COUPLED FEATURE MAPPING AND CORRELATION MINING FOR CROSS-MEDIA RETRIEVAL

Mengdi Fan, Wenmin Wang, Ronggang Wang*

School of Electronic and Computer Engineering, Shenzhen Graduate School, Peking University
Lishui Road 2199, Nanshan District, Shenzhen, China 518055
fanmengdi@sz.pku.edu.cn, {wangwm, rgwang}@ece.pku.edu.cn

ABSTRACT

Cross-media retrieval aims to integrate and analyze the features of various modalities (e.g., text, image and video) to mine their potential semantic information. In this paper, we propose a novel cross-media retrieval framework, which performs coupled feature mapping and correlation mining successively. Our method first learns two projection matrices to map the multimodal features into a common category space, in which homo- and hetero-correlation techniques can be applied easily. Homo-correlation focuses on the semantic category information within the same media type, while hetero-correlation focuses on the semantic category information between different media types. The two could complement and reinforce each other. Experiments on two different datasets, Wikipedia dataset and Pascal Voc dataset, demonstrate that the proposed framework gives promising results compared to the related state-of-the-art approaches.

Index Terms— Cross-media retrieval, feature mapping, homo-correlation, hetero-correlation

1. INTRODUCTION

In the age of information, with the rise of Web 2.0 technology and the widespread use of the Internet, there has been an explosive growth of multimedia content including texts, images, audios and videos. The traditional content-based single media retrieval can not adapt to the complex Internet environment, especially for e-commerce, and meet the increasingly diverse user needs, such as finding the texts that can best match a given picture (e.g., to find some textual descriptions or user's evaluation of a 'dress') or searching pictures that can best illustrate a given text (e.g., to search some 'dress' pictures that can optimally match the query sentence). Therefore, the cross-media retrieval problem has been proposed, which comprises two tasks: 1) predicting text documents in response to an image query ("*Img2Text*"), and 2) predicting images in response to a text query ("*Text2Img*"). In general, media objects would be transformed into low-level feature vectors for representation, including the discrete sparse word count for texts and SIFT histogram for images, or even some high-level deep-net features [1][2]. No matter what kind of features are used, the common goal of this subject is to build bilateral semantic associations between images and texts [3]. Due to the heterogeneity and non-comparability among different media features, the first problem of cross-media retrieval is how to learn a unified representation space to make all data isomorphic and comparable. But the semantic gap between features and human understanding still exists. Thus, the second problem of how to measure the semantic similarity based on cross-modal features is considered. For the first problem, some dimensionality reduction approaches only consider the direct correlation between images and texts, ignoring the diversity of representations and the correlational structures hiding in the category information. For the second problem, the existing heterogeneous similarity measure can only consider the nearest neighbor of the same modality and has difficulties in discovering the semantic relationship across different modalities. What's more, if the features are very sparse or

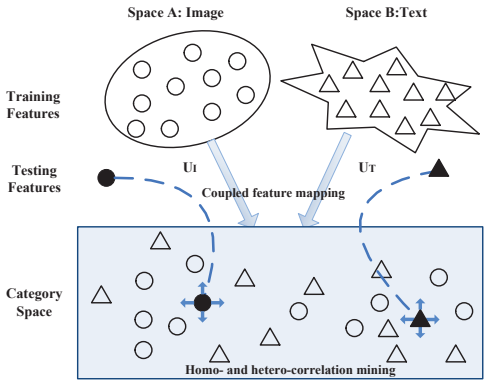

Fig. 1. The framework of our proposed method. All training features are mapped into the category space by coupled feature mapping. $\mathbf{U}_I$ and $\mathbf{U}_T$ are projection matrices. Testing features of the query samples are projected into the same category space by $\mathbf{U}_I$ or $\mathbf{U}_T$. Then perform the homo-correlation mining and the hetero-correlation mining respectively.

*Corresponding author

have too low or too high dimension, the performance of some method may not agree to its assumption.

