METRIC BASED ON MULTI-ORDER SPACES FOR CROSS-MODAL RETRIEVAL

Liang Zhang[1,3], *Bingpeng Ma*[1,2,3⋆], *Guorong Li*[1,3], *Qingming Huang*[1,2,3]

[1] School of Computer and Control Engineering, University of Chinese Academy of Sciences, China
[2] Key Lab of Intell. Info. Process., Inst. of Comput. Tech., CAS, China
[3] Key Laboratory of Big Data Mining and Knowledge Management, CAS, China
zhangliang14@mails.ucas.ac.cn {bpma, liguorong, qmhuang}@ucas.ac.cn

ABSTRACT

This paper proposes a novel method for cross-modal retrieval. Different from vector (text)-to-vector (image) framework of the traditional cross-modal methods, we adopt a vector (text)-to-matrix (image) framework. We assume that compared with vectors, matrices can directly represent images and characterize the structure of feature space. Furthermore, we propose a **M**etric based on **M**ulti-order **s**paces (MMs). Multi-order statistic features are used to represent images for enriching the semantic information, and metrics among the multi-spaces are jointly learned to measure the similarity between two different modalities. Specifically, there are three steps for MMs. First, we jointly use the bags of visual features (zero-order), mean (first-order) and covariance (second-order) to characterize each image. Second, considering that covariance matrices and vectors lie on a Riemannian manifold and an Euclidean space respectively, we embed multi-order spaces into their corresponding Hilbert spaces to reduce the heterogeneity among the original spaces. Finally, the similarity between two different modalities can be measured by learning multiple transformations from the different Hilbert spaces to a common subspace. The performance of the proposed method over the state-of-the-art has been demonstrated through the experiments on two public datasets.

Index Terms— Cross-modal retrieval, Vector-to-matrix metric, Information fusion

1. INTRODUCTION

As the major component of big data, the applications of multi-modal data have become more and more widespread in recent years. Therefore, relevance of multi-modal data has drawn the increasing attentions in many works. The goal of cross-modal retrieval is to provide a correct prediction on whether a pair of samples from two different modalities belong to the same semantic category. There are numerous practical applications of cross-modal retrieval. For example, in Google search engine, we often search the most relevant images (or videos) in response to a textual description, and vice versa. In this paper, we focus on the cross-modal matching problem and apply it to match images and text documents, which has been actively investigated in recent years [1, 2, 3, 4, 5, 6].

As is well known, the fundamental problem of cross-modal retrieval is how to model the correlations among the multi-modal data since different modalities lie on different feature spaces. To alleviate this problem, a lot of algorithms have been proposed to learn a common subspace for different modalities. Generally speaking, there are three kinds of dominant directions to learn common subspace [7].

First, some algorithms learn the maximal correlations among different modalities [2, 3, 8, 9, 10]. Canonical Correlation Analysis (CCA) [8] learned a pair of linear transformations to maximize the correlations between two modalities. Generalized Multiview Analysis (GMA) [10] was a supervised extension of CCA, which solved a joint, relaxed quadratic constrained quadratic program over the different feature spaces to obtain a single (non-)linear subspace.

Second, other algorithms rely on the manifold learning to obtain the common subspace [11, 12]. Since the high dimensional data may embed in a lower dimensional intrinsic space, the manifold learning methods project different modalities into a common manifold by learning their underlying manifold representations. For instance, Mao *et al.* [11] projected the relevance between two different modalities into a latent semantic space during the process of manifold alignment.

Finally, the methods of learning to rank are also applied in cross-modal retrieval to find the common subspace [4, 5, 13, 14, 15]. These methods obtain the common subspace via large margin learning and certain ranking criteria. Specifically, Wu *et al.* [4] proposed Bi-directional Cross-Media Semantic Representation Model (Bi-CMSRM), which obtained a latent space embedding by learning the structural large margin to optimize the bi-directional listwise ranking loss.

All the above methods represent each image (text docu-

⋆Co-corresponding authors

This work was supported in part by National Natural Science Foundation of China (NSFC): 61332016, 61572465, 61429201, 61620106009, U1636214 and 61650202, in part by National Basic Research Program of China (973 Program): 2015CB351800, in part by Key Research Program of Frontier Sciences, CAS: QYZDJ-SSW-SYS013.

