LEARNING A MID-LEVEL FEATURE SPACE FOR CROSS-MEDIA REGULARIZATION

Yunchao Wei, Yao Zhao, Zhenfeng Zhu, Yanhui Xiao, Shikui Wei

Institute of information Science, Beijing Jiaotong University, Beijing, 100044, China
Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing, 100044, China
wychao1987@gmail.com

ABSTRACT

In this paper, we propose a cross-media regularization framework to enhance image understanding which can benefit image retrieval, classification and so on. The goal of cross-media regularization is to find regularization projections by exploiting the correlations between visual features and textual features. Thus, the original noisy distribution of visual features can be refined by leveraging the discriminative distribution of the corresponding textual features. Within the proposed cross-media regularization framework, a mid-level representation is built by jointly projecting both visual and textual features into a shared feature subspace, which leads to transferring of the discriminative semantic characteristic embedded in the textual modality into the corresponding visual modality. Meanwhile, the discriminative characteristic of textual features can also be boosted simultaneously. The experimental results demonstrate that the proposed mid-level space learning process can remarkably improve the search quality and outperform the existing semantic regularization methods.

Index Terms— Cross-media, Image Search, Subspace Learning, Knowledge Transfer

1. INTRODUCTION

Content-based image search aims to retrieve semantic relevant images from a large image database by using similarity matching of visual features. Since existing image understanding techniques [1, 2, 3, 4] only exploit visual features, the performance is far from satisfactory. Therefore, it is natural to improve image search performance by introducing more distinguishable cues. In fact, with the rapid development of the Internet and information techniques, the cross-media (i.e., cross-modal) data which represent the same semantics have been widely available in the real world. For example, an image usually co-occurs with text on a web page to describe the same object, action or state. Although they are in different modalities, the reflected semantics is consistent. In general, the textual modality is provided with a more discriminative semantic distribution characteristic than the visual modality due to the bottleneck of visual feature extraction techniques. Hence, making full use of the discriminative characteristic of textual features can potentially benefit image understanding and finally facilitate the image search.

Some approaches [5, 6, 7, 8, 9] have been developed to model the relationships among different modalities. As two classical methods, Canonical Correlation Analysis (CCA) [10] and Partial Least Squares (PLS) [11] are usually employed to find multi-variable correlations. CCA attempts to find the directions which maximize the correlations between two multidimensional variables, while PLS aims to find the directions of maximum covariance. In addition, the problem of transferring knowledge from text to image is also studied in [7, 8] to improve the performance of image classification. However, the above-mentioned methods do not take advantage of the discriminative distribution characteristic of textual features to refine the noisy distribution of visual features. Recently, a regularized image semantics (RIS) method [9] is proposed to refine the visual features by using the discriminative distribution characteristic of their corresponding textual features. RIS firstly estimates the high-level semantic features of images and text from their low-level feature spaces respectively, and then the high-level semantic features of text are exploited to regularize the corresponding high-level semantic features of images. However, since the high-level semantic features are directly derived from low-level features, the semantic gap will result in noisy high-level semantic features and then decrease the effectiveness of regularization. Therefore, it is necessary to construct a mid-level feature space to bridge the semantic gap as well as take knowledge transferring into account.

In this paper, we propose a cross-media regularization framework to enhance image understanding which can benefit image classification, retrieval and so on. Within the proposed framework, a mid-level representation is built by jointly projecting both visual and textual features into a shared mid-level feature space. In this mid-level feature space learning (MSL) approach, there is the transferring of the discriminative semantic characteristic embedded in the textual modality into the corresponding visual modality. Meanwhile, the discriminative characteristic of textual features can also be boosted simultaneously. Then, more efficient high-level sematic features for images and text can be calculated from the mid-level feature space rather than the low-level feature space, and the

