Cross-modal Retrieval by Real Label Partial Least Squares

Jianfeng He[1,3], Bingpeng Ma[1,2,3]*, Shuhui Wang[2], Yugui Liu[1,3], Qingming Huang[1,2,3]
[1] School of Computer and Control Engineering, University of Chinese Academy of Sciences, China
[2] Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences(CAS), China
[3] Key Laboratory of Big Data Mining and Knowledge Management, CAS, China
{jianfeng.he, shuhui.wang}@vipl.ict.ac.cn {bpma, liuyg, qmhuang}@ucas.ac.cn

ABSTRACT

This paper proposes a novel method named Real Label Partial Least Squares (RL-PLS) for the task of cross-modal retrieval. Pervious works just take the texts and images as two modalities in PLS. But in RL-PLS, considering that the class label is more related to the semantics directly, we take the class label as the assistant modality. Specially, we build two KPLS models and project both images and texts into the label space. Then, the similarity of images and texts can be measured more accurately in the label space. Furthermore, we do not restrict the label indicator values as the binary values as the traditional methods. By contraries, in RL-PLS, the label indicator values are set to the real values. Specially, the label indicator values are comprised by two parts: positive or negative represents the sample class while the absolute value represents the local structure in the class. By this way, the discriminate ability of RL-PLS is improved greatly. To show the effectiveness of RL-PLS, the experiments are conducted on two cross-modal retrieval tasks (Wiki and Pascal Voc2007), on which the competitive results are obtained.

Keywords

Cross-modal retrieval, multimedia, partial least squares, images and documents.

1. INTRODUCTION

With the development of Internet technology, the quantity of multimedia has increased dramatically. It is a great challenge to retrieve the information from the different modalities. Recently, more and more people pay their attention to the task of cross-modal retrieval [1] [9] [12] [20] [22] [24] [25]. The key problem of cross-modal retrieval is matching one modal feature to the other modal feature in the content-based area [13]. For example, in the image-text retrieval, given an image query, find the texts that best describe the image; or given a text document, return the most related images. However, the text features and the image features are in the different feature space so that they can't be matched directly with each other.

To solve the retrieval problem of the different feature spaces, recent works are devoted to the common subspace learning [8] [9] [13] [19]. These methods try to learn a common feature space so as to match the image and text features directly and preserve the correlations in the image-text pairs. As one of possible solutions, Canonical Correlation Analysis (CCA) projects two modal features to a shared latent space which maximizes the correlations between two modal features [7] [10] [13] [17]. Many extensions of CCA have been used in the similar area. For example, Semantic Correlation Match (SCM) was proposed to get the semantic subspace by using the logistic regressor based on CCA [13]. In [20], without no assumption on specific form of text, Correlated Semantic Representation (CSR) obtains a joint image-text representation and an unified formulation by learning a compatible function based on structural SVM.

Another classical method of subspace learning is Partial Least Squares (PLS) [2] [6] [14] [15] [16], aiming at learning two latent spaces by maximizing the correlations between them. Sharma *et al.* [17] and Kang *et al.* [9] applied PLS to build the relation between the latent variables of images and texts. Besides cross-modal retrieval, PLS and its extensions have been applied successfully on other problems. For an instance, in the task of cross-pose face recognition, the relation between the coupled faces are constructed by PLS [5] [11] [23]. In [18], bridge PLS was proposed by adding ridge-parameter to improve the efficiency in each iteration. Rosipal *et al.* [15] proposed the kernel PLS by mapping the input variables into high dimension space so as to solve the nonlinear problem in linear algorithm.

Moreover, most of the above algorithms don't use the label information. Generally speaking, cross-modal retrieval problem always exists semantic gap between the different modalities [4] [13] [21]. But by using the label information, the semantic gap can be decreased theoretically [4] [21]. Specially, using labels for extracting muliti-view features, GMLDA and GMMFA, constructed by framework based on Generalized Multiview Analysis (GMA), shown the competitive performance on the cross-modal retrieval problem [17]. In [9], which gains the state-of-art performance on cross-modal retrieval problem, the label information are used to close different modalities within the same class and far away between different modalities.

*Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

MM '16, October 15-19, 2016, Amsterdam, Netherlands
© 2016 ACM. ISBN 978-1-4503-3603-1/16/10...$15.00
DOI: http://dx.doi.org/10.1145/2964284.2967216

In this paper, we propose a supervised algorithm named Real Label Partial Least Squares (RL-PLS), which projects two modal feature into a common space and reserves the label information and local structure while at the same time. Different with the above works [16] [17], we regard the label information as the assistant modality. Thus, we achieve a three modality shared feature space to obtain better performance than only using two modalities. Further, considering that the local structure of samples is ignored when all the samples existed in the different regions have the same label [9], RL-PLS constructs the class indicator matrix using real values to reserve the local structure. The label values in the matrix are comprised by two parts: character and value. Character represents the relationship between sample and class while the absolute value represents the local structure in the class. By using the probability belonging to the class, the different samples are given the different weights. Finally, the discriminative ability of RL-PLS is improved greatly.

2. KERNEL PARTIAL LEAST SQUARES

PLS can construct the relation between two different modalities by maximizing the correlation between the latent variables. Recently, it has gained great success in many areas [2] [6] [16]. Kernel partial least squares(KPLS), the extension of PLS, maps original feature into a new high dimension feature space to solve the nonlinear problem with linear algorithm. Because of reserving more information in nonlinear problem, KPLS often get better performance.

For the original feature of training set, its kernel feature as the input variables can be denoted by $\boldsymbol{X} \in \mathbb{R}^{n \times n}$, where n is the number of training samples. The latent variables of $\boldsymbol{X}$ are represented as $\boldsymbol{V} = [\boldsymbol{v}_1, \cdots, \boldsymbol{v}_n]^T \in \mathbb{R}^{n \times p}$, where p is the dimension of latent variables and far smaller than n. Similarly, $\boldsymbol{Z} = [\boldsymbol{z}_1, \cdots, \boldsymbol{z}_n]^T \in \mathbb{R}^{n \times d}$ denotes the output variables of training set, where d is the number of dimensions. It is represented by $\boldsymbol{U} = [\boldsymbol{u}_1, \cdots, \boldsymbol{u}_n]^T \in \mathbb{R}^{n \times p}$. Finally, KPLS can be built as:

$$\begin{cases} \boldsymbol{X} = \boldsymbol{V}\boldsymbol{P}^T + \boldsymbol{\varepsilon}_x \\ \boldsymbol{Z} = \boldsymbol{U}\boldsymbol{Q}^T + \boldsymbol{\varepsilon}_z \end{cases} \tag{1}$$

where the matrices $\boldsymbol{P}$ and $\boldsymbol{Q}$ represent the loading matrices, the matrices $\boldsymbol{\varepsilon_x}$ and $\boldsymbol{\varepsilon_z}$ are the residuals matrices. And according to [15], by means of the low dimension latent variables $\boldsymbol{V}, \boldsymbol{U}$, we can further get a regression coefficient matrix $\boldsymbol{B} \in \mathbb{R}^{n \times d}$ to get the relation between $\boldsymbol{X}$ and $\boldsymbol{Z}$:

$$\begin{cases} \boldsymbol{B} = \boldsymbol{X}^T\boldsymbol{U}(\boldsymbol{V}^T\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{U})^{-1}\boldsymbol{V}^T\boldsymbol{Z} \\ \boldsymbol{Z} = \boldsymbol{X}\boldsymbol{B}^T + \boldsymbol{\varepsilon}_B \end{cases} \tag{2}$$

where $\boldsymbol{\varepsilon}_B$ is the matrix of residuals.

KPLS can be solved by Nonlinear Partial Least Squares (NIPALS) algorithm. More details about KPLS and NIPALS can be found in [15]. In Fig. 1 (a), we show the relation between different variables in KPLS model.

3. REAL LABEL PARTIAL LEAST SQUARES

In this section, we introduce the proposed method in details. In Fig. 1 (b), we show the structure diagram of the proposed method.

3.1 Labels in KPLS

From Eq. 2, we can know that in KPLS, the input modality $\boldsymbol{X}$ is projected into the space of the output modality $\boldsymbol{Z}$

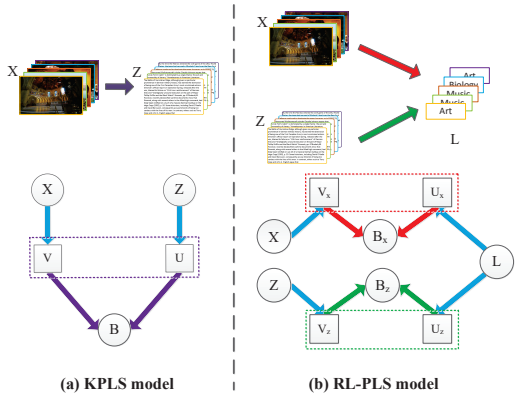

Figure 1: (a) The structure diagram of PLS. (b) The structure diagram of RL-PLS. The rectangles mean latent space.

and the similarity of the different modalities are measured in this space. Considering the semantic gap and the great discrepancy between texts and images in cross-model retrieval, the direct projection from the input modality to the output modality inevitably causes the information loss of the input modality and decreases the retrieval performance.

Compared with texts and images, label information is more related to the semantic information. Many pairs of two modalities share the common labels in cross-modal problem. So, for the semantic gap between texts and images in cross-model retrieval, we argue that label information can help to build the relation between texts and images.

In this paper, we use $\boldsymbol{L}$ to denote the label indicator matrix. We take labels as the auxiliary modality and propose the RL-PLS method. In RL-PLS, we project images and texts into the semantic space simultaneously. By this way, the similarity of the different modalities can be measured more effectively in the semantic space. Specially, taking images as the input variables and labels as the output variables, $KPLS_{XL}$ constructs the relation between images and labