoother. Furthermore, the proposed method obtains the discriminability across all modalities of data, which is of benefit to improve the performance.

Methods	Image query	Text query	Average
PLS	0.275	0.199	0.237
BLM	0.266	0.240	0.253
CCA	0.265	0.221	0.243
GMMFA	0.309	0.230	0.269
GMLDA	0.242	0.204	0.223
JGRMSL	**0.346**	**0.265**	**0.305**

Table I
MAP COMPARISON ON THE PASCAL VOC DATASET.

Further analysis of the results is presented in Figure 1, which shows the corresponding precision-scope curves and precision-recall curves of all approaches. The scope (i.e., the number of top retrieved items) for the precision-scope curves varies from K=50 to 1000. The top row shows the precision-scope curves of our method and its several counterparts for both forms of cross-modal retrieval tasks, i.e., Image query vs. Text database (left) and Text query vs. Image database (right). We observe that compared with its several counterparts, our method obtains better results for both tasks. The bottom row shows the performance of all methods based on the precision-recall curves, and our method again outperforms other algorithms for both forms of cross-modal retrieval.

C. Results on Wiki Dataset

The Wiki image-text dataset [2], generated from Wikipedia's "featured article", consists of 2866 image-text pairs. In each pair, the image is related to a complete text article, not just a few keywords. Each pair is annotated with a

Methods	Image query	Text query	Average
PLS	0.240	0.163	0.202
BLM	0.256	0.202	0.229
CCA	0.254	0.184	0.219
GMMFA	0.276	**0.213**	0.245
GMLDA	0.275	0.210	0.243
JGRMSL	**0.304**	0.211	**0.258**

Table II
MAP COMPARISON ON THE WIKI DATASET.

label from the vocabulary of 10 semantic classes. The dataset is split into a training set of 1300 pairs (130 pairs per class) and a testing set of 1566 pairs. The representation of the text with 10 dimensions is derived from a latent Dirichlet allocation model [14]. And each image is represented by a 128-dimensional SIFT [15] descriptor.

Table II shows the MAP scores achieved by all approaches on the Wiki dataset. We can observe that for the text query, the proposed method performs comparably to GMMFA and GMLDA, but better than the other methods. However, for the image query, the proposed method outperforms its several counterparts and achieves the highest MAP for their average.

The corresponding precision-scope curves and precision-recall curves are plotted in Figure 2. The top row shows the performance of all methods based on the precision-scope curves for both forms of cross-modal retrieval tasks. Similarly, it can be observed that our method performs comparably to GMMFA and GMLDA for the text query, and obtains better results than its several counterparts for the image query. The bottom row shows the precision-recall curves of our method and its several counterparts, and our method again obtains similar results for both forms of cross-modal retrieval.

IV. CONCLUSION

In this paper, we have proposed a joint graph regularized multi-modal subspace learning (JGRMSL) algorithm to learn a common latent space, in which the content similarity between heterogeneous data can be measured. The proposed method explores inter-modality similarities and intra-modality similarities through a joint graph regularization term to better explore the cross-modality correlation and the local manifold structure. To obtain good class separation, the proposed method maximizes the between-class covariance of all projected data, meanwhile minimizing the within-class covariance of all projected data. And the joint graph regularizer, the between-class covariance and the within-class covariance are integrated into a unified formulation, which can be solved as a general eigenvalue problem. Experimental results on two cross-modal datasets have shown that the proposed method outperformed several state-of-the-art methods. In the future, we will establish (or seek) a dataset containing more modalities of data and test our method in the multi-modal case for the cross-modal retrieval task.

ACKNOWLEDGMENT

This work is jointly supported by National Basic Research Program of China (2012CB316300), National Natural Science Foundation of China (61175003, 61135002, 61202328, 61103155), Hundred Talents Program of CAS.

REFERENCES

[1] D. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Cannonical correlation analysis: an overview with application to learning methods," *Neural Computation*, vol. 16, no. 12, pp. 2639–2664, 2004.

[2] N. Rasiwasia, J. C. Pereira, E. Coviello, G. Doyle, G. Lanckriet, R. Levy, and N. Vasconcelos, "A new approach to cross-modal multimedia retrieval," *In ACM MM*, pp. 251–260, 2010.

[3] S. J. Hwang and K. Grauman, "Accounting for the relative importance of objects in image retrieval." *In BMVC*, pp. 1–12, 2010.

[4] R. Udupa and M. Khapra, "Improving the multilingual user experience of wikipedia using cross-language name search," *In NACACL-HLT*, pp. 492–500, 2010.

[5] A. Li, S. Shan, X. Chen, and W. Gao, "Face recognition based on non-corresponding region matching," *In ICCV*, pp. 1060–1067, 2011.

[6] A. Sharma and D. W. Jacobs, "Bypassing synthesis: PLS for face recognition with pose, low-resolution and sketch," *In CVPR*, pp. 593–600, 2011.

[7] J. B. Tenenbaum and W. T. Freeman, "Separating style and content with bilinear models," *Neural Computation*, vol. 12, no. 6, pp. 1247–1283, 2000.

[8] A. Sharma, A. Kumar, H. Daume, and D. W. Jacobs, "Generalized multiview analysis: a discriminative latent space," *In CVPR*, pp. 2160–2167, 2012.

[9] Y. Chen, L. Wang, W. Wang, and Z. Zhang, "Continuum regression for cross-modal multimedia retrieval," *In ICIP*, pp. 1949–1952, 2012.

[10] J. Jagarlamudi, R. Udupa, and H. Daume, "Generalization of cca via spectral embedding." *In The Learning Workshop along with AISTATS 2011*, 2011.

[11] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen, "Multi-view discriminant analysis." *In ECCV*, pp. 808–821, 2012.

[12] S. Hwang and K. Grauman, "Reading between the lines: object localization using implicit cues from image tags," *IEEE TPAMI*, vol. 34, no. 6, pp. 1145–1158, 2012.

[13] N. Rasiwasia, P. J. Moreno, and N. Vasconcelos, "Bridging the gap: query by semantic example." *IEEE TMM*, vol. 9, no. 5, pp. 923–938, 2007.

[14] D. Blei, A. Ng, and M. Jordan, "Latent dirichlet allocation." *JMLR*, vol. 3, pp. 993–1022, 2003.

[15] D. Lowe, "Distinctive image features from scale-invariant keypoints." *IJCV*, vol. 60, no. 2, pp. 91–110, 2004.