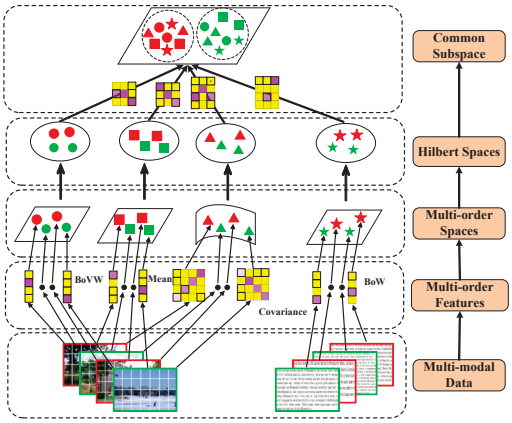


Fig. 1. The flowchart of the proposed method.

ment) as a vector. Then, the similarity between two different modalities is measured by learning the optimal transformations for different feature vectors. We call this framework as vector(text)-to-vector(image). However, since an image is represented by a matrix naturally, it inevitably loses some structural information when we transform a 2D image to 1D vector [16]. In this paper, we use the covariance matrices to characterize the structure of feature space. Different from the framework of vector-to-vector, we first propose the framework of vector(text)-to-matrix(image) to improve the performance of cross-modal retrieval. Furthermore, we propose a novel **M**etric based on **M**ulti-order **s**paces (MMs) to effectively model the correlation between two different modalities. In MMs, the images are represented by the multi-order statistics and the similarity between two modalities is measured by the metrics among the multiple spaces. We show the flowchart of the proposed MMs in Fig. 1.

The novelty and advantages of the proposed approach include: (1) For characterizing the structure information of image feature space, we adopt the covariance matrices as the features of images. Different from the traditional framework of vector(text)-to-vector(image), we propose the framework of vector(text)-to-matrix(image) to model the correlation between two modalities. To the best of our knowledge, it is the first time that the vector-to-matrix framework is proposed in cross-modal retrieval. (2) The images are jointly represented by the zero-, first- and second-order statistical features. Since different order statistics can characterize image from different perspectives, fusing these statistics can provide the complementary information for image. (3) We propose a novel metric to measure the similarity between different modalities based on the multi-order statistical features of images. It combines the metrics of Euclidean-to-Euclidean (E2E) and Euclidean-to-Riemannian (E2R). The weight of the different metrics can be generated automatically.

To show the effectiveness of MMs, we evaluate it on two benchmark datasets, Wiki and NUS-WIDE. Experimental results demonstrate that the performance of MMs is much better than the state-of-the-art methods.

2. THE PROPOSED METHOD

This section introduces the three parts of the proposed method: feature extraction, from multi-order spaces to Hilbert spaces and metric learning.

2.1. Feature Extraction

Statistics of images: We first divide an image into many patches with constant size. Then we compute its SIFT feature [17]. For the ith image, we can obtain its SIFT features $S_i = \{S_{ij}, j = 1, \ldots, k_i\}$, where k_i denotes the number of SIFT features of the ith image.

Based on all the SIFT features of all images, we first learn a codebook by using the k-means algorithm. Then, the SIFT features of each image are quantized by this codebook. Finally, for the ith image, we obtain a histogram feature vector (zero-order statistic feature) h_i, and the dimension of h_i is decided by the number of words in the codebook.

We also compute the first- and second-order statistic features for images. For the ith image, the mean vector m_i and the covariance matrix C_i can be computed as follows:

$$\begin{aligned} m_i &= \frac{1}{k_i}\sum_{j=1}^{k_i} S_{ij} \\ C_i &= \frac{1}{k_i - 1}\sum_{j=1}^{k_i}(S_{ij} - m_i)(S_{ij} - m_i)^T \end{aligned} \tag{1}$$

There are two advantages for selecting the covariance matrices as the representation of images [18]. First, since the covariance matrix doesn't assume the distribution of the features, the image with any number of SIFT features can obtain a natural representation. This representation can discriminate images from the different categories by encoding the feature correlation information specific to each category. Second, as a statistic of all the features in one image, the covariance matrix can largely filter out the noise-corrupting features with an average filter during the computation of the covariance matrix.

Different order statistical information characterizes images from different perspectives. For example, the histogram vector can reflect the distribution of the key words from the codebook, the mean vector roughly reflects the position of the image in the high-dimensional space. The covariance matrix reflects the variance of each individual feature in the diagonal entries and the correlations of the different features in the non-diagonal entries. Hence, these statistical features can provide complementary information to represent images.

