CROSS-MODAL LEARNING TO RANK WITH ADAPTIVE LISTWISE CONSTRAINT

Guangzhuo Qu[1], *Jing Xiao**[1], *Jia Zhu*[1], *Yang Cao*[1], *Changqin Huang*[2]

[1] School of Computer Science, South China Normal University
[2] School of Information Technology in Education, South China Normal University
szqugz@163.com, xiaojing@scnu.edu.cn, {jzhu, caoyang}@m.scnu.edu.cn, cqhuang@scnu.edu.cn

ABSTRACT

Multi-modal data lies on heterogeneous feature spaces, which brings a significant challenge to cross-modal retrieval. Some works have been proposed to cope with this problem by learning a common subspace. However, previous methods often learn the common subspace by enhancing the relation between embedded features and relevant class labels but ignore the relation between embedded features and irrelevant class labels. Additionally, most methods assume that irrelevant samples are of equal importance. Considering this, we propose to train an optimal common embedding space via cross-modal learning to rank with adaptive listwise constraint (CMAL^{2}R) based on two-branch neural networks. The listwise loss function in CMAL^{2}R adaptively assigns larger margins to harder irrelevant samples, strengthening the relation between embedded features and irrelevant class labels. Experiments on Wikipedia and Pascal datasets demonstrate the effectiveness for bi-directional image-text retrieval.

Index Terms— Cross-modal retrieval, common space, adaptive listwise theory, cross-modal learning to rank

1. INTRODUCTION

This paper focuses on solving bi-directional image-text retrieval problem, which attracts increasing attention in cross-modal retrieval tasks [1], [2], [3], [4], [5]. However, because of the heterogeneity gap between data from different modalities, we cannot directly calculate cross-modal similarities. There have been many methods proposed for alleviating this problem by learning a common subspace [6]. The learning of common space has been the prevailing method in cross-modal retrieval, which can be mainly classified into unsupervised methods and supervised methods.

Unsupervised cross-modal methods learn the common subspace by utilizing paired samples between two different modalities. Canonical correlation analysis (CCA) [7] embeds image and text features respectively into a common space, which is learned by maximizing the pairwise correlation between the projected vectors of two different modalities. Deep CCA [8] extends the traditional CCA and simultaneously learns two deep nonlinear mappings by combining the autoencoder with CCA. In these methods, the feature vectors of each highly relevant image-text pair are respectively represented in the common subspace so as to calculate the similarity across different modalities.

Supervised cross-modal methods exploit class information to boost the learning of the common space of multi-modal data. These methods enforce inter samples to be mapped far apart while the intra samples lie as close as possible for obtaining a more discriminative common representation. For example, generalized multiview analysis (GMA) [9] , a supervised extension of CCA, learns a discriminative space by exploiting the class information. In [10], the authors propose a learning algorithm which aims to learn two projection matrices to map the data of the coupled spaces into the common space defined by class labels. Besides, multi-label Canonical Correlation Analysis (ml-CAA) [11] is proposed to cope with the problem of multi-label annotations. As an extension of CAA, ml-CAA effectively incorporates the multi-label information to learn a better shared subspace.

Although the above methods have made some contributions for cross-modal retrieval tasks, their performance still cannot meet the need in many real-word applications. The reason is that most methods focus on exploiting the relation between embedding features and relevant class labels but neglect the relation between embedded features and irrelevant class labels. That is to say, they pull each sample toward the direction of its relevant samples but pay little attention to pushing each sample far away from the directions of its irrelevant samples. In fact, the relation between embedded features and irrelevant class labels can provide abundant information for learning a more discriminative common subspace.

In order to overcome the above problem, we utilize learning to rank framework to learn the common space. Learning to rank, a kind of supervised learning, aims to train a ranking-based loss function to preserve the orders of the retrieved documents according to a given query. A lot of works have been proposed to learn a common subspace using learning to rank [12], [13], [14], [15], [16], [17]. In this paper,

* Corresponding author. This work was partially supported by National Natural Science Foundation of China (NSFC) projects No. 61202296, 61750110516 and Natural Science Foundation of Guangdong Province project No. S2012030006242.

we propose CMAL^{2}R which incorporates adaptive listwise constraint into cross-modal learning to rank to learn the common subspace. CMAL^{2}R can exploit a great deal of irrelevant samples to enhance the correlation between embedded features and irrelevant class labels. Furthermore, CMAL^{2}R can assign larger margins to harder irrelevant samples and capture the importance of irrelevant samples ranked in different positions. Since the learning algorithm is more concerned with the hard irrelevant samples than those irrelevant samples ranked behind, the learned common subspace will be more generalized. Finally, the proposed bi-directional loss function utilizes the semantic information from two directions, which motivates the algorithm to search better common subspace.

2. PROPOSED METHOD

2.1. Problem Description

Suppose the training dataset consists of n image-text pairs, i.e. $\mathcal{D} = \{(\mathbf{x_i}, \mathbf{y_i})\}_{i=1}^{n}$, where $\mathbf{x}_i \in \mathcal{X} \subseteq R^p$ denotes a p-dimensional visual feature vector from the i-th image and $\mathbf{y}_i \in \mathcal{Y} \subseteq R^q$ refers to a q-dimensional feature vector from the i-th text. The image set and text set are denoted as $\mathcal{X} = \{\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_n}\}$ and $\mathcal{Y} = \{\mathbf{y_1}, \mathbf{y_2}, ..., \mathbf{y_n}\}$ respectively. Note that the i-th image $\mathbf{x}_i \in \mathcal{X}$ and the i-th text $\mathbf{y}_i \in \mathcal{Y}$ come from same pair.

To investigate the latent correlation between relevant image and text, a common representation is learned for image and text from different modalities. Then the relevance between image and text can be measured by calculating their cosine similarity in the learned common multi-modal embedding space $\mathcal{E} \subseteq R^d$. For simplicity, we treat images as queries and texts as documents in following sections. Given an image query $\mathbf{x}_i \in \mathcal{X}$, we define a two-layer neural network to map image feature into a common multi-modal embedding space via $f^I : \mathcal{X} \rightarrow \mathcal{E}$, where $f^I(\cdot)$ is the image embedding function. Similarly, we map each text feature into a shared embedding space by $f^T : \mathcal{Y} \rightarrow \mathcal{E}$, where $f^T(\cdot)$ is the text embedding function. Through the embedding function $f(\cdot)$, the similarity measurement between image query $\mathbf{x}_i$ and retrieved text $\mathbf{y}_j$ can be simply obtained by computing the inner product in the shared embedding space, and then we normalize the similarity score to [0,1], i.e.,

$$s(f_i^I, f_j^T) = \frac{1}{2}(1 + {f_i^I}^T f_j^T). \quad (1)$$

In the above formula, $s(f_i^I, f_j^T)$ is simply represented as s_j^i which refers to the similarity score between i-th image query and j-th retrieved text. Then the embedded features from i-th image query and j-th retrieved text are denoted as f_i^I and f_j^T respectively. As f_i^I and f_j^T are normalization by the L2 norm, the inner product ${f_i^I}^T f_j^T$ actually equals to cosine similarity. In this case, the latent correlation between relevant image and text depends on the embedding parameters $\mathbf{W}$ from two-branch networks. Therefore, the major goal of our method is to effectively learn the embedding parameters, which will be detailed in the following subsections.

