Semi-supervised Distance Consistent Cross-modal Retrieval

Xiao Dong[†], En Yu[†], Min Gao[†], Lei Zhu[‡], Jiande Sun[†*], Huaxiang Zhang[†*]

[‡]The University of Queensland, Brisbane, Australia
[†]Shandong Normal University, Jinan, China
dx.icandoit@gmail.com,{sdnu_enyu,mingao2017}@hotmail.com,leizhu0608@gmail.com,jiandesun@hotmail.com,
huaxzhang@163.com

ABSTRACT

Most of existing cross-modal retrieval approaches only exploit labeled data to train coupled projection matrices for supporting retrieval tasks across heterogeneous modalities. However, the valuable information involved in unlabeled data is unfortunately ignored. In this paper, we propose a novel Semi-Supervised Distance Consistent method (SSDC) to solve the problem. Our approach firstly models the initial correlation between different modalities by constructing the pseudo label and corresponding data of unlabeled query. Then our method learns projection matrices by adaptively optimizing the pseudo label of unlabeled data. In this way, SSDC could learn discriminative projection matrices. Experimental results on two publicly available datasets demonstrate the superior performance of the proposed approach.

KEYWORDS

Cross-modal retrieval, semi-supervised, pseudo label

1 INTRODUCTION

With the rapid growth of multimedia data, content-based multimedia retrieval has become increasingly important [1–3] . In multimedia representation, the data can be characterized with multiple modalities that describe the same semantic information. For instance, when we look at a scenic place, we can write a paragraph words, take pictures or shoot a video to record. To facilitate multimedia understanding, it is essential to explore the connection of different modalities describing the same sematnic information. However, different modalities are represented with features located in space with different dimensions. Thus, the similarity between low-level features and high-level semantics among different modalities can't be directly measured. This challenge is called as semantic-gap [4, 5] . To address this problem, many methods have been proposed. One kind of the important and effective methods are based on subspace learning. In this paper, we take cross-modal retrieval task between image and text as an example to introduce our solution. Note that our method can be easily extended when more modalities are involved.

As mentioned above, cross-modal retrieval is designed to address semantic-gaps. The subspace learning methods aim at learning a latent semantic space, where the similarity of heterogeneous data can be measured. Based on the subspace learning, traditional methods learn a couple of projection matrices, and project different modalities into a common latent semantic space, so that the similarity can be measured directly. For instance, Canonical Correlation Analysis (CCA) [6, 7] intends to project different modal features into a common latent space that maximizes the correlation between heterogeneous data [8, 9] . In addition, based on CCA, Semantic Correlation Match (SCM) [6] is proposed to learn the semantic subspace by logistic regression [10, 11]. Another approaches are Partial Least Squares (PLS) [12–17] . It learns two latent spaces by maximizing the correlations among heterogeneous data. GMLDA [18] and GMMFA [18] based on generalized multi-view analysis (GMA) [19–22] obtain better performance by using tags to extract multi-view features. However, the cross-modal retrieval task includes multiple sub-tasks. For example, giving an image to retrieve relevant texts (I2T) and giving a text description to retrieve relevant images (T2I). Therefore, learning a couple of projection matrices ignores the importance of query on different sub-tasks (e.g. I2T and T2I) when learning the projection matrices. Specifically, for the I2T task, images may play a more important role when learning projection. Similarly, text may be more essential in T2I task (Typical I2T and T2I tasks are shown in Fig. 1). Therefore, as analyzed above, existing projection matrix learning methods still can not achieve satisfactory performance.

In order to consider the semantic association of query data on different sub-tasks, Wei et al. propose a method named modality-dependent cross-media retrieval method (MDCR) [23] . It is designed to learn two couples of projection matrices for different retrieval sub-tasks. This method considers the importance of query on different sub-tasks when learning the projection matrices. Hence, the performance on different datasets can achieve certain improvement. However, MDCR is a supervised method using only labeled data for training. It fails to exploit the potential valuable information of unlabeled data. To address this limitation, in this paper, we combine labeled data with unlabeled data to improve the performance. In order to make full use of the unlabeled data, our method assigns a pseudo label for each unlabeled datum, then pseudo labels are fitting to the true label in the optimization by using the correspondence between similarities and labels. We propose a novel semi-supervised cross-modal retrieval method. Specifically, in the

[*] Corresponding author

VSCC'17, October 23, 2017, Mountain View, CA, USA.
© 2017 ACM. ISBN 978-1-4503-5506-3/17/10…$15.00
DOI: https://doi.org/10.1145/3132734.3132735

A boy doing a wheelie on a plank with the beach in the background.

A man is doing a wheelie on a bike.

A man on a bicycle doing a wheelie at the end of a diving board.

A man wearing a bike helmet does a trick on his dirt bike near a beach.

Person on bike popping a wheelie from a rooftop

......

(a)Image query

A man knees on the ground while talking on his cell phone next to a bike and car.

A man on a cell phone kneels on the sidewalk next to a red bike to write something down.

A man talking on his cellphone is crouched on the sidewalk near a bike and a car.

Beside a pot of yellow flowers, a man is kneeling on a sidewalk using a cell phone and writing on a piece of paper.

Man kneels on sidewalk and takes notes while talking on a cell phone.

......

(b)Text query

Figure 1: Typical examples of two sub-tasks in cross-modal retrieval: (a) I2T: using image to retrieve texts. (b) T2I: using text to retrieve images.

I2T, we first project all image features into the text feature space and then calculate the class-centers of all the labeled classes in different modalities. At the same time, the weights of unlabeled images and text class-centers are learned. The pseudo labels of unlabeled images are generated by optimizing the weight matrix. Finally, we learn the projection matrices by regarding all the original training data and new unlabeled images as training set. The basic framework of proposed method is shown in Fig. 2 . The contributions of this paper can be summarized as follows:

(1) The SSDC approach is proposed to combine both labeled data and unlabeled data to learn two couples of projection matrices. It specially considers the importance of different queries when learning the projection matrices for two different sub-tasks.

(2) We propose a novel semi-supervised method to assign a pseudo label for each unlabeled datum. The automatically generated pseudo labels are further optimized to gradually match their semantics with the labeled data.

(3) The proposed method effectively learns the structural relations of different modalities by projecting one modality data into another.

The rest of the paper is structured as follows. In Section 2, the related methods of cross-modal retrieval are introduced. In Section 3, our method and an iterative algorithm are showed in detail. Comparative analysis of experimental results is presented on two popular datasets in Section 4. The conclusion is presented in Section 5.

2 RELATED WORK

Recently, some promising methods are proposed for cross-modal retrieval. The basic idea of them is jointly modeling different modalities. Most of the methods learn a common latent subspace for different modalities. Representations of different modalities in this common subspace are compared directly for cross-modal retrieval. As one of possible solutions, Canonical Correlation Analysis(CCA) [6] projects two modalities to a shared latent space by maximizing the correlations between different modalities. Partial Least Squares (PLS) [12] builds the relation between the latent variables of different modalities for cross-modal retrieval. Based on the single multi-view analysis, Generalized Multiview Analysis (GMA) [18] is presented for feature extraction, and three-view CCA is proposed in [24] . The common idea of these methods is that they directly map the native feature space of each modality into the common latent space.

However, these aforementioned methods ignore the characteristics of different retrieval tasks. Therefore, in order to solve this problem, some methods based on two couples of projection matrices are proposed. Specifically, the main idea of these methods is learning two couples of projection matrices for two sub-tasks (I2T and T2I) respectively. They project the multi-modal data features into a latent semantic space, so that the similarity can be measured. The most typical representative method is Modality-Dependent Cross-media Retrieval (MDCR), which is proposed in [23] . However, it only exploits the labeled data for training, ignoring the important information of unlabeled data. In contrast, in this paper, our method not only investigates the significance of the query data when learning the projection matrices for two sub-tasks, but also comprehensively considers using information of both unlabeled and labeled data

3 SEMI-SUPERVISED DISTANCE CONSISTENT CROSS-MODAL RETRIEVAL

In this section, we describe the details of our proposed method. Fig. 2 shows the flowchart of the proposed method. It can be seen from Fig. 2 , our approach consists of three parts, which include obtaining class center, constructing pseudo labels and corresponding data, and training projection matrices.

3.1 Data Description

Let I=$[I_l; I_u] \in R^{n\times q}$ and T=$[T_l; T_u] \in R^{n\times p}$ denote all the image and text, respectively, where $I_l = [I_1, ..., I_k] \in R^{k\times q}$ and $I_u = [I_{k+1}, ..., I_n] \in R^{(n-k)\times q}$ represent labeled and unlabeled data separately in image modality, $T_l = [T_1, ..., T_k] \in R^{k\times p}$ and $T_u = [T_{k+1}, ..., T_n] \in R^{(n-k)\times p}$ represent labeled data and unlabeled data separately in text modality, q is the dimension of I and p is the dimension of T. $S = [S_l; S_u] \in R^{n\times c}$ is used to indicate the semantic matrix where S_i is represented by the one-hot code, and c is the number of classes.