After feature extraction, we obtain the training features of images $X = \{x_1, x_2, \ldots, x_n\}$ with the class labels $\{l_1^x, l_2^x, \ldots, l_n^x\}$. The ith image is represented as $x_i =$

(h_i, m_i, C_i). h_i and m_i lie on the different Euclidean spaces $\mathbb{R}^{d_h}$ and $\mathbb{R}^{d_m}$ with the dimensionalities d_h and d_m respectively, and C_i lies on the Riemannian manifold $\mathcal{M}$.

Statistics of texts: For text documents, we extract high dimensional feature vectors using the BoW representation with the TF-IDF weighting scheme. The training features of text documents are denoted as $Y = \{y_1, y_2, \ldots, y_n\}$ with the class labels $\{l_1^y, l_2^y, \ldots, l_n^y\}$, and y_i is the feature of the ith text document. It is obvious that y_i lies on an Euclidean space $\mathbb{R}^{d_y}$, where d_y is the dimension of the text space.

Since the zero-order statistic of texts, the zero-, first-, and second-order statistics of images lie on different spaces, we call these spaces as the multi-order spaces.

2.2. Multi-order Spaces to Hilbert Spaces

After the feature extraction for images and texts, the similarity between the ith image and the jth text can be transformed to calculate the distance $d(x_i, y_j)$ in cross-modal retrieval. In MMs, $d(x_i, y_j)$ is designed as follows:

$$d(x_i, y_j) = \alpha_1 d_1(h_i, y_j) + \alpha_2 d_2(m_i, y_j) + \alpha_3 d_3(C_i, y_j) \quad (2)$$

where $\sum_{r=1}^{3} \alpha_r = 1$, $\alpha_r > 0$, $d_1(h_i, y_j)$ and $d_2(m_i, y_j)$ are Euclidean-to-Euclidean distance (E2E), $d_3(C_i, y_j)$ is Euclidean-to-Riemannian distance (E2R). It is difficult to directly compute $d(x_i, y_j)$ because h_i, m_i and y_j lie on the different Euclidean spaces while C_i and y_j lie on the Riemannian manifold and the Euclidean space, respectively.

However, it is difficult to find the common subspace straightforwardly because of the large heterogeneity gap among the different Euclidean spaces and the Riemannian manifold. Thus, we rely on latent intermediate spaces to reduce the heterogeneity. In MMs, the multi-order spaces are embedded into their corresponding reproducing kernel Hilbert spaces.

We embed the Euclidean spaces into the Hilbert spaces by the non-linear mappings, $\varphi_h : \mathbb{R}^{d_h} \to \mathcal{H}_h$ and $\varphi_m : \mathbb{R}^{d_m} \to \mathcal{H}_m$ for images, and $\varphi_y : \mathbb{R}^{d_y} \to \mathcal{H}_y$ for texts:

$$\begin{aligned} \varphi_h : K_1(i,j) &= exp(-\|h_i - h_j\|^2 / 2\sigma_h^2) \\ \varphi_m : K_2(i,j) &= exp(-\|m_i - m_j\|^2 / 2\sigma_m^2) \\ \varphi_y : K_y(i,j) &= exp(-\|y_i - y_j\|^2 / 2\sigma_y^2) \end{aligned} \quad (3)$$

Specially, the Riemannian manifold $\mathcal{M}$ is embedded into a high dimensional Hilbert space $\mathcal{H}_C$ by φ_C, which uses Log-Euclidean Distance to measure the similarity between the covariance matrices [19].

$$\varphi_C : K_3(i,j) = exp(-\|\log(C_i) - \log(C_j)\|^2 / 2\sigma_C^2) \quad (4)$$

2.3. Hilbert Spaces to the Common Subspace

As is well known, the kernelized features lie on different kernel spaces. It is still difficult to measure the similarity between the heterogeneous data. Therefore, we learn multiple transformations $\mathbf{U} = \{U_1, U_2, U_3\}$ and $\mathbf{V} = \{V\}$ which map the different Hilbert spaces $\mathcal{H}_h$, $\mathcal{H}_m$, $\mathcal{H}_C$ and $\mathcal{H}_y$ into the common subspace $\mathbb{R}^d$, where the distance between the heterogeneous intra-class samples should be minimized, and the inter-class samples should be maximized simultaneously.

2.3.1. Objective function

To obtain the multiple transformations $\mathbf{U}$ and $\mathbf{V}$, we formulate a general objective function $\mathcal{O}(\mathbf{U}, \mathbf{V})$ by combining the metrics of E2E and E2R:

$$\min_{\mathbf{U},\mathbf{V}} \mathcal{O}(\mathbf{U}, \mathbf{V}) = \min_{\mathbf{U},\mathbf{V}} \{L(\mathbf{U}, \mathbf{V}) + \lambda T(\mathbf{U}, \mathbf{V})\}$$

where $L(\mathbf{U}, \mathbf{V})$ is the loss function defined on the sets of similarity and dissimilarity constraints. $T(\mathbf{U}, \mathbf{V})$ is the regularizer defined on the target transformations $\mathbf{U}$ and $\mathbf{V}$. λ is the trade-off parameter and its value is greater than zero.

Compared with the objective function in [6, 20], we improve the objective function in two aspects: (1) We fuse three distance measures from the different Hilbert spaces to constrain L. Generally speaking, the different order statistics characterize the image from different perspectives. Hence, fusing these statistics can provide complementary information for images as that of image set [21]. Furthermore, we also modify the way of solving the weights to explore the complementary property of multi-feature. In our case, it is beneficial to match the images and texts; (2) We ignore the discriminant geometry constraint in [20] and joint graph regularization in [6], because they only constrain the transformations for individual modality. In fact, in the task of matching the images and texts, we need the constraint to enhance the relevance between two different modalities. In MMs, fusing three distance measures is equal to increasing the constraint on transformations for two modalities. Additionally, this also reduces the computation complexity.

Loss function: The goal of $L(\mathbf{U}, \mathbf{V})$ is to minimize the distances of samples with the same label and maximize the distances of samples with the different labels simultaneously. We adopt the weighted sum of the squared distances to constrain E2E and E2R:

$$L(\mathbf{U}, \mathbf{V}) = \sum_{r=1}^{3} \alpha_r L(U_r, V) \quad (5)$$

$$\begin{aligned} L(U_r, V) &= \frac{1}{2} \sum_{i,j=1}^{n} Z_{ij} \left\| U_r^T K_r(:, i) - V^T K_y(:, j) \right\|^2 \\ &= \mathrm{Tr}(U_r^T K_r Q_x K_r^T U_r + V^T K_y Q_y K_y^T V - 2 U_r^T K_r Z K_y^T V) \end{aligned} \quad (6)$$

$$Z_{ij} = \begin{cases} 1 & \text{if } l_i^I = l_j^T \\ -1 & \text{if } l_i^I \neq l_j^T \end{cases} \quad (7)$$

where Z_{ij} represents whether x_i and y_j belong to the same class. We normalize Z by the number of pairs with the same

(different) labels. Q_x, Q_y are the diagonal matrices with $Q_x(i,i) = \sum_{j=1}^{n} Z_{ij}$ and $Q_y(j,j) = \sum_{i=1}^{n} Z_{ij}$.

Transformation regularization: $T(\mathbf{U}, \mathbf{V})$ is defined to control the scale of the transformations and reduce overfitting. Its formulation is defined as follows:

$$T(\mathbf{U}, \mathbf{V}) = \frac{1}{2} \sum_{r=1}^{3} \left\| U_r^T K_r \right\|_F^2 + \frac{1}{2} \left\| V^T K_y \right\|_F^2 \quad (8)$$

2.3.2. Iterative Optimization

To obtain the optimal solution for $\mathcal{O}(\mathbf{U}, \mathbf{V})$, we design a iterative strategy to minimize the objective function. We initialize $\mathbf{U}$ and $\mathbf{V}$ with the within- and between-class analysis, and α = [α_1, α_2, α_3] with equal value $1/3$. Since $\mathbf{U}$ and $\mathbf{V}$ are independent to each other, we first update V, and fix $\mathbf{U}$ as well as α. Then, we update U_1, U_2 and U_3 simultaneously when $\mathbf{V}$ and α are given. After obtaining $\mathbf{U}$ and $\mathbf{V}$, we solve α by applying the conventional Lagrange multiplier.

Update V: Differentiating $\mathcal{O}(\mathbf{U}, \mathbf{V})$ with respect to V, we can get the following formulation:

$$\frac{\partial \mathcal{O}(\mathbf{U}, \mathbf{V})}{\partial V} = K_y Q_y K_y^T V + 2\lambda K_y K_y^T V - \sum_{r=1}^{3} \alpha_r K_y Z K_r^T U_r \quad (9)$$

Then, setting Eq. (9) to 0, we can obtain V as:

$$V = \left(K_y Q_y K_y^T + 2\lambda K_y K_y^T \right)^{-1} \left(\sum_{r=1}^{3} \alpha_r K_y Z K_r^T U_r \right) \quad (10)$$

Update U: Differentiating $\mathcal{O}(\mathbf{U}, \mathbf{V})$ with respect to U_r and setting it to zero, we can achieve U_r, $r = 1, 2, 3$:

$$U_r = (K_r Q_x K_r^T + \frac{2\lambda}{\alpha_r} K_r K_r^T)^{-1} K_r Z K_y^T V \quad (11)$$

Update α: After obtaining the transformations, we apply the Lagrange multiplier to explore the complementary information of the different features. we modify α_r as α_r^p, where $p > 1$, and construct the Lagrange function $\bar{\mathcal{O}}(\alpha, \eta)$ as:

$$\bar{\mathcal{O}}(\alpha, \eta) = \sum_{r=1}^{3} \alpha_r^p L(U_r, V) + \lambda T(\mathbf{U}, \mathbf{V}) + \eta \left(\sum_{r=1}^{3} \alpha_r - 1 \right) \quad (12)$$

Differentiating $\bar{\mathcal{O}}(\alpha, \eta)$ with respect to η and α_r, and setting them to zero, we have:

$$p\alpha_r^{p-1} L(U_r, V) - \eta = 0 \quad s.t. \sum_{r=1}^{3} \alpha_r - 1 = 0 \quad (13)$$

According to the Eq. (13), α_r is obtained as follows:

$$\alpha_r = \frac{(1/L(U_r, V))^{1/(p-1)}}{\sum_{r=1}^{3} (1/L(U_r, V))^{1/(p-1)}} \quad (14)$$

To obtain a locally optimal solution, we update $\mathbf{U}$, $\mathbf{V}$ and α until the performance on the validation set decreases. In practice, MMs can converge to the stable and desirable solution after several iterations.

3. EXPERIMENT

The section is utilized to demonstrate the effectiveness of the proposed MMs for cross-model retrieval. We compare different methods on two publicly available datasets.

3.1. Experimental Settings

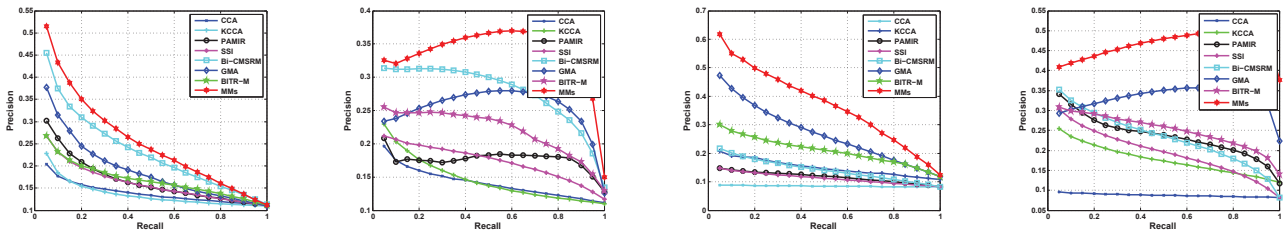
We compare MMs with the following methods: CCA and KCCA [8], Bi-CMSRM [4], GMA [10], BITR [13], PAMIR [14] and SSI [15]. Since MMs adopts the multi-order statistics to represent images, we also report the performance of MMs-h, MMs-m and MMs-C. They just use one kind of image' statistic feature to measure the similarity between images and texts. For example, MMs-h denotes that we just use the images' zero-order features to match images and texts.

In this paper, the mean average precision (MAP) [3] and the precision-recall curve [3] are used to evaluate the performance of all methods. For all methods, parameters are set by 5-fold cross validation to achieve the best performance. For MMs, the parameters p, λ are respectively set to 2, 0.001. On both datasets, the dimensionality of the latent space is set to 10 for all the methods. For fairness, we follow the experimental protocol of samples selection in [4] for both datasets. We conduct experiments 10 times by randomly selecting training/validation/testing combinations, and compute and compare the average MAP of different methods.

3.2. Results on the Wiki Dataset

The Wiki dataset contains 2,866 articles generating from Wikipedia's featured articles [3]. Each article consists of an image and a text description which is categorized into 10 semantic classes. We randomly choose 1,500 pairs of the data for training, 500 pairs for validation and 866 pairs for testing. The MAP scores of different methods are shown in Tab. 1, from which we can draw the following conclusions:

First, in both directional retrieval, the MAP scores of MMs-h outperform those of other comparative methods except MMs-C and MMs. For example, the MAP scores of MMs-h achieve 0.2165 and 0.2856 for text query and image query, while Bi-CMSRM obtains 0.2123 and 0.2528, respectively. We attribute the improvement of MMs-h to the metric framework. In MMs, the metric framework takes both the positive and negative pairs to constrain the loss function, while Bi-CMSRM focuses on the ranking function to optimize the listwise ranking loss. This experiment demonstrates that our metric framework can effectively measure the similarity between two different modalities.



(a) Wiki Text Query (b) Wiki Image Query (c) NUS Text Query (d) NUS Image Query

Fig. 2. Precision-Recall curves on the Wiki and NUS-WIDE datasets.

Table 1. The MAP scores of different methods on Wiki and NUS-WIDE.The items in bold are the best results.

Methods \ Tasks	Wiki			NUS-WIDE		
	Text query	Image query	Average	Text query	Image query	Average
SSI	0.1664	0.1759	0.1712	0.1140	0.1992	0.1566
CCA	0.1433	0.1451	0.1442	0.0851	0.0883	0.0867
GMA	0.1987	0.2566	0.2277	0.2768	0.3328	0.3048
BITR	0.1752	0.2230	0.1991	0.2143	0.2543	0.2343
KCCA	0.1421	0.1503	0.1462	0.1680	0.1806	0.1743
PAMIR	0.1734	0.1779	0.1757	0.1184	0.2410	0.1797
Bi-CMSRM	0.2123	0.2528	0.2326	0.1453	0.2380	0.1917
MMs-h	0.2165	0.2856	0.2511	0.2670	0.3349	0.3010
MMs-m	0.1767	0.2494	0.2131	0.1992	0.2653	0.2323
MMs-C	0.2453	0.3244	0.2849	0.2715	0.3427	0.3071
MMs	**0.2689**	**0.3504**	**0.3097**	**0.3712**	**0.4404**	**0.4058**

Second, the performance of MMs-C is much better than that of MMs-h. Specially, the MAP scores of MMs-C can obtain 0.2453 and 0.3244 for text query and image query. Considering that the difference between MMs-h and MMs-C is that images in MMs-C are represented by the 2D covariance matrices, the results show that compared with the 1D vectors, the 2D covariance matrices can characterize the structure of feature space, which is beneficial to match texts and images.

Finally, by integrating MMs-h, MMs-m and MMs-C, the performance of MMs can be further improved. For MMs, the MAP scores of text and image query are 0.2689 and 0.3504 respectively. Compared with other methods, MMs achieves the best performance on the Wiki dataset. These results show that fusing the zero-, first- and second- statistical features of image can enrich the semantic information such that the correlations between two modalities are enhanced.

The Precision-Recall curves on both directional retrieval are shown in Fig. 2(a) and 2(b). The curves further validate the superiority of MMs for cross-modal retrieval.

3.3. Results on the NUS-WIDE Dataset

The NUS-WIDE dataset consists of 269,648 paired samples with 81 concepts [22]. Each image with its annotated tags can be treated as a pair of image-text data while the concepts are regarded as the labels. We randomly select 6,664 images that have at least one tag and one concept from the 10 largest categories. Then 2,664 paired samples are used for training, 2,000 for validation and 2,000 for testing. The MAP scores of all methods are shown in Tab. 1. The Precision-Recall curves are shown in Fig. 2(c) and 2(d).

On this dataset, MMs-C achieves the comparable performance with GMA, which obtains the best performance except that of MMs. The performance of MMs is still the best of all the methods and it improves 34.10% and 32.33% compared with GMA. These results validate that the above analyses on the Wiki dataset are reasonable and MMs can achieve the stable performance on different datasets.