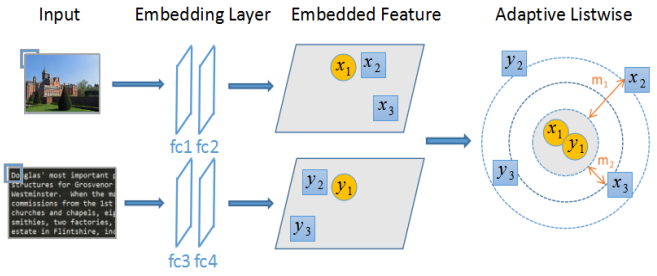

Fig. 1. Flowchart of the proposed CMAL^{2}R. For a given image set $\{\mathbf{x}_i\}_{i=1}^{K+1}$ and text set $\{\mathbf{y}_j\}_{j=1}^{K+1}$, we firstly embed their preprocessed features into the common space by two-branch networks with four full-connected layers(fc) to obtain their feature $\{f_i^I\}_{i=1}^{K+1}$ and $\{f_j^T\}_{j=1}^{K+1}$, then the pairwise similarity scores are calculated as $\{s_j^I\}_{j=1}^{K+1}$ and $\{s_i^T\}_{i=1}^{K+1}$ for the image query and text query respectively. Next, we map each score list to the distribution of the top one group. Finally, we define the bi-directional listwise loss function as the negative log likelihood of the ground-truth top one group: $-\log P(\mathbf{\Omega}_1|\mathbf{x}_1) - \log P(\mathbf{\Omega}_1|\mathbf{y}_1)$. The circle with yellow color represents an image query or text query and the squares with blue color represent the documents from different classes. The bi-directional listwise loss function can assign adaptive margin according to the rank of different documents. The nearer the document is, the larger the value m_k is assigned.

2.2. Loss Function

To learn the parameters of two-branch networks as shown in Fig.1, the cross-modal retrieval task is formulated as a listwise learning-to-rank problem. Firstly, we give some important concepts. Suppose that all document texts are identified with the numbers 1,2,...,n. A permutation π on the texts is defined as a bijection from 1,2,...,n to itself. We write the permutation as $\pi = \langle \pi(1), \pi(2), ..., \pi(n) \rangle$, where $\pi(j)$ denotes the text index at position j and $\pi^{-1}(j)$ denotes the rank of text $\mathbf{y}_j$ in permutation π.

In order to predict the permutation correctly, the crucial issue is how to design a listwise loss function to measure the difference between the predicted permutation and the ground-truth permutation. Based on the consideration, we introduce the ListMLE [18] method to define such loss function, which can transform ranking similarity scores to a probability distribution. And then we can maximize the negative log likelihood

of ground-truth permutations as the loss function.

$$L_{total} = -\sum_{i=1}^{n} \log P(\pi_i|\mathbf{x}_i) \tag{2}$$

where π_i denotes the text permutation corresponding to the image query $\mathbf{x}_i$. Given an image query $\mathbf{x}_i$, however, we only know the ground-truth text should be ranked in the first position and no information about orders within the irrelevant texts is provided. That is to say, we can't obtain the total order of all the retrieved texts which is needed to generate the ground-truth permutation. Thus, the above loss function can't be directly applied to deal with cross-modal retrieval task.

To alleviate the above problem, the top one probability model is employed which only forces the ground-truth document to be ranked before all the irrelevant documents. Given an image query $\mathbf{x}_i$, we use $\mathbf{\Omega}_i$ to denote the set of all ground-truth permutations consistent with this restriction

$$\mathbf{\Omega}_i = \{\pi|\pi(1) = i\} \tag{3}$$

which denotes that the top document in all the permutations of $\mathbf{\Omega}_i$ is exactly $\mathbf{y}_i$. Next, we can easily obtain the top one probability by

$$P(\mathbf{\Omega}_i|\mathbf{x}_i) = \frac{\phi(s_i^i)}{\sum_{k=1}^{n} \phi(s_k^i)} \tag{4}$$

where $\phi(\cdot)$ is an increasing and strictly positive mapping function. For each i ranging from 1 to n, obviously, the top one probabilities $P(\mathbf{\Omega}_i|\mathbf{x}_i)$ form a probability distribution over set $\mathbf{\Omega}_i$. So given the training set $\{\mathbf{x}_i, \mathbf{\Omega}_i\}_{i=1}^n$, we can define the loss function as the negative log likelihood of the ground-truth top one group as follows:

$$L_{top\ one} = -\sum_{i=1}^{n} \log P(\mathbf{\Omega}_i|\mathbf{x}_i). \tag{5}$$

Different from conventional listwise learning-to-rank methods such as ListMLE [18] and ListNet [19] which pay identical attention to different training samples, we introduce an adaptive listwise constraint [20] to cope with the problem. In the following section, we will detail the adaptive listwise method.

2.3. An Adaptive Margin Listwise Loss

We employ the mapping function as follows:

$$\phi(s_k^i) = \exp(\frac{s_k^i + m_k^i}{\beta}) \tag{6}$$

where m_k^i refers to the adaptive margin between i-th query and k-th retrieved document and β is a sharpness parameter. Now, we can rewrite the loss of a single training sample as follows:

$$l = \log(1 + \sum_{k \neq i} \exp(\frac{s_k^i - s_i^i + m_k^i - m_i^i}{\beta})). \tag{7}$$

Since what we only concern about is the difference between m_k^i and m_i^i, m_i^i is always set to zero and m_k^i is referred as an adaptive margin which can improve the generalization performance. The adaptive margin is simply defined as follows [20]:

$$m_k^i = \begin{cases} \frac{3}{4} - \frac{\tilde{\pi}_i^{-1}(k) - 2}{2(n-2)}, & k \neq i \\ 0, & k = i \end{cases} \tag{8}$$

where $\tilde{\pi}_i^{-1}(k)$ is the rank of k-th retrieved document in permutation $\tilde{\pi}_i$. m_k^i is limited to the range $[\frac{1}{4}, \frac{3}{4}]$. $\tilde{\pi}_i$ is obtained by keeping the relevant text at the top position, and sorting irrelevant texts in descending order according to their similarity scores

$$\tilde{\pi}_i = \left\langle i, sort(s_k^i) \right\rangle, k \neq i. \tag{9}$$

According to (8), the nearer irrelevant texts can be assigned larger margins than those further ones, which means the hard irrelevant samples will be pushed a larger margin away from the image query. Specially, the negative samples from diverse classes are pushed different distance away from the given query, which can be interpreted as exploring the relation between embedded features and irrelevant class labels.

2.4. Sampling Algorithm

In order to reduce the computation, we propose a more scalable method to learn the network's parameters. The main idea is that we randomly sample T image queries and T corresponding text queries. Then we randomly sample K irrelevant texts and K irrelevant images for each image query and text query respectively (the irrelevant documents are all indexed with $N = \{n_i\}_{i=1}^K$). Thus, there are $T(K+1)$ images and $T(K+1)$ texts in each mini-batch. Instead of sampling one training sample with a long permutation, our proposed method utilizes more training samples with shorter permutations in each iteration, which is more efficient in computation. So when we treat images as queries, the loss function in each iteration is as follows:

$$L_1(s) = \sum_{i=1}^{T} \left\{ \log \sum_{k \in N \setminus i} \exp(\frac{s_k^i + m_k^i}{\beta}) - \frac{s_{i^+}^i + m_{i^+}^i}{\beta} \right\}. \tag{10}$$

Similarly, when we treat texts as queries, the loss function in each iteration is as follows:

$$L_2(s) = \sum_{j=1}^{T} \left\{ \log \sum_{k \in N \setminus j} \exp(\frac{s_k^j + m_k^j}{\beta}) - \frac{s_{j^+}^j + m_{j^+}^j}{\beta} \right\}. \tag{11}$$

Table 1. MAP score comparison with state-of-the-art methods on two benchmark datasets.

Method	Img2Txt	Txt2Img	Average	Dataset
CCA	0.2160	0.1872	0.2016	Wiki
SCM	0.2759	0.2336	0.2548	
LCFS	0.2711	0.2043	0.2377	
MvDA	0.2971	0.2319	0.2645	
LGCFL	0.3775	0.3160	0.3467	
ml-CCA	0.3527	0.2873	0.3120	
GMLDA	0.3159	0.2885	0.3022	
GMMFA	0.3155	0.2964	0.3060	
CMAL^{2}R	**0.4251**	**0.3377**	**0.3814**	
CMAL^{2}R(bi.)	**0.4316**	**0.3416**	**0.3866**	
CCA	0.3073	0.2945	0.3009	Pascal
LCFS	0.4278	0.3355	0.3816	
LGCFL	0.4362	0.3440	0.3901	
ml-CAA	0.4303	0.3885	0.4094	
CMAL^{2}R	**0.5112**	**0.4792**	**0.4952**	
CMAL^{2}R(bi.)	**0.5172**	**0.4831**	**0.5002**	

Finally, the bi-directional listwise loss is as follows:

$$L(s) = \alpha L_1(s) + (1-\alpha)L_2(s) + \frac{\lambda}{2}\sum_{m=1}^{4}\|\mathbf{W}^{(m)}\|_F^2 \quad (12)$$

where α balances the strength of the listwise loss in each direction, λ controls the strength of regularization, and $\{\mathbf{W}^{(m)}\}_{m=1}^{4}$ are the parameters of two-branch networks.