3.2 Obtaining Class Center

In cross-modal retrieval, the similarities of different modalities can't be measured directly. Therefore, in order to correlate different modalities, our method learns the representation of the query data in the retrieved data space by learning a projection matrix. We utilize I2T as an example to learn the mapping matrix that is used to project images to texts, and the formula is given as:

$$\hat{W} = \arg\min \|I_l W - T_l\|_F^2 + \lambda\|W\|_F^2 \quad (1)$$

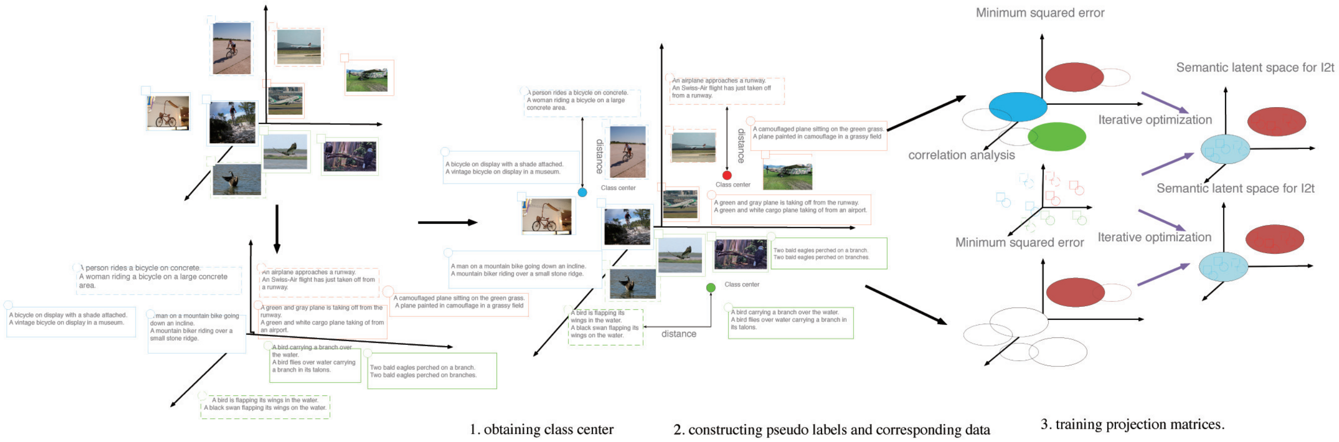

A person rides a bicycle on concrete.
A woman riding a bicycle on a large concrete area.
A bicycle on display with a shade attached.
A vintage bicycle on display in a museum.
A man on a mountain bike going down an incline.
A mountain biker riding over a small stone ridge.
An airplane approaches a runway.
An Swiss-Air flight has just taken off from a runway.
A green and gray plane is taking off from the runway.
A green and white cargo plane taking of from an airport.
A camouflaged plane sitting on the green grass.
A plane painted in camouflage in a grassy field
A bird carrying a branch over the water.
A bird flies over water carrying a branch in its talons.
Two bald eagles perched on a branch.
Two bald eagles perched on branches.
A bird is flapping its wings in the water.
A black swan flapping its wings on the water.
distance
Class center
distance
Class center
Class center
distance
Minimum squared error
correlation analysis
Semantic latent space for I2t
Iterative optimization
Semantic latent space for I2t
Minimum squared error
Iterative optimization
1. obtaining class center
2. constructing pseudo labels and corresponding data
3. training projection matrices.

Figure 2: The flowchart of our method.

where W is the projection matrix that makes I_l map into the T_l feature space. The representation of an image is obtained as

$$I_{l_p} = I_l \times \hat{W} \tag{2}$$

The jth class center of I_{l_p} and T_l is represented as $C_{I_{lpi}} = \frac{\sum_{i=1}^{n} \sum_{c=j} I_{lp_i}}{k_j}$ and $C_{Ti} = \frac{\sum_{i=1}^{n} \sum_{c=j} T_{l_i}}{k_j}$, where k_j is the labeled data number of the jth class. The class matrix $C_{I_{l_p}} \in R^{C \times p}$ and $C_T \in R^{C \times p}$ with $C_{I_{lpi}}$ and C_{Ti} are calculated by this way.

3.3 Constructing Pseudo Labels and Corresponding Data

In many cross-modal retrieval methods, unlabeled data are ignored in the training process, which may be helpful for the retrieval performance. In order to fully use the information of the unlabeled data, we calculate the center of each class based on the labeled data, and construct the pseudo labels for the unlabeled data.

I_{l_p} and T_l are in the same feature space, and the similarity $S_{u_{ji}}$ between each unlabeled data I_j and the ith class center C_{T_i} of the labeled data is measured as follows

$$S_{u_{ji}} = e^{\gamma \|I_{lp_j} - C_{T_i}\|_2^2} \tag{3}$$

where γ is the kernel coefficient.