4. CONCLUSIONS

Motivated by the fact that images should be represented by the 2D matrices to preserve spatial information, this paper adopts a novel framework for cross-modal retrieval. Under this framework, we further propose a novel method named MMs. In MMs, the representation of images is enriched by the multi-order statistics and the metrics among the multi-spaces are jointly used to measure the similarity between two modalities. Experiments on two benchmark datasets (Wi-

ki and NUS-WIDE) have shown that the proposed method achieves the state-of-the-art performance.

5. REFERENCES

[1] C. Kang, S. Xiang, S. Liao, C. Xu, and C. Pan, "Learning consistent feature representation for cross-modal multimedia retrieval," *IEEE Transactions on Multimedia*, vol. 17, no. 3, pp. 370–381, 2015.

[2] V. Ranjan, N. Rasiwasia, and C. Jawahar, "Multi-label cross-modal retrieval," in *IEEE International Conference on Computer Vision*, 2015.

[3] N. Rasiwasia, J. Pereira, E. Coviello, G. Doyle, G. Lanckriet, R. Levy, and N. Vasconcelos, "A new approach to crossmodal multimedia retrieval," in *ACM International Conference on MultiMedia*, 2010.

[4] F. Wu, X. Lu, Z. Zhang, S. Yan, Y. Rui, and Y. Zhuang, "Cross-media semantic representation via bi-directional learning to rank," in *ACM International Conference on MultiMedia*, 2013.

[5] L. Zhang, B. Ma, G. Li, Q. Huang, and Q. Tian, "Pl-ranking: A novel ranking method for cross-modal retrieval," in *ACM International Conference on MultiMedia*, 2016.

[6] X. Zhai, Y. Peng, and J. Xiao, "Heterogeneous metric learning with joint graph regularization for cross-media retrieval," in *AAAI Conference on Artificial Intelligence*, 2013.

[7] J. Pereira, E. Coviello, G. Doyle, N. Rasiwasia, G. Lanckriet, R. Levy, and N. Vasconcelos, "On the role of correlation and abstraction in cross-modal multimedia retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 3, pp. 521–535, 2014.

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

[9] N. Rasiwasia, D. Mahajan, V. Mahadevan, and G. Aggarwal, "Cluster canonical correlation analysis," in *International Conference on Artificial Intelligence and Statistics*, 2014.

[10] A. Sharma, A. Kumar, D. Hal, and D. Jacobs, "Generalized multiview analysis: A discriminative latent space," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[11] X. Mao, B. Lin, D. Cai, X. He, and J. Pei, "Parallel field alignment for cross media retrieval," in *ACM International Conference on MultiMedia*, 2013.

[12] C. Wang and S. Mahadevan, "Heterogeneous domain adaptation using manifold alignment," in *International Joint Conference on Artificial Intelligence*, 2011.

[13] Y. Verma and C. Jawahar, "Im2text and text2im: Associating images and texts for cross-modal retrieval," in *British Machine Vision Conference*, 2014.

[14] D. Grangier and S. Bengio, "A discriminative kernel-based approach to rank images from text queries," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 8, pp. 1371–1384, 2008.

[15] B. Bai, J. Weston, D. Grangier, R. Collobert, K. Sadamsda, Y. Qi, O. Chapelle, and K. Weinberger, "Learning to rank with (a lot of) word features," *Information Retrieval*, vol. 13, no. 3, pp. 291–314, 2010.

[16] Z. Liang, B. Ma, G. Li, Q. Huang, and Q. Tian, "Cross-modal retrieval using multi-ordered discriminative structured subspace learning," *IEEE Transactions on Multimedia*, 2016.

[17] D. Lowe, "Distinctive image features from scale invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[18] R. Wang, H. Guo, L. Davis, and Q. Dai, "Covariance discriminative learning: A natural and efficient approach to image set classification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[19] V. Arsigny, P. Fillard, X. Pennec, and N. Ayache, "Geometric means in a novel vector space structure on symmetric positive-definite matrices," *SIAM Journal of Matrix Analysis and Applications*, vol. 29, no. 1, pp. 328–347, 2007.

[20] Z. Huang, R. Wang, S. Shan, and X. Chen, "Learning euclidean-to-riemannian metric for point-to-set classification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[21] J. Lu, G. Wang, and P. Moulin, "Image set classification using holistic multiple order statistics features and localized multi-kernel metric learning," in *IEEE International Conference on Computer Vision*, 2013.

[22] T. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, "Nus-wide: A real-world web image database from national university of singapore," in *ACM International Conference on Image and Video Retrieval*, 2009.