In this paper, we propose an effective framework to address the above problems. In the training stage, we use coupled linear regression method to map all the training data from heterogeneous feature space to the common space defined by class labels (can also be called category space) and get the projection matrices for each modality respectively. In the testing stage, we first project the heterogeneous features of the query into the category space by projection matrices, then calculate the probability vectors of the query belonging to each category. The probability is obtained by analyzing the nearest neighbors of the same modality and the other. Finally we get the weighted category similarity matrices for image queries or text queries.

The main contributions of our work can be summarized as follows: 1) We propose a novel cross-media retrieval framework that evolves learning the category space based on coupled linear regression, and mining the homo- and hetero-correlations based on top-k nearest neighbors (KNN) simultaneously. The proposed framework has the ability to deal with the diversity of different features, and distinctly improve the cross-media retrieval efficiency. 2) We explore the correlation under category information, as we map the heterogeneous data of the coupled spaces into the common space defined by category labels. 3) We extend the HSNN method [4], which can only seek the semantic correlation within the same media type, to the case between the same or different media types under the circumstances of category space.

The rest of the paper is organized as follows. Section 2 overviews the related work of cross-media retrieval. Section 3 describes our proposed framework, including common space mapping based on coupled linear regression and homo- and hetero-correlation mining based on KNN. Section 4 presents the experimental results. Finally, we conclude the paper in section 5.

2. RELATED WORK

There are a few successful approaches to deal with cross-media retrieval problems currently. And their exploration of the improvement mainly focus on the following two aspects: subspace learning and feature comparison.

The goal of subspace learning is to learn a common latent space to jointly model images and its associated texts. Rasiwasia et al. [5] proposed to learn projections from the original feature spaces to a common semantic space in a supervised manner. Canonical Correlation Analysis (CCA) are applied to learn a common subspace by maximizing the correlation between images and texts. And CCA has also been developed to support more effective cross-modal image clustering for large-scale annotated image collections [6]. Chen et al. [7] pointed out that CCA and its variants may cause information dissipation when switching the modals, and they used the Partial Least Squares (PLS) to transform the image features to text space, then learned a semantic space to measure the similarity of two modalities. Wang et al. [8] proposed a generic minimization formulation by coupled linear regressions, l_{21}-norm and trace norm to achieve common subspace learning and coupled feature selection. Xu et al. [9] proposed a framework which firstly performed the coupled dictionary learning method to generate homogeneous sparse representations for different modalities, then used a coupled feature mapping scheme to project the derived sparse representations into a common subspace.

Feature comparison as a major component of the retrieval task aims to select the most similar results in response to the query by a similarity measure. Zhai et al. [4] proposed a heterogeneous similarity measure with nearest neighbors by computing the probability of two media objects belonging to the same semantic category. The method could directly compute the similarity between media objects of different types. And they also proposed a novel correlation propagation approach to simultaneously deal with positive correlation and negative correlation between media objects of different modalities in [10, 11]. Zhuang et al. [12] proposed a method of transductive learning to mine the semantic correlations in different modalities. According to the co-occurrence information of media objects, they constructed a uniform cross-media correlation graph and assigned a positive score to perform cross-media retrieval. And the relevance feedback is required to boost the performance.

We can learn from these methods that both subspace learning and feature comparison are crucial for cross-media retrieval problems. Subspace learning method is mapping the data of different modalities into a common latent space. In the process of subspace learning, the diversity of feature representations and the necessity of category information are usually neglected. For feature comparison methods, the correlations within each media type and the correlations between different media types are rarely associated with each other. Accordingly, we aim to combine subspace learning and feature comparison through coupled feature mapping and homo- and hetero-correlation mining for cross-media retrieval problem. The details of our proposed framework will be described in section 3.

3. PROPOSED METHOD

In this section, we present our proposed framework for cross-media retrieval as shown in Fig.1, which contains two procedures: Coupled feature mapping, and Homo- and Hetero-correlation mining.