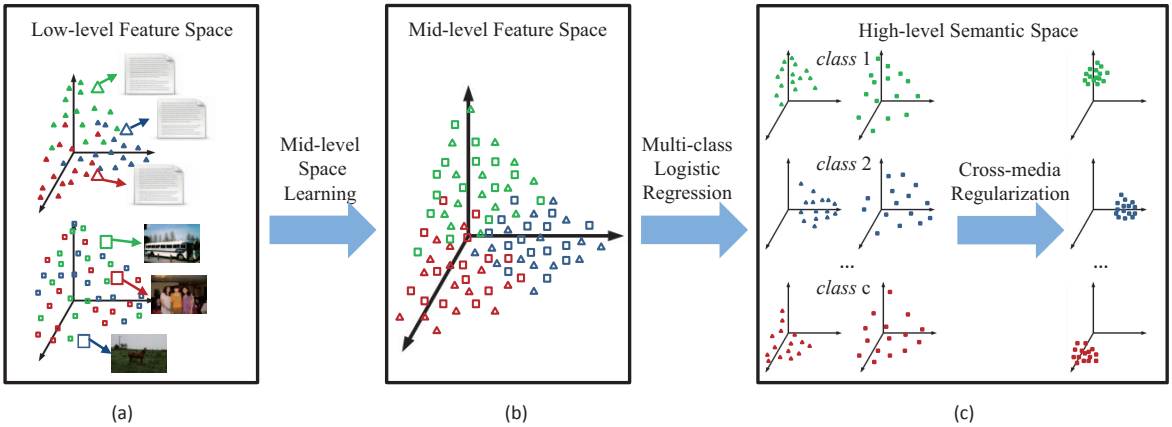

Fig. 1. The overall framework of the proposed cross-media regularization method. Images are represented by square icons and text is represented by triangle icons. Different colors indicate different classes. (a) Images and text are represented by low-level features. (b) Through the proposed mid-level space learning method, images and text are projected into a common mid-level feature space. (c) High-level semantic features derived from the mid-level feature space are used for cross-media regularization.

cross-media regularizing process proposed in [9] can be carried out more naturally and smoothly.

The remainder of this paper is organized as follows. In Section 2, a general framework of the proposed cross-media regularization is introduced. Section 3 shows the mid-level feature space learning method in detail. Comparative analysis of experimental results is presented in Section 4. The conclusion and summary is presented in Section 5.

2. FRAMEWORK OF CROSS-MEDIA REGULARIZATION

The goal of cross-media regularization is to find regularization projections by exploiting the correlations between visual features and textual features. Thus, the original noisy distribution of visual features can be refined by leveraging the discriminative distribution of their corresponding textual features. Figure 1 illustrates the overall framework of the proposed cross-media regularization. As shown in Fig. 1(a), both images and text are individually represented in their low-level feature spaces, and class information is associated with each image and text. To better understand our work, we will firstly make a brief introduction of RIS [9].

In RIS, high-level semantic features of images and text are independently calculated from their low-level features. Actually, the high-level semantic feature is a probability vector that each component is a posterior probability under each concept (i.e., class). Each vector spanning the high-level semantic space needs to satisfy the condition that all the components are non-negative and added to one. The high-level semantic features can be learned via multi-class logistic regression as indicated in Fig. 1. Assume there are c classes in the training set. The ith class contains n_i pairs of training data, each of which is associated with a c-d visual feature and a c-d textual feature. Denote $\mathbf{A}_i = [\mathbf{a}_1^{(i)}, ..., \mathbf{a}_{n_i}^{(i)}]^T \in \mathbb{R}^{n_i \times c}$ and $\mathbf{D}_i = [\mathbf{d}_1^{(i)}, ..., \mathbf{d}_{n_i}^{(i)}]^T \in \mathbb{R}^{n_i \times c}$ as two feature matrices of text and images in the ith class, respectively. The goal of cross-media regularization proposed in RIS is to calculate a set of regularization operators $\mathbf{B}_i \in \mathbb{R}^{c \times c}$ $(i = 1, ..., c)$, which makes the distribution characteristic of visual features as similar as possible to that of the corresponding textual features for each class. As shown in Fig. 1(c), the high-level semantic features derived from text documents have much smaller variance than those derived from images, That is, the high-level semantic features of the text are more distinguishable than those of images. This is the reason why the textual features can be used to regularize their corresponding visual features. For the ith class, the cross-media regularization problem is formulated as:

$$\min_{\mathbf{B}_i} \|\mathbf{D}_i\mathbf{B}_i - \mathbf{A}_i\|_F^2 \qquad (1)$$

$$s.t.\ \ (\mathbf{d}_j^{(i)})^T\mathbf{b}_l^{(i)} \geq 0, (\mathbf{d}_j^{(i)})^T\mathbf{B}_i\mathbf{1} = 1, \forall j = 1...n_i, \forall l = 1...c$$

where $\mathbf{b}_l^{(i)}$ is the lth column of $\mathbf{B}_i$ and the constraint is used to keep the transformed features satisfying the distribution condition in the high-level semantic space. Once the regularization operators are obtained, a class-sensitive regularization step is further employed to regularize images via a non-linear regularizer. More details of class-sensitive regularization can be found in [9].