Then we have the new similarity matrix $S = [S_l; S_u]$.

Furthermore, in order to obtain the corresponding pseudo text T_{u_p} data of I_u. For each I_{u_i}, the corresponding element $T_{u_{p_i}}$ of T_{u_p} is expressed as follows

$$T_{u_{p_i}} = C_T(\max_{location} S_{u_i}) \tag{4}$$

In this step, we use the projected class center to create the pseudo labels of I_u and corresponding data T_{u_p} of I_u. According to S_u, the class center data can be used to replace the unknown corresponding retrieval data of I_u. Therefore, our scheme above can bridge the semantic relevance between different modalities.

3.4 Training Projection Matrix

As described above, the best performance can’t be obtained if only one couple of projections are used to make the correlation maximized between different modals. Unlabeled data are more easy to be obtained than labeled ones. Therefore, our method uses I and new $T = [T_l; T_{u_p}]$ to achieve good performance.

Based on Section 3.2 and 3.3, we define the objective function (I2T) as

$$F_1(V, W) = \min_{V,W} \lambda \|IV^T - TW^T\|_F^2 + (1 - \lambda)\|IV^T - S\|_F^2 + \mu_1 \|V\|_F^2 + \mu_2 \|W\|_F^2 \tag{5}$$

Similarly, objective function(T2I) is defined as

$$F_2(V, W) = \min_{V,W} \lambda \|IV^T - TW^T\|_F^2 + (1 - \lambda)\|TV^T - S\|_F^2 + \mu_1 \|V\|_F^2 + \mu_2 \|W\|_F^2 \tag{6}$$

where V and W are projection matrices, λ is a balance coefficient, μ_1 and μ_2 are regularized coefficients.

In Eq. 5, $\|IV^T - TW^T\|_F^2$ is used to keep feature consistent in the common space, $\|IV^T - S\|_F^2$ is used to keep semantic consistent in the latent semantic space, and the rest regularized terms are used to prevent overfitting.

Eq. 5 is a nonconvex problem, but it is convex with respect to either V or W, while fixing the other. Therefore, the minimization over each one can be solved by the gradient descent.

The partial derivation of F_1 with respect to V and W are as follows:

$$\begin{aligned} \frac{\partial F_1}{\partial V} &= VI^T I + 2[\mu_1 V - \lambda W T^T I - (1 - \lambda) S^T I] \\ \frac{\partial f_1}{\partial W} &= 2[\mu_2 W + \lambda(W T^T T - V I^T T] \end{aligned} \tag{7}$$

Similarly, the partial derivation of F_2 with respect to V and W are as follows:

$$
\begin{aligned}
\frac{\partial f_2}{\partial V} &= 2[\mu_1 V + \lambda(VI^T I - WT^T I)] \\
\frac{\partial f_2}{\partial W} &= WT^T T + 2[\mu_2 W - \lambda V X^T T - (1-\lambda) S^T T]
\end{aligned}
\tag{8}
$$

The solution of V and W can be obtained by alternatively updating their values, and stop the process until a termination condition is met. Different from the previous methods, we take into account the semantic information of the unlabeled samples during the updating process of V and W. The learning process is described in Algorithm 1. Since the algorithm SSDC (T2I) for T2I is similar to SSDC (I2T), we don't repeat it.

Algorithm 1 SSDC(I2T)

```
Require:
    $I$ and $T$: Image-text pairs;
    $S$:semantic matrix;
    $\lambda$:balance coefficient;
    $\sigma$:learning rate;
    $\mu_1$ and $\mu_2$: regularized coefficient;
    $\epsilon_1$ and $\epsilon_2$:convergence coefficient;
Ensure:
    $V \in R^{c \times q}$ and $W \in R^{c \times p}$
    Initialize $\lambda$, $\sigma$, $\mu_1$ and $\mu_2$, $\epsilon_1$ and $\epsilon_2$, the entries of $S$ concerning
    the unlabeled samples is set to 0;
    repeat
        repeat
            $V_{i+1} = V_i - \sigma \frac{\partial f_1(V^i, W^j)}{\partial V^i}$
            i ← i+1
        until $F_1(V_i, W_j)$-$F_1(V_{i+1}, W_j)$<$\epsilon_1$
        repeat
            $W_{j+1} = W_j - \sigma \frac{\partial f_1(V_i, W_j)}{\partial W_j}$
            j ← j+1
        until $F_1(V_i, W_j)$-$F_1(V_i, W_{j+1})$<$\epsilon_1$
        $S_u = e^{-(I_u V^T - T_u W^T)}$;
        $St = (S_l, S_u)$
        t ← t+1
    until $F_1(V_i, W_j)$-$F_1(V_i, W_{j+1}$<$\epsilon_2$ or $t$ >maximal iteration number
    Output: Projection matrix $W$ and $V$.
```

By constructing the pseudo label of unlabeled data and iteratively updating the projection matrices, our method can adaptively adapt the projection distribution of the unlabeled data to the distribution of the labeled ones.