3.1. Coupled feature mapping

We define the multimedia dataset as $D = \{D_1, D_2, \ldots, D_n\}$, in which $D_i = (\mathbf{X}_i^I, \mathbf{X}_i^T)$ denotes the original features from

two modalities: d^I dimensional image features and d^T dimensional text features. Here, we denote data from image modality as $\mathbf{X}_I = [\mathbf{X}_1^I, \mathbf{X}_2^I, \ldots, \mathbf{X}_n^I]^\top \in \mathbb{R}^{n \times d^I}$ and data from text modality as $\mathbf{X}_T = [\mathbf{X}_1^T, \mathbf{X}_2^T, \ldots, \mathbf{X}_n^T]^\top \in \mathbb{R}^{n \times d^T}$ respectively, where n stands for the number of the samples. Let $\mathbf{L} = [\mathbf{L}_1, \mathbf{L}_2, \ldots, \mathbf{L}_n]^\top \in \mathbb{R}^{n \times c}$ be the class label matrix, where c is the number of classes.

Coupled feature mapping aims to learn two projection matrices $\mathbf{U}_I \in \mathbb{R}^{d^I \times c}$ and $\mathbf{U}_T \in \mathbb{R}^{d^T \times c}$ to map the data of the coupled feature spaces into the common space defined by class labels, which can also be called the "category space". Inspired by [8, 9], we adopt the ridge regression method to minimize the errors of projecting the original representations of each modality to the category space. That is, the generic minimization problem is derived in the following form:

$$\begin{aligned} \min_{\mathbf{U}_I, \mathbf{U}_T} & \left\|\mathbf{X}_I\mathbf{U}_I - \mathbf{L}\right\|_F^2 + \left\|\mathbf{X}_T\mathbf{U}_T - \mathbf{L}\right\|_F^2 \\ & + \lambda\left(\left\|\mathbf{U}_I\right\|_F^2 + \left\|\mathbf{U}_T\right\|_F^2\right) \end{aligned} \tag{1}$$

The solutions of $\mathbf{U}_I$ and $\mathbf{U}_T$ in (1) can be derived as:

$$\begin{aligned} \mathbf{U}_I &= \left(\mathbf{X}_I^\top\mathbf{X}_I + \lambda I\right)^{-1}\mathbf{X}_I^\top\mathbf{L} \\ \mathbf{U}_T &= \left(\mathbf{X}_T^\top\mathbf{X}_T + \lambda I\right)^{-1}\mathbf{X}_T^\top\mathbf{L} \end{aligned} \tag{2}$$

where I is the unitary matrix. Then, we use $\mathbf{U}_I$ and $\mathbf{U}_T$ to map the original features of images and texts into the category space,

$$\mathbf{I} = \mathbf{X}_I\mathbf{U}_I \qquad \mathbf{T} = \mathbf{X}_T\mathbf{U}_T \tag{3}$$

where $\mathbf{I} = [\mathbf{I}_1, \mathbf{I}_2, \ldots, \mathbf{I}_n]^\top \in \mathbb{R}^{n \times c}$ and $\mathbf{T} = [\mathbf{T}_1, \mathbf{T}_2, \ldots, \mathbf{T}_n]^\top \in \mathbb{R}^{n \times c}$ are represented for the isomorphic feature matrices of images and texts respectively.

3.2. Homo- and Hetero-correlation mining

Homo-correlation focuses on mining the semantic category information within each media type. For one testing feature point $\mathbf{I}_i \in \mathbb{R}^c$ (or $\mathbf{T}_i \in \mathbb{R}^c$) obtained in the above process, we firstly use KNN classification to find k-nearest training samples of the same modality in the category space defined by labels. The probability of an image query $\mathbf{I}_i$ belonging to category c is defined as follows:

$$p_{i,c} = p(L_i = c|\mathbf{I}_i) = \frac{\sum\limits_{\mathbf{I}_k \in KNN(\mathbf{I}_i) \wedge L_k = c} \sigma(sim(\mathbf{I}_i, \mathbf{I}_k))}{\sum\limits_{\mathbf{I}_k \in KNN(\mathbf{I}_i)} \sigma(sim(\mathbf{I}_i, \mathbf{I}_k))} \tag{4}$$

$\mathbf{I}_k \in KNN(\mathbf{I}_i)$ stands for searching k-nearest images of $\mathbf{I}_i$ in training set, each of which is represented as $\mathbf{I}_k$. $L_k = c$ means the class label of image $\mathbf{I}_k$ equals to c. $\sigma(z) = (1 + exp(-z))^{-1}$ is the sigmoid function, and $sim(\mathbf{I}_i, \mathbf{I}_k)$ is the similarity measure between two data points.