The main difference between our proposed framework and RIS lies in the mid-level feature space learning step, i.e., Fig. 1(b). As discussed above, the high-level semantic space is independently calculated from the low-level feature space in RIS. This procedure not only fails to exploit the discriminative distribution characteristic of textual features, but also results in a large semantic gap which will decrease the effectiveness of the cross-media regularization. In this paper, a

mid-level feature space learning method is proposed to provide higher abstract for low-level features, especially for visual features. Two classic techniques, Linear Discriminant Analysis (LDA) [12] and Canonical Correlation Analysis (CCA) [10] are jointly involved to learn the mid-level features. By using the discriminant analysis on text, the discriminative semantic characteristic of textual features can be further enhanced. Simultaneously, this characteristic of textual features can be transferred to their corresponding visual features via the correlation analysis process. Then the high-level semantic space can be smoothly calculated from the mid-level feature space, which will benefit the cross-media regularization.

3. MID-LEVEL FEATURE SPACE LEARNING

Suppose we are given a dataset of n data instances, i.e., $\mathcal{G} = \{(\mathbf{x}_i, \mathbf{t}_i, y_i)\}_{i=1}^{n}$, where $\mathbf{x}_i \in \mathbb{R}^{d_1}$ and $\mathbf{t}_i \in \mathbb{R}^{d_2}$ are low-level features of image and text, respectively. $y_i \in \{1, ..., c\}$ is the corresponding class label, which is shared by $\mathbf{x}_i$ and $\mathbf{t}_i$. Let $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_n]^T \in \mathbb{R}^{n \times d_1}$ be the feature matrix of image data, and $\mathbf{T} = [\mathbf{t}_1, ..., \mathbf{t}_n]^T \in \mathbb{R}^{n \times d_2}$ be the feature matrix of text data.

In this section, mid-level feature space learning (MSL) algorithm, which is the key step to improve the effectiveness of cross-media regularization, is presented in details. Before we introduce the MSL method, we firstly give a quick review of two classic techniques, i.e., Linear Discriminant Analysis (LDA) [12] and Canonical Correlation Analysis (CCA) [10].

3.1. Background

3.1.1. Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) explicitly attempts to model the difference between the classes of data. Denote $\bar{\boldsymbol{m}}_{\boldsymbol{j}}$ as the class mean of the textual features in the jth class. Denote $\bar{\boldsymbol{m}}$ as the mean of all the textual features. The within-class scatter matrix is defined as $\mathbf{S}_w = \sum_{j=1}^{c} \sum_{y_i=j} {}^1\!/_n \, (\mathbf{t}_i - \bar{\mathbf{m}}_j)(\mathbf{t}_i - \bar{\mathbf{m}}_j)^T$, and the total scatter matrix is defined as $\mathbf{S}_t = \sum_{i=1}^{n} {}^1\!/_n \, (\mathbf{t}_i - \bar{\mathbf{m}}) \, (\mathbf{t}_i - \bar{\mathbf{m}})^T$. The objective function that has been widely used for LDA is formulated as follows:

$$\min_{\mathbf{W}\mathbf{W}^T = \mathbf{I}_k} \frac{tr\left(\mathbf{W}\mathbf{S}_w\mathbf{W}^T\right)}{tr\left(\mathbf{W}\mathbf{S}_t\mathbf{W}^T\right)} \tag{2}$$

where $tr(\cdot)$ denotes the trace of a square matrix, $\mathbf{W} = [\mathbf{w}_1, ..., \mathbf{w}_k]^T \in \mathbb{R}^{k \times d_2}$ is a mapping matrix composed of k basis vectors and $\mathbf{I}_k$ denotes the identity matrix.

3.1.2. Canonical Correlation Analysis

Canonical Correlation Analysis (CCA) is a well known technique for data analysis and dimensionality reduction. CCA is often used to analyze the linear relationships between two multidimensional variables. In particular, canonical correlation aims to choose two directions, $\mathbf{v}$ and $\mathbf{w}$, to maximize the correlation between the two vectors, i.e., $\max_{\mathbf{v},\mathbf{w}} corr\,(\mathbf{X}\mathbf{v}, \mathbf{T}\mathbf{w})$. By CCA, two lists of basis vectors can be learned to project low-level features of images and text into a common subspace. As indicated by [13], CCA is equivalent to the following constrained optimization problem:

$$\min_{\mathbf{V},\mathbf{W}} \left\|\mathbf{X}\mathbf{V}^T - \mathbf{T}\mathbf{W}^T\right\|_F^2 \tag{3}$$

$$s.t.\ \mathbf{V}\mathbf{X}^T\mathbf{X}\mathbf{V}^T = \mathbf{I}_k \quad \mathbf{W}\mathbf{T}^T\mathbf{T}\mathbf{W}^T = \mathbf{I}_k$$

where $\mathbf{V} = [\mathbf{v}_1, ..., \mathbf{v}_k]^T \in \mathbb{R}^{k \times d_1}$ and $\mathbf{W} = [\mathbf{w}_1, ..., \mathbf{w}_k]^T \in \mathbb{R}^{k \times d_2}$ are two mapping matrices composed of k basis vectors, respectively.

3.2. Mid-level Feature Space Learning Method

The key idea of MSL is to find a shared feature space $\mathbb{R}^k$ for images and text, in which the distribution of visual features has a similar characteristic as that of their corresponding textual features. To this end, MSL learns two mapping matrices, i.e., $\mathbf{W}$ and $\mathbf{V}$, for textual features and visual features respectively, which can be formulated as the following optimization framework:

$$\min_{\mathbf{W},\mathbf{V}} C(\mathbf{W}, \mathbf{V}, \mathbf{X}, \mathbf{T}) + L(\mathbf{W}, \mathbf{S}_w, \mathbf{S}_t) \tag{4}$$

where, $C(\mathbf{W}, \mathbf{V}, \mathbf{X}, \mathbf{T})$ is a correlation analysis term and $L(\mathbf{W}, \mathbf{S}_w, \mathbf{S}_t)$ is a discriminant analysis term.

For the first item, CCA is adopted to analyze the correlation between visual and textual modalities. For the second item, LDA is employed to enhance the discriminative characteristic of textual features. To overcome the computational difficulty with LDA, we transform the traditional formulation into the minimum margin form indicated by [14]. Meanwhile, in order to keep the accuracy of the correlation analysis term, we employ the orthogonally constrained CCA (OCCA) [15] method by loosening the constraints. Then, the optimization framework (4) can be formulated as:

$$\min_{\mathbf{V},\mathbf{W}} \underbrace{\left\|\mathbf{X}\mathbf{V}^T - \mathbf{T}\mathbf{W}^T\right\|_F^2}_{C(\mathbf{W},\mathbf{V},\mathbf{X},\mathbf{T})} + \underbrace{tr\left(\mathbf{W}\mathbf{S}_w\mathbf{W}^T\right) - \lambda tr\left(\mathbf{W}\mathbf{S}_t\mathbf{W}^T\right)}_{L(\mathbf{W},\mathbf{S}_w,\mathbf{S}_t)} \tag{5}$$

$$s.t.\ \mathbf{W}\mathbf{W}^T = \mathbf{I}_k$$

where λ is the balance parameter.

As indicated in (5), if we fix the projection directions $\mathbf{W}$, $\mathbf{V}$ can be obtained via a linear regression operation from $\mathbf{X}$ to $\mathbf{T}\mathbf{W}^T$. For the fixed $\mathbf{W}$, the optimal $\mathbf{V}$ can be computed by an approximate solution $\mathbf{V}^T = \mathbf{X}^+\mathbf{T}\mathbf{W}^T$, where $\mathbf{X}^+ = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T$ is the pseudo-inverse of $\mathbf{X}$. If we define $\mathbf{H} = \mathbf{X}\mathbf{X}^+$, the optimization problem (5) can then be formulated

as:

$$\begin{aligned}\min_{\mathbf{W}\mathbf{W}^T=\mathbf{I}_k} & tr\left(\mathbf{W}\mathbf{T}^T(\mathbf{H}-\mathbf{I}_k)^T(\mathbf{H}-\mathbf{I}_k)\mathbf{T}\mathbf{W}^T\right)+ \\ & tr\left(\mathbf{W}\mathbf{S}_w\mathbf{W}^T\right)-\lambda tr\left(\mathbf{W}\mathbf{S}_t\mathbf{W}^T\right) \\ \Rightarrow & \\ \max_{\mathbf{W}\mathbf{W}^T=\mathbf{I}_k} & tr(\mathbf{W}\mathbf{M}\mathbf{W}^T)\end{aligned} \tag{6}$$

where $\mathbf{M}=\lambda\mathbf{S}_t-\mathbf{S}_w-\mathbf{T}^T(\mathbf{H}-\mathbf{I}_k)^T(\mathbf{H}-\mathbf{I}_k)\mathbf{T}$. In this paper, λ is experimentally set as 0.5. $\mathbf{W}$ can be obtained by taking the eigenvectors corresponding to the top-k largest eigenvalues of $\mathbf{M}$. Once $\mathbf{W}$ has been obtained, $\mathbf{V}$ can also be calculated through a linear regression operation. Afterwards, the mid-level features of images and text can be obtained by $\mathbf{x}_i^{'}=\mathbf{V}\mathbf{x}_i$, $\mathbf{t}_i^{'}=\mathbf{W}\mathbf{t}_i$.

Algorithm 1 Mid-level Feature Space Learning

Input: The feature matrix of image data $\mathbf{X}=[\mathbf{x}_1,...,\mathbf{x}_n]^T\in\mathbb{R}^{n\times d_1}$, the feature matrix of text data $\mathbf{T}=[\mathbf{t}_1,...,\mathbf{t}_n]^T\in\mathbb{R}^{n\times d_2}$ and the parameter λ.

1. Compute the within-class scatter matrix $\mathbf{S}_w$ and the total scatter matrix $\mathbf{S}_t$ of text data;
2. Compute the matrix $\mathbf{M}$;
3. Perform SVD on matrix $\mathbf{M}$ ($\mathbf{M}=\mathbf{U}^T\Sigma\mathbf{U}$). $\mathbf{W}$ corresponds to the top-k rows of $\mathbf{U}$;
4. Compute $\mathbf{V}$ through a linear regression operation according to $\mathbf{V}=\mathbf{W}\mathbf{T}^T(\mathbf{X}^+)^T$;
5. Project visual features and textual features from the low-level feature space into the mid-level feature space via $\mathbf{x}_i^{'}=\mathbf{V}\mathbf{x}_i$. $\mathbf{t}_i^{'}=\mathbf{W}\mathbf{t}_i$;

output: $\mathbf{V}$, $\mathbf{W}$, $\mathcal{G}=\{(\mathbf{x}_i^{'},\mathbf{t}_i^{'},y_i)\}_{i=1}^n$

4. EXPERIMENTAL RESULTS

In this section, we conduct two experiments to test the performance of the proposed algorithm. To verify the effectiveness of mid-level feature space learning (MSL) method, we firstly compare our approach with other subspace learning methods. Then, we study the performance of our proposed framework compared with regularized image semantics (RIS) [9] through image retrieval task.

4.1. Experiment Setup

Our selection of datasets is motivated by the consideration that image is accompanied both by text and ground-truth label. The first dataset is Wikipedia dataset used in [16], which contains total 2866 image-text pairs from 10 categories. The whole dataset is randomly split into two parts, one for training with 2173 pairs and the other for test with 693 pairs. The second dataset is Pascal sentences described in [17], which contains total 1000 pairs from 20 categories. 80% of Pascal sentences are randomly selected as the training set, and the others are treated as the test set.

To represent image content, we first densely extract SIFT descriptors from images and further build a visual dictionary with 1024 visual words by k-means vector quantization. Using the visual dictionary, each image can be represented by a 1024-d vector of visual words. For textual information, we firstly obtain the feature vectors based on 500 tokens through Natural Language Toolkit (NLTK) [18]. All tokens are extracted and stemmed from text documents. Then, the Latent Dirichlet Allocation (LDA) model is used to compute the probability of each text document under 100 hidden topics. The probability vectors are used for text representation. To learn high-level semantic space from low-level space or mid-level space, the multi-class logistic regression [19] is employed in our experiments.

4.2. Performance Evaluation

4.2.1. Image Classification

To verify the effectiveness of MSL, we compare our approach with two dimensionality reduction techniques, i.e., PCA [20] and CCA. For each method, we utilize logistic regression as the classifier. Table 1 shows the image classification results for different dimensionality reduction methods. As shown in Table 1, the best performance is achieved by the proposed method on Wikipedia and Pascal Sentence. For PCA, the textual information is not taken into account. Although CCA exploits the textual information, it only cares about pair-wise closeness in the latent feature space and is not applicable for the classification task.