4 EXPERIMENT

In this section, we test SSDC on two popular datasets to show its performance on cross-modal retrieval.

4.1 Experimental Databases

Wikipedia [6] : This dataset has 2866 image-text pairs labeled by 10 topics. In Wikipedia, 2173 pairs of data are training samples and the rest are test samples. In our experiments, we use public dataset provided by [6] where images are represented by 128 dimensional bags of visual SIFT [25] features and texts are represented by 10 dimensional Latent Dirichlet Allocation (LDA) [26] .

Table 1: Retrieval performance (MAP %)

Wikipedia			
method	I2T	T2I	average
PLS	23.75	17.23	20.49
CCA	24.14	19.71	21.93
SM	22.64	21.84	22.24
SCM	26.62	22.57	24.59
GMMFA	23.09	20.34	21.72
GMLDA	24.64	19.52	22.08
MDCR	26.19	21.03	23.61
SSDC	**28.49**	**24.36**	**26.43**
Pascal Sentence			
method	I2T	T2I	average
PLS	36.53	37.63	37.08
CCA	37.99	37.20	37.59
SM	44.98	43.39	44.19
SCM	40.71	39.35	40.03
GMMFA	37.32	34.70	36.01
GMLDA	40.80	38.77	39.79
MDCR	43.22	46.22	44.72
SSDC	**44.50**	**48.91**	**46.71**

Pascal Sentence dataset [27] : This dataset consists of 1000 image-text pairs with 20 categories. We randomly choose 30 pairs from each category as training samples and the rest as test samples. Images are represented by 4096 dimensional CNN features [28, 29] , and texts are represented by 100 dimensional LDA [26] features.

4.2 Experimental Setting

SSDC is compared with PLS [12] , CCA [6] , SM [6] , SCM [6] , GMMFA [18] , GMLDA [18] and MDCR [23] , CCA [6] and PLS [12] are unsupervised learning methods, and the other methods are supervised learning methods. In PLS, the image feature is considered as input, while the text feature is considered as output. PLS projects the input variables into the space of output variables to directly compare the similarity of different modalities. SM and SCM use CCA to learn two maximally correlated subspaces to achieve retrieval task. Both GMMFA and GMLDA are based on the framework of GMA. MDCR is a double couple of mapping methods that use semantic information and pair-wise information.

In our experiment, the Mean Average Precision (MAP) [30] , the Precision-Recall curve and the Precision of each class are used to evaluate the performance of the different methods.

4.3 Experimental Results

The MAP values for different methods on Wikipedia and Pascal dataset are shown in Tab. 1.

As can be seen from the Tab. 1, supervised learning approaches (SM, SCM, GMMFA, GMLDA, MDCR) outperform unsupervised learning ones (PLS, CCA) . By exploiting the labelled data, the performance of cross-modal retrieval can be improved.

In addition, in the first step of SSDC, it projects the query data into the retrieval data space. We can find in T2I that our method has the best performance. This experimenal results indicate that the consistency of data dimension is an important factor in cross-modal retrieval.

Moreover, though the above supervised methods can achieve relatively good performance, they cannot exploit relation between unlabeled data features and class labels. Therefore, it may degrade the retrieval performance.

Fig. 3 and Fig. 4 show the precision-Recall curves and the precision of each class on Wikipedia and Pascal dataset. We find from Fig. 3 and Fig. 4 that, double couples of projection matrices are more competitive than one couple of projection matrices. To sum up, our method outperforms the compared baselines and achieves superior performance on both two databases.

5 CONCLUSIONS

This paper proposes a novel semi-supervised method for cross-modal retrieval. In our approach, in order to learn a common space, the pseudo label and pseudo data are generated to train two couples of projection matrices in an adaptive optimization. We take into account the full advantage valuable unlabeled data for better performance. The experimental results on two public databases show that SSDC can achieve superior performance compared with several competitive methods. Current work will be further extended following two directions. First, we intend to find a better solution to obtain more discriminative pseudo label and pseudo data. Second, we will integrate our approach into deep learning framework to further improve the retrieval results.

ACKNOWLEDGMENTS

The work is partially supported by the National Natural Science Foundation of China (61373081, 61572298, 61772322) , the Key Research and Development Foundation of Shandong Province (2016G GX101009) and the Natural Science Foundation of Shandong China (BS2014DX006, ZR2014FM012) . We also gratefully acknowledge the support of NVIDIA Corporation with the donation of the TITAN X GPU used for this research. The corresponding authors are Jiande Sun (jiandesun@hotmail.com) and Huaxiang Zhang (huaxzhang@163.com)

REFERENCES

[1] Chia-Hung Wei and Chang-Tsun Li. Content-based multimedia retrieval. *Proc. Encyclopedia of Multimedia Technology and Networking*, pages 116–122, 2005.

[2] Hugo Jair Escalante, Carlos A Hérnadez, Luis Enrique Sucar, and Manuel Montes. Late fusion of heterogeneous methods for multimedia image retrieval. In *the 1st ACM international conference on Multimedia information retrieval*, pages 172–179. ACM, 2008.

[3] Jiwoon Jeon, Victor Lavrenko, and Raghavan Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In *the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 119–126. ACM, 2003.

[4] Marc Ehrig. *Ontology alignment: bridging the semantic gap*, volume 4. Springer Science & Business Media, 2006.

[5] Fei Wu, Hong Zhang, and Yueting Zhuang. Learning semantic correlations for cross-media retrieval. In *Image Processing, 2006 IEEE International Conference on*, pages 1465–1468. IEEE, 2006.

[6] Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert RG Lanckriet, Roger Levy, and Nuno Vasconcelos. A new approach to cross-modal multimedia retrieval. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 251–260. ACM, 2010.

[7] David R Hardoon, Sandor Szedmak, and John Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16(12):2639–2664, 2004.

[8] Yangqing Jia, Mathieu Salzmann, and Trevor Darrell. Learning cross-modality similarity for multinomial data. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2407–2414. IEEE, 2011.

[9] Cuicui Kang, Shiming Xiang, Shengcai Liao, Changsheng Xu, and Chunhong Pan. Learning consistent feature representation for cross-modal multimedia retrieval. *IEEE Transactions on Multimedia*, 17(3):370–381, 2015.

[10] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression*, volume 398. John Wiley & Sons, 2013.

[11] Jerome Friedman, Trevor Hastie, Robert Tibshirani, et al. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The annals of statistics*, 28(2):337–407, 2000.

[12] Jianfeng He, Bingpeng Ma, Shuhui Wang, Yugui Liu, and Qingming Huang. Cross-modal retrieval by real label partial least squares. In *2016 ACM on Multimedia Conference*, pages 227–231. ACM, 2016.

[13] Abhishek Sharma and David W Jacobs. Bypassing synthesis: Pls for face recognition with pose, low-resolution and sketch. In *Computer Vision and Pattern Recognition, 2011. CVPR 2009. IEEE Conference on*, pages 593–600. IEEE, 2011.

[14] Beiying Ding and Robert Gentleman. Classification using generalized partial least squares. *Journal of Computational and Graphical Statistics*, 14(2):280–298, 2005.

[15] Murad Al Haj, Jordi Gonzalez, and Larry S Davis. On partial least squares in head pose estimation: How to simultaneously deal with misalignment. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2602–2609. IEEE, 2012.

[16] Roman Rosipal and Nicole Krämer. Overview and recent advances in partial least squares. *Lecture notes in computer science*, 3940:34, 2006.

[17] Roman Rosipal and Leonard J Trejo. Kernel partial least squares regression in reproducing kernel hilbert space. *Journal of machine learning research*, 2(Dec):97–123, 2001.

[18] Abhishek Sharma, Abhishek Kumar, Hal Daume, and David W Jacobs. Generalized multiview analysis: A discriminative latent space. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2160–2167. IEEE, 2012.

[19] Meina Kan, Shiguang Shan, Haihong Zhang, Shihong Lao, and Xilin Chen. Multi-view discriminant analysis. *IEEE transactions on pattern analysis and machine intelligence*, 38(1):188–194, 2016.

[20] Kamalika Chaudhuri, Sham M Kakade, Karen Livescu, and Karthik Sridharan. Multi-view clustering via canonical correlation analysis. In *the 26th annual international conference on machine learning*, pages 129–136. ACM, 2009.

[21] Lei Zhu, Jialie She, Xiaobai Liu, Liang Xie, and Liqiang Nie. Learning compact visual representation with canonical views for robust mobile landmark search. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pages 3959–3965, 2016.

[22] L. Zhu, Z. Huang, X. Liu, X. He, J. Sun, and X. Zhou. Discrete multi-modal hashing with canonical views for robust mobile landmark search. *IEEE Transactions on Multimedia*, 2017.