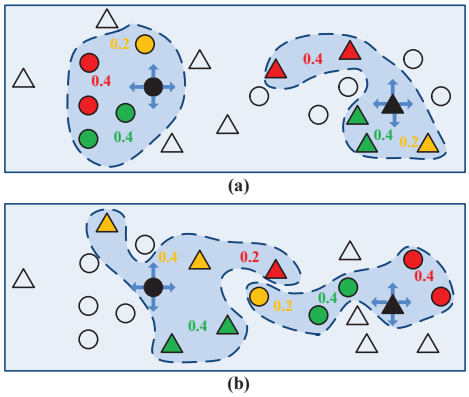

Fig. 2. The illustration of homo- and hetero-correlation mining. (a) homo-correlation mining, aiming at finding 5-nearest neighbors of the same modality which belong to three semantic categories. (b) hetero-correlation mining, aiming at finding 5-nearest neighbors of the other modality. The black circle and triangle represent the image query and text query respectively. The color of yellow, red and green represent three categories. The category similarity matrices of image and text queries are represented as $\mathbf{S}_I = t1 \cdot [0.2, 0.4, 0.4] \cdot [0.2, 0.4, 0.4]^\top + t2 \cdot [0.4, 0.2, 0.4] \cdot [0.2, 0.4, 0.4]^\top$ and $\mathbf{S}_T = t1 \cdot [0.2, 0.4, 0.4] \cdot [0.2, 0.4, 0.4]^\top + t2 \cdot [0.2, 0.4, 0.4] \cdot [0.4, 0.2, 0.4]^\top$ respectively.

As the heterogeneous features of different media types have been projected into the common category space in the procedure of coupled feature mapping, so we can measure the similarity across different media types. Hetero-correlation focuses on mining the semantic category information between media types. For the same image query $\mathbf{I}_i$, the Eq.4 can be rewritten as:

$$p\prime_{i,c} = p\prime(L_i = c|\mathbf{I}_i) = \frac{\sum\limits_{\mathbf{T}_k \in KNN(\mathbf{I}_i) \wedge L_k = c} \sigma(sim(\mathbf{I}_i, \mathbf{T}_k))}{\sum\limits_{\mathbf{T}_k \in KNN(\mathbf{I}_i)} \sigma(sim(\mathbf{I}_i, \mathbf{T}_k))} \tag{5}$$

$\mathbf{T}_k \in KNN(\mathbf{I}_i)$ stands for searching k-nearest texts of $\mathbf{I}_i$ in training set, each of which is represented as $\mathbf{T}_k$. $L_k = c$ means the class label of text $\mathbf{T}_k$ equals to c. The illustration of homo- and hetero-correlation mining is explicitly shown in Fig.2.

$\mathbf{P}_{homo}^I \in \mathbb{R}^{m \times c}$ and $\mathbf{P}_{hetero}^I \in \mathbb{R}^{m \times c}$ represent the homo- and hetero-correlation matrix of m testing images respectively:

$$\mathbf{P}_{homo}^I = \begin{pmatrix} p_{1,1} & \cdots & p_{1,c} \\ \vdots & \ddots & \vdots \\ p_{m,1} & \cdots & p_{m,c} \end{pmatrix} \mathbf{P}_{hetero}^I = \begin{pmatrix} p\prime_{1,1} & \cdots & p\prime_{1,c} \\ \vdots & \ddots & \vdots \\ p\prime_{m,1} & \cdots & p\prime_{m,c} \end{pmatrix} \tag{6}$$

The definition of $\mathbf{P}^T_{homo}$ and $\mathbf{P}^T_{hetero}$ can be obtained similarly. Actually, both the homo- and hetero-correlation are important, and their fusion can complement each other. For *Img2Text* and *Text2Img*, the category similarity matrices $\mathbf{S}_I$, $\mathbf{S}_T \in \mathbb{R}^{m\times m}$ can be computed respectively:

$$\begin{aligned}\mathbf{S}_I &= t1 \cdot \mathbf{P}^I_{homo} \cdot \mathbf{P}^{T\ \top}_{homo} + t2 \cdot \mathbf{P}^I_{hetero} \cdot \mathbf{P}^{T\ \top}_{hetero} \\ \mathbf{S}_T &= t1 \cdot \mathbf{P}^T_{homo} \cdot \mathbf{P}^{I\ \top}_{homo} + t2 \cdot \mathbf{P}^T_{hetero} \cdot \mathbf{P}^{I\ \top}_{hetero} \\ &s.t. \quad t1 + t2 = 1\end{aligned} \tag{7}$$

where $t1$ and $t2$ are empirical weights in accordance with the experimental results.

Algorithm 1 summarizes the procedure of our proposed framework.

Algorithm 1 Algorithm of our proposed framework

Input: $\mathbf{X}_I \in \mathbb{R}^{n\times d^I}$, $\mathbf{X}_T \in \mathbb{R}^{n\times d^T}$ and $\mathbf{L} \in \mathbb{R}^{n\times c}$

1: Compute $\mathbf{U}_I, \mathbf{U}_T$ according to Eq.2.
2: Use $\mathbf{U}_I, \mathbf{U}_T$ to map $\mathbf{X}_I, \mathbf{X}_T$ into the category space $\mathbf{I}, \mathbf{T}$ in Eq.3.
3: For m image queries $\mathbf{I}_i \in \mathbf{I}$ or text queries $\mathbf{T}_j \in \mathbf{T}$ $(i, j = 1, \ldots, m)$, calculate homo-correlation $p_{i,c}$, $p_{j,c}$ and hetero-correlation $p\prime_{i,c}$, $p\prime_{j,c}$ of each sample.
4: Compute $\mathbf{S}_I$ and $\mathbf{S}_T$ according to Eq.6∼7. $\mathbf{S}_I$ and $\mathbf{S}_T$ are the weighted category similarity matrices which mean assigning $\mathbf{I}_i$ and $\mathbf{T}_j$ to the same semantic category.

Output: $\mathbf{S}_I \in \mathbb{R}^{m\times m}$ and $\mathbf{S}_T \in \mathbb{R}^{m\times m}$

4. EXPERIMENTAL RESULTS

4.1. Experimental setting

In this section, we apply the proposed method to two cross-media retrieval tasks: *Img2Text* and *Text2Img*. Given an image (or text) query, the goal is to find the nearest neighbors from text (or image) database. We compare with the state-of-the-art methods, CCA [5], LCFS [8], CDLFA (Advanced) [9] and HSNN [4] on two publicly available datasets: Wikipedia dataset [5] and Pascal Voc dataset [13]. All of the compared cross-media retrieval methods adopt the same features, parameters and experimental conditions in training and testing stages for fair comparison purpose.

The mean average precision (MAP) score is used to evaluate the performance of the schemes. $\mathbf{S}_I \in \mathbb{R}^{m\times m}$ and $\mathbf{S}_T \in \mathbb{R}^{m\times m}$ are sorted in descending order. The *average precision* (AP) [14] of N retrieved target objects is defined as

$$AP = \frac{1}{T}\sum_{j=1}^{N} Prec(j)\delta(j) \tag{8}$$

where T is the number of relevant objects in the retrieved set, $Prec(j)$ is the precision of the top j retrieved objects and $\delta(j)$ is an indicator function equaling to 1 if the jth retrieved object belongs to the same class of the query, zero otherwise. The MAP score is calculated by averaging the AP values from all the queries in the query set. Besides, to evaluate the performance pictorially, *Precision-Recall* curves are also adopted in all the approaches.

For our method, the parameter k in KNN classification is empirically set to 80. The setting of weights $t1$ and $t2$ are determined by their performance curves. For the trade-off parameter λ, we referred to the implementation in [9]. We choose such as [10, 1, 0.1, 0,01, 0.001 ...] for validation and finally select λ=0.1 for Wikipedia dataset and λ=1 for Pascal Voc dataset with considerable performance.