Table 1. Image classification accuracy of different dimensionality reduction methods on Wikipedia and Pascal sentence. The best results are highlighted in bold.

Classifier	PCA	CCA	MSL
Wikipedia	30.1%	29.3%	**31.3%**
Pascal	18%	17.5%	**19.5%**

In Figure 2, we illustrate the image classification accuracy of different dimensionality reduction methods with varying reduction dimensions. As shown in Fig. 2, MSL gives a better performance compared with PCA and CCA in most of the cases.

To further illustrate the effectiveness of MSL, we compare the uni-model classification accuracy for both image and text in low-level feature space (abbreviated as LF) and mid-level

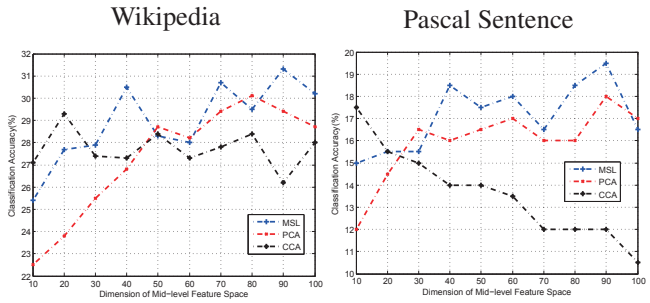

Fig. 2. Image classification accuracy of different dimensionality reduction methods on different feature dimensions.

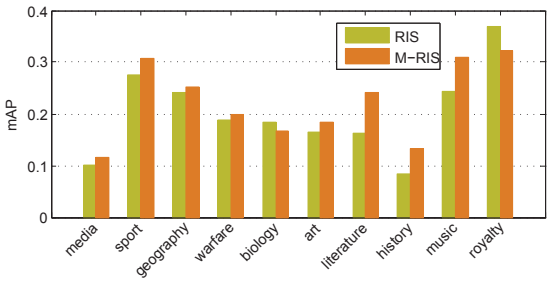

Fig. 3. mAP scores of each class on Wikipedia dataset.

feature space (abbreviated as MF). The experimental results are shown in Table 2. As expected, the MF scheme clearly outperforms the LF scheme, especially for the image modality. This means that our proposed method is indeed effective for enhancing the discriminative capability of the two kinds of features.

Table 2. Uni-modal classification accuracy for both image and text in LF and MF.

Classifier	Wikipedia		Pascal	
	LF	MF	LF	MF
Image	29.7%	**31.3%**	17.5%	**19.5%**
Text	73.6%	**74.8%**	74.5%	**75.0%**

Table 3. Summary of mAP scores. These scores are averaged over all queries. Gains in mAP scores towards our proposed scheme (M-RIS) are shown in (%).

Method	Wikipedia		Pascal	
	mAP	%	mAP	%
M-RIS	**24.2%**	-	**21.8%**	-
RIS	22.5%	8	19.3%	13

4.2.2. Image retrieval

To evaluate the effectiveness of MSL on image regularization, the proposed regularization scheme (M-RIS, for short) is compared with the original RIS scheme on the image retrieval task. The mean average precision (mAP) is used to evaluate the performance. First of all, both images and text are projected into the mid-level feature space from the low-level feature space via the learned mapping matrices, i.e., $\mathbf{V}$ and $\mathbf{W}$. Then, the high-level semantic features for images and text are calculated from their corresponding mid-level features, which can be implemented by multi-class logistic regression [19]. Finally, a set of regularization operators for each class are learned through cross-media regularization as indicated by (1), and then a class-sensitive regularization is further employed to regularize images. For details of class-sensitive regularization and and retrieval method please refer to [9]. The experimental results are summarized in Table 3. The mAP gains on Wikipedia and Pascal sentence are 8% and 13%, respectively.

The detailed mAP scores of each class on Wikipedia and Pascal sentence are illustrated in Fig. 3 and Fig. 4. Obviously, M-RIS performs better than RIS in almost all classes. Therefore, the proposed mid-level feature space learning method is indeed effective and efficient on cross-media regularization.

5. CONCLUSION

In this paper, we present a cross-media regularization framework to enhance image understanding. Within the framework, a mid-level feature space learning approach is proposed to transfer the discriminative characteristic embedded in the textual modality into their corresponding visual modality. Specifically, images and text documents are firstly projected into the mid-level feature space from the low-level feature space, and then high-level semantic features can be calculated from the mid-level feature space to implement cross-media regularization. The experiments on two commonly used datasets demonstrate that introducing the mid-level space learning remarkably improves the effectiveness of cross-media regularization. In the future, our focus will be on the application of the proposed framework for the scenarios of multi-modal by exploring the correlations among them.