[23] Yunchao Wei, Yao Zhao, Zhenfeng Zhu, Shikui Wei, Yanhui Xiao, Jiashi Feng, and Shuicheng Yan. Modality-dependent cross-media retrieval. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 7(4):57, 2016.

[24] Yunchao Gong, Qifa Ke, Michael Isard, and Svetlana Lazebnik. A multi-view embedding space for modeling internet images, tags, and their semantics. *International Journal of Computer Vision*, 106(2):210–233, 2012.

[25] Yan Ke and Rahul Sukthankar. Pca-sift: a more distinctive representation for local image descriptors. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, pages II–506–II–513 Vol.2, 2004.

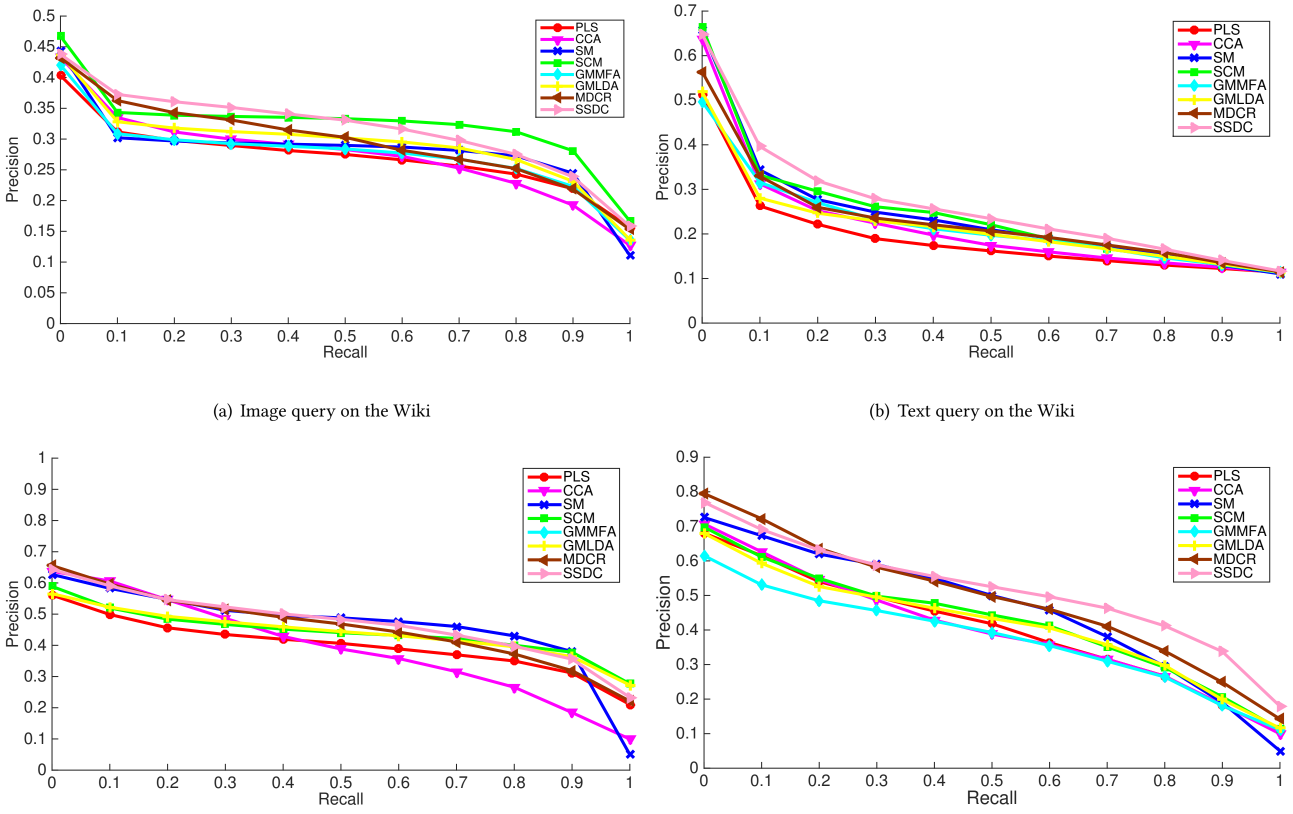
[26] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. *Latent dirichlet allocation*. JMLR.org, 2003.

[27] Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. Collecting image annotations using amazon's mechanical turk. In *the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 139–147. Association for Computational Linguistics, 2010.

[28] Liang Zheng, Yali Zhao, Shengjin Wang, Jingdong Wang, and Qi Tian. Good practice in cnn feature transfer. *arXiv preprint arXiv:1604.00133*, 2016.