4.2. Results on Wikipedia dataset

We first evaluate the proposed approach on Wikipedia dataset[1], which is generated from Wikipedia's featured articles. The dataset consists of 2866 text-image pairs, annotated with a label from 10 most populated semantic classes. A random split was used to produce a training set of 2173 documents and 693 documents according to [5]. The representation of the text with 10 dimensions is derived from a latent Dirichlet allocation (LDA) model. And each image is represented by the 128 dimensional SIFT descriptor histograms.

Table 1. MAP scores on Wikipedia dataset

Methods	Img2Text	Text2Img	Average
CCA	0.2449	0.1929	0.2189
LCFS	0.2340	0.2122	0.2231
CDLFA	0.2628	0.2335	0.2482
HSNN[2]	0.2839	0.2018	0.2429
Proposed (Coupled)	0.2885	0.2182	0.2534
Proposed (Chi)	0.3240	0.2374	0.2807
Proposed (NC)	0.3235	0.2372	0.2804
Proposed (CC)	0.3240	0.2374	0.2807
Proposed (HI)	**0.3249**	**0.2374**	**0.2812**

Table 1 shows MAP scores of different methods on Wikipedia dataset. And the performance of the proposed method without the correlation mining part is reported in Proposed (Coupled). Besides, we explored the effect of different similarity measures in Equation 4 and 5, such as Chi-square distance (Chi), Normalized Correlation (NC), Centered Correlation (CC) and Histogram Intersection (HI). We can see the MAP scores without the correlation mining part are 0.2885 (*Img2Text*) and 0.2182 (*Text2Img*). And the optimal MAP scores 0.3249 (*Img2Text*) and 0.2374 (*Text2Img*) are achieved by using Histogram Intersection measurement.

[1] http://www.svcl.ucsd.edu/projects/crossmodal/

[2] As the author did not provide the source code of HSNN, we implemented it without the step of AdaRank.

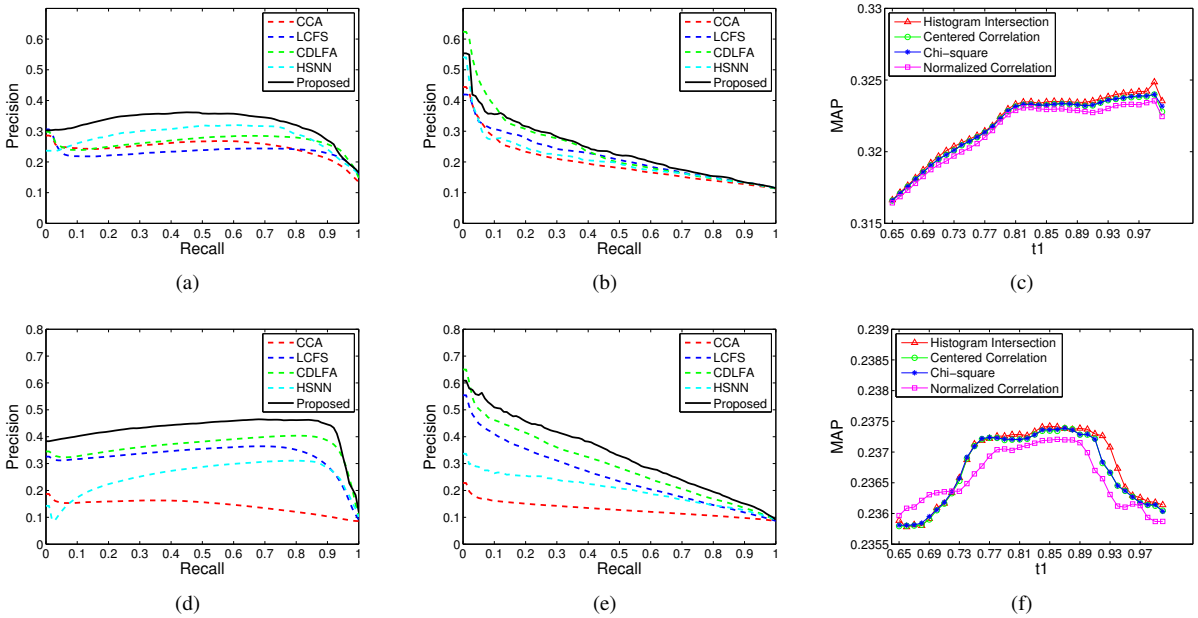

Fig. 3. Precision recall curves: (a) *Img2Text* on Wikipedia dataset. (b) *Text2Img* on Wikipedia dataset. (d) *Img2Text* on Pascal Voc dataset. (e) *Text2Img* on Pascal Voc dataset. Performance curves of $t1$: (c) *Img2Text* on Wikipedia dataset. (f) *Text2Img* on Wikipedia dataset.

We can see that our methods performs better than HSNN both in *Img2Text* and *Text2Img*. This is because HSNN only finds the nearest neighbors of the same modality, and our method integrates both hetero-correlation and homo-correlation altogether. The performance curves of $t1$ in Fig.3(c) and 3(f) show the contribution of explored cross modality correlation. $t1$ equals to 1 means only homo-correlation takes effect and $t1$ equals to 0 vice versa. The MAP of *img2Text* increases to maximum when $t1$ equals to 0.99. The MAP of *Text2Img* first increases, then keeps optimal results in a relatively wide range and finally decreases. It means hetero-correlation mining enhances the performance of *Text2Img*, which is a rather difficult issue in cross-media retrieval. For the representative subspace learning methods, such as CCA and LCFS, the performance improvements of our method are clear seen. The method CDLFA improves the performance of *Text2Img* based on LCFS by transforming the raw features into sparse representations, and our method still achieves better than it.

Comparing the PR curves in Fig.3(a), we can see that our method attains higher precision at almost all levels of recall for *Img2Text*. For text queries in Fig.3(b), the improvement is not substantial at lower levels of recall compared with CDLFA, but is noticeable at higher levels of recall.

Fig.4 shows two examples of text queries and the top five images retrieved by our method. It can be observed that our method finds the closet matches at semantic level. The first text query belongs to "Warfare" category and the five retrieved images belong to the category of "Warfare", "Warfare", "History", "Geography"&"places" and "Warfare" respectively. The second text query belongs to "Biology" category and the five retrieved images belong to the category of "Biology", "Biology", "History", "Warfare" and "Biology" respectively. The fact that "History" and "Geography&places" are misclassified as "Warfare" is not surprising, since these three categories share similar words and imagery. But there are still some unreasonable misclassifications, such as confusing "Biology" and "Warfare" with each other.

4.3. Results on Pascal Voc dataset

We then perform comparison on Pascal Voc dataset[3][13]. The dataset consists of 5011 training and 4952 testing image-tag pairs, which can be categorized into 20 different classes. The image features are 512-dimensional Gist features, and the text features are 399-dimensional word frequency features. As some images are multi-labeled, we select 2808 training and 2841 testing pairs with only one object each image. Since the Pascal Voc dataset has higher dimensional features than Wikipedia dataset, and the text features are sparse, so the performance of various methods will be entirely different.

Table 2 shows the MAP scores of different methods on Pascal Voc dataset. We can see our proposed method with centered correlation outperforms the others for *Img2text*, and Histogram Intersection achieves the best results for *Text2Img*. The original CCA method is almost invalid due to many redundant and irrelevant features. And the HSNN method has

[3]https://github.com/peaceful-lukas/BMVC2010/tree/master/data

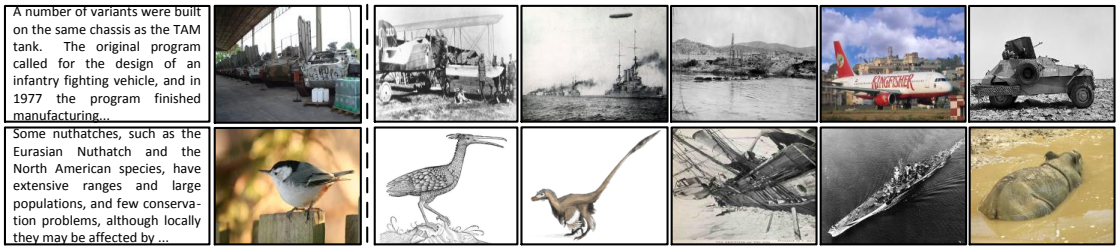

Fig. 4. Two examples of *Text2Img*. The first column contains the text queries. The second column represents the paired images of the text queries. Column 3-7 are the retrieved images of our method on Wikipedia dataset.

not worked so well in original heterogeneous space. It shows that the procedure of coupled feature mapping is necessary, which not only maps the heterogeneous features into a common category space but also provides the ways to find cross-modal correlations of different media types. Compared with the LCFS and CDLFA method, we can see that the retrieval performance of our correlation mining method is more effective than their feature selection methods.

Table 2. MAP scores on Pascal Voc dataset

Methods	Img2Text	Text2Img	Average
CCA	0.1426	0.1297	0.1362
LCFS	0.3438	0.2674	0.3056
CDLFA	0.3741	0.2944	0.3342
HSNN[4]	0.2535	0.2021	0.2278
Proposed (Coupled)	0.2907	0.2768	0.2838
Proposed (HI)	0.3595	**0.3302**	0.3449
Proposed (Chi)	0.4265	0.3290	0.3778
Proposed (CC)	**0.4266**	0.3291	**0.3779**

The corresponding PR curves are plotted in Fig.3(d) and 3(e). It is clear that our method outperforms other methods at all levels of recall for both forms of cross-media retrieval.

5. CONCLUSION

In this paper, we have proposed a framework to solve cross-media retrieval problem. Firstly, the features of different modalities are projected into a common subspace defined by label information. Then, for each query, we calculate the category probabilities by analyzing the nearest neighbors of the same modality and the other. Our framework integrates coupled feature mapping and homo- and hetero-correlation mining subtly, and achieves better performance over related state-of-the-art methods on two public challenging datasets. In the future, we will jointly model other media types, such as video and audio, and further improve the accuracy of *Text2Img*.

[4]As [4] did not test on Pascal Voc dataset, so we use the results of our implementation for consistency's sake.

6. ACKNOWLEDGEMENT

This project was supported by Shenzhen Peacock Plan.

7. REFERENCES

[1] A. Frome, G. Corrado, J. Shlens, et al., "Devise: A deep visual-semantic embedding model," in *NIPS*, 2013.

[2] J. Dong, X. Li, S. Liao, J. Xu, D. Xu, and X. Du, "Image retrieval by cross-media relevance fusion," in *ACM MM*, 2015.

[3] Y. Verma and CV. Jawahar, "Im2text and text2im: Associating images and texts for cross-modal retrieval," in *BMVC*, 2014.

[4] X. Zhai, Y. Peng, and J. Xiao, "Effective heterogeneous similarity measure with nearest neighbors for cross-media retrieval," in *MMM*, 2012.

[5] N. Rasiwasia, J. Costa Pereira, E. Coviello, et al., "A new approach to cross-modal multimedia retrieval," in *ACM MM*, 2010.

[6] C. Jin, W. Mao, R. Zhang, Y. Zhang, and X. Xue, "Cross-modal image clustering via canonical correlation analysis," in *AAAI*, 2015.

[7] Y. Chen, L. Wang, W. Wang, and Z. Zhang, "Continuum regression for cross-modal multimedia retrieval," in *ICIP*, 2012.

[8] K. Wang, R. He, W. Wang, L. Wang, and T. Tan, "Learning coupled feature spaces for cross-modal matching," in *ICCV*, 2013.

[9] X. Xu, A. Shimada, R. Taniguchi, and L. He, "Coupled dictionary learning and feature mapping for cross-modal retrieval," in *ICME*, 2015.

[10] X. Zhai, Y. Peng, and J. Xiao, "Cross-modality correlation propagation for cross-media retrieval," in *ICASSP*, 2012.

[11] X. Zhai, Y. Peng, and J. Xiao, "Cross-media retrieval by intra-media and inter-media correlation mining," *Multimedia systems*, 2013.

[12] Y. Zhuang, Y. Yang, and F. Wu, "Mining semantic correlation of heterogeneous multimedia data for cross-media retrieval," *TMM*, 2008.

[13] S. Hwang and K. Grauman, "Reading between the lines: object localization using implicit cues from image tags.," *TPAMI*, 2012.

[14] X. Lu, F. Wu, S. Tang, et al., "A low rank structural large margin method for cross-modal ranking," in *SIGIR*, 2013.