6. ACKNOWLEDGMENT

This work was supported in part by National Basic Research Program of China (No.2012CB316400), National Natural Science Foundation of China (No.61025013, No.61172129, No.61202241), Program for Changjiang Scholars and Innovative Research Team in University (No.IRT201206), Program for New Century Excellent Talents in University (No.13-0661).

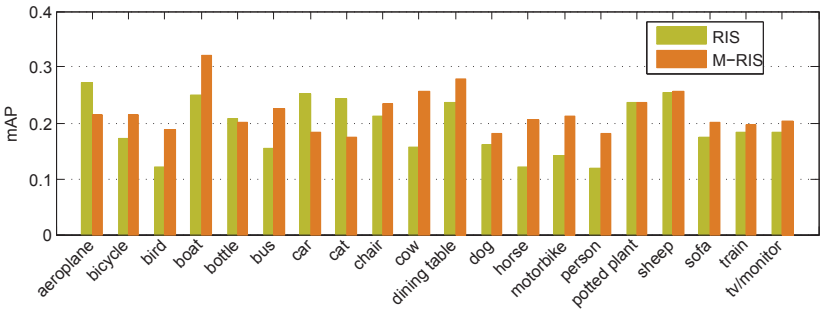

Fig. 4. mAP scores of each class on Pascal sentence dataset.

7. REFERENCES

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

[2] Navneet Dalal and Bill Triggs, "Histograms of oriented gradients for human detection," in *CVPR*, 2005, vol. 1, pp. 886–893.

[3] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories," *CVIU*, vol. 106, no. 1, pp. 59–70, 2007.

[4] S.K. Wei, Y. Zhao, Z.F. Zhu, and N. Liu, "Multimodal fusion for video search reranking," *TKDE*, vol. 22, no. 8, pp. 1191–1199, 2010.

[5] Yue-Ting Zhuang, Yi Yang, and Fei Wu, "Mining semantic correlation of heterogeneous multimedia data for cross-media retrieval," *TMM*, vol. 10, no. 2, pp. 221–229, 2008.

[6] Yi Yang, Dong Xu, Feiping Nie, Jiebo Luo, and Yueting Zhuang, "Ranking with local regression and global alignment for cross media retrieval," in *MM*, 2009, pp. 175–184.

[7] G.J. Qi, C. Aggarwal, and T. Huang, "Towards semantic knowledge propagation from text corpus to web images," in *WWW*, 2011, pp. 297–306.

[8] Y. Zhu, Y. Chen, Z. Lu, S.J. Pan, G.R. Xue, Y. Yu, and Q. Yang, "Heterogeneous transfer learning for image classification," in *AAAI*, 2011.

[9] J.C. Pereira and N. Vasconcelos, "On the regularization of image semantics by modal expansion," in *CVPR*, 2012, pp. 3093–3099.

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

[11] Hervé Abdi, "Partial least square regression (pls regression)," *Encyclopedia for research methods for the social sciences*, pp. 792–795, 2003.

[12] R.O. Duda, P.E. Hart, and D.G. Stork, "Pattern classification. 2nd," *Edition. New York*, 2001.

[13] Jon R Kettenring, "Canonical analysis of several sets of variables," *Biometrika*, vol. 58, no. 3, pp. 433–451, 1971.

[14] Haifeng Li, Tao Jiang, and Keshu Zhang, "Efficient and robust feature extraction by maximum margin criterion," *TNN*, vol. 17, no. 1, pp. 157–165, 2006.

[15] Y.N. Chen and H.T. Lin, "Feature-aware label space dimension reduction for multi-label classification," in *NIPS*, 2012, pp. 1538–1546.

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

[17] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier, "Collecting image annotations using amazon's mechanical turk," in *Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, 2010, pp. 139–147.

[18] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python*, O'Reilly Media, Incorporated, 2009.

[19] R.E. Fan, K.W. Chang, C.J. Hsieh, X.R. Wang, and C.J. Lin, "Liblinear: A library for large linear classification," *JMLR*, vol. 9, pp. 1871–1874, 2008.

[20] Lindsay I Smith, "A tutorial on principal components analysis," *Cornell University, USA*, vol. 51, pp. 52, 2002.