[29] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016.

[30] Kazuaki Kishida. Property of mean average precision as performance measure in retrieval experiment. *Ipsj Sig Notes*, 2001:97–104, 2001.



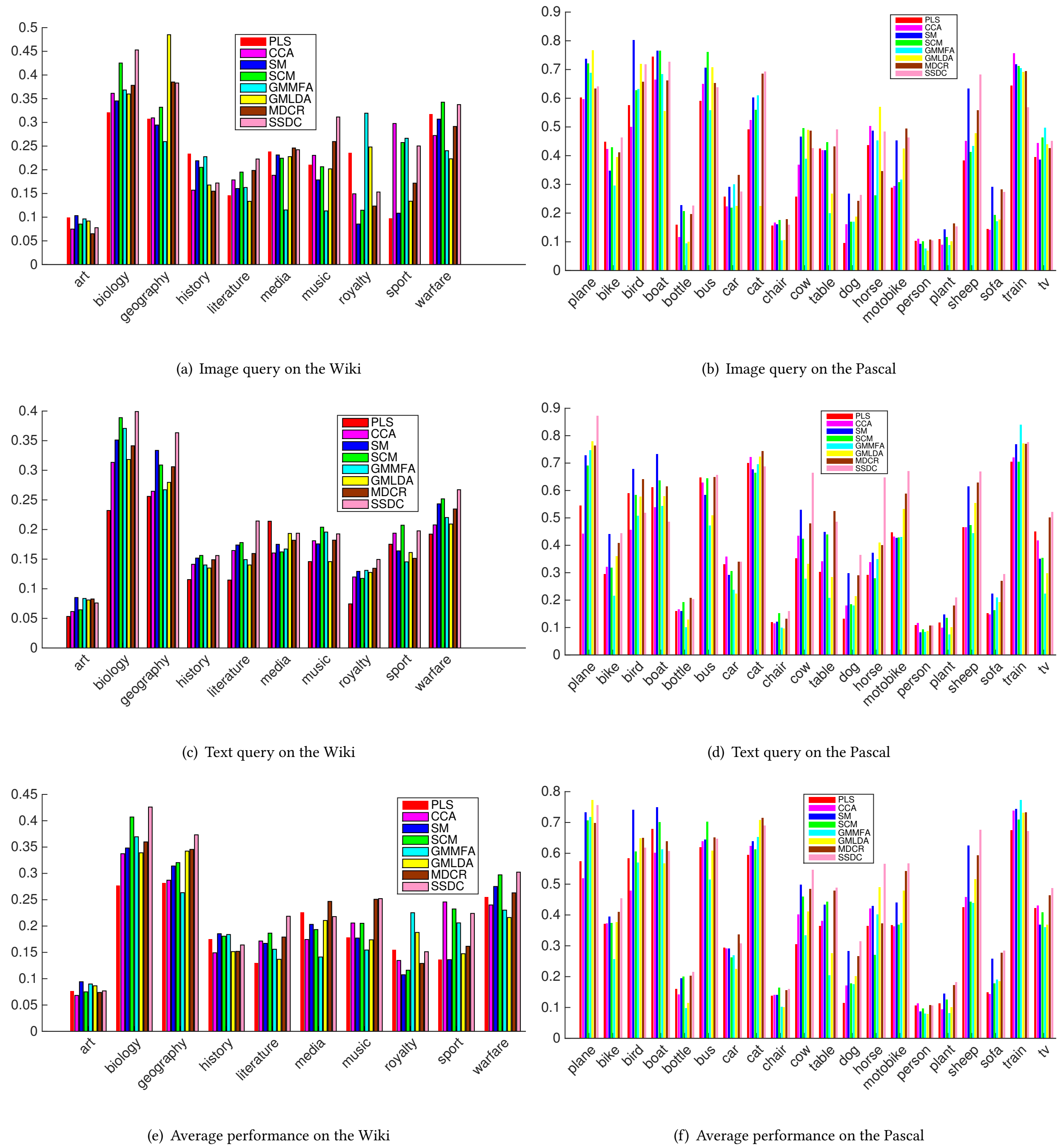
(a) Image query on the Wiki

(b) Text query on the Wiki

(c) Image query on the Pascal

(d) Text query on the Pascal

Figure 3: Precision-Recall curves of the proposed SSDC and compared methods.



(a) Image query on the Wiki

(b) Image query on the Pascal

(c) Text query on the Wiki

(d) Text query on the Pascal

(e) Average performance on the Wiki

(f) Average performance on the Pascal

Figure 4: mAP performance for each class on the Wikipedia dataset and the Pascal Sentence dataset.