3. EXPERIMENTAL RESULTS

3.1. Datasets

1)Wikipedia Dataset [21]: This dataset consists of 2866 image-text pairs and it is divided into 10 semantic categories. For fair comparison, we use the same image features, which are represented into a 4096 dimensional vectors from the fc7 layer of CNN [22], and text features, which are generated into the 100 dimensional skip-gram word vectors by the word2vec model and a simple average calculation. Moreover, we follow a same dataset partition according to [23] and 2000 pairs and 866 pairs are selected for training and testing respectively.

2)Pascal Dataset [24]: This dataset consists of 5011 and 4952 image-tag pairs from 20 different semantic classes for training and testing respectively [25]. Each pair belongs to one or more of 20 semantic classes. The image and text features are provided by this dataset, which are represented as the 512-dimensional GIST features and 399-dimensional word frequency features respectively. We follow the origin training-test split and remove some images without the corresponding tags. Eventually, 5000 pairs are used for training and 4919 pairs for testing.

3.2. Experimental Results

The proposed method is compared with several the state-of-the-arts, such as CCA & SCM [21], GMLDA & GMMFA [9], LCFS [10], LGCFL [26], ml-CCA [11] and MvDA [27]. We set α to 0.4 and set β to 0.5 in all experiments. For fairness, the same image-text features and train/test division are used in all methods. The mean average precision(MAP) [21] scores on Wikipedia and Pascal datasets of all the methods are shown in Table 1.

As we can see from Table 1, our proposed method outperforms all the compared methods by a large margin. The reason is that the proposed method can exploit a great deal of unpair data to enhance the correlation between embedded features and irrelevant class labels. Furthermore, it pays different attention to the irrelevant samples according to the positions in which those samples are ranked. Hence, a discriminative and generalized common subspace can be learned in our framework.

For Wiki dataset, our method achieves the best average MAP of 0.3866, which is about 4% higher than the second best result from LGCFL. This is because LGCFL only utilizes the label information as interlinkage to model the image space and text space and ignores the importance of irrelevant samples ranked in different positions. However, our method can pay different attention to the irrelevant samples by the adaptive margins.

For Pascal dataset, we can see that the proposed method achieves the best performance with the average MAP at 0.5002, which is about 9% higher than the second best result from ml-CCA. ml-CCA only utilizes semantic information, in the form of multi-label information, to establish the correlation across the modalities but ignores the relation between embedded features and irrelevant class labels.

Finally, bi-directional retrieval results have a little improvement over those single-directional results. The reason is that bi-directional listwise loss can make the best of semantic information from two directions, which motivates the algorithm to search better common subspace.

4. CONCLUSION

In this paper, we employ two-branch networks to transfer heterogeneous feature spaces to a common embedding space and introduce adaptive listwise constraint into cross-modal learning to rank to train a discriminative multi-modal embedding space. This method can exploit the relation between embedded features and irrelevant class labels. Furthermore, the proposed bi-directional listwise loss function can adaptively pay unequal attention to the irrelevant samples according to the ranks of those samples. The comprehensive experiment results on two cross-modal datasets demonstrate the effectiveness of the proposed method for bi-directional image-text retrieval.

5. REFERENCES

[1] F. yan and k. Mikolajczyk, "Deep correlation for matching images and text," in *CVPR*, 2015, pp. 3441–3450.

[2] Z. Ren, H. Jin, Z. L. Lin, C. Fang, and A. L. Yuille, "Joint image-text representation by gaussian visual-semantic embedding," in *ACMMM*, 2016, pp. 207–211.

[3] L. Zhang, B. P. Ma, G. R. Li, Q. M. Huang, and Q. Tian, "Pl-ranking: a novel ranking method for cross-modal retrieval," in *ACMMM*, 2016, pp. 1355–1364.

[4] M. K. Yuan: X. Huang, Y. X. Peng, "Cross-modal common representation learning by hybrid transfer network," in *IJCAI*, 2017, pp. 1893–1900.

[5] J. L. Wu, Z. C. Lin, and H. B. Zha, "Joint latent subspace learning and regression for cross-modal retrieval," in *SIGIR*, 2017, pp. 917–920.

[6] X. Y. Jiang, F. Wu, X. Li, Z. Zhao, W.M. Lu, S.L. Tang, and Y.T. Zhuang, "Deep compositional cross-modal learning to rank via local-global alignment.," in *ACMMM*, 2015, pp. 69–78.

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

[8] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *ICML*, 2013, pp. 1247–1255.

[9] A. Sharma, A. Kumar, D. Hal, and D. Jacobs, "Generalized multiview analysis: a discriminative latent space," in *CVPR*, 2012, pp. 2160–2167.

[10] K. Wang, R. He, W. Wang, L. Wang, and T. Tan, "Learning coupled feature spaces for cross-modal matching," in *IEEE International Conference on Computer Vision*, 2013, pp. 2088–2095.

[11] V. Ranjan, N. Rasiwasia, and C. Jawahar, "Multi-label cross-modal retrieval," in *ICCV*, 2015, pp. 4094–4102.

[12] J. Weston, S. Bengio, and N. Usunier, "Wsabie: Scaling up to large vocabulary image annotation," in *IJCAI*, 2011, pp. 2764–2770.

[13] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, and T. Mikolov, "Devise: A deep visual-semantic embedding model," in *NIPS*, 2013, pp. 2121–2129.

[14] L. Wang, Y. Li, and S. Lazebnik, "Learning deep structure-preserving image-text embeddings," in *CVPR*, 2016, pp. 5005–5013.

[15] Y. Yue, T. Finley, F. Radlinski, and T. Joachims, "A support vector method for optimizing average precision," in *SIGIR*, 2007, pp. 271–278.

[16] T. Qin, T.-Y. Liu, X.-D. Zhang, D.-S. Wang, and H. Li, "Global ranking using continuous conditional random fields," in *NIPS*, 2008, pp. 1281–1288.

[17] F. Wu, X. Y. Jiang, X. Li, S. L. Tang, W. M. Lu, Z. F. Zhang, and Y. T. Zhuang, "Cross-modal learning to rank via latent joint representation," *IEEE Trans. Image Processing*, vol. 24, no. 5, pp. 1497–1509, 2015.

[18] F. Xia, T. Y. Liu, J. Wang, W. S. Zhang, and H. Li, "Listwise approach to learning to rank: theory and algorithm," in *ICML*, 2008, pp. 1192–1199.

[19] Z. Cao, T. Qin, T.-Y. Liu, M.-F. Tsai, and H. Li, "Learning to rank: From pairwise approach to listwise approach," in *ICML*, 2007, pp. 129–136.

[20] J. Wang, Z. Wang, C. X. Gao, N. Sang, and R. Huang, "Deeplist: learning deep features with adaptive listwise constraint for person reidentification," *IEEE Trans. Circuits Syst. Video Techn.*, vol. 27, no. 3, pp. 513–524, 2017.

[21] N. Rasiwasia, J. C. Pereira, and E. Coviello, "A new approach to cross-modal multimedia retrieval," in *ACMMM*, 2010, pp. 251–260.

[22] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: convolutional architecture for fast feature embedding," in *ACMMM*, 2014, pp. 675–678.

[23] L. Zhang, B. Ma, J. F. He, G. R. Li, Q. M. Huang, and Q. Tian, "Adaptively unified semi-supervised learning for cross-modal retrieval," in *IJCAI*, 2017, pp. 3406–3412.

[24] S. J. Hwang and K. Grauman, "Reading between the lines: Object localization using implicit cues from image tags," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 6, pp. 1145–1158, 2012.

[25] M. Everingham, V. Gool, C. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.

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

[27] C. Kang, S. Xiang, S. Liao, C. Xu, and C. Pan, "Multiview discriminant analysis," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 38, no. 1, pp. 188–194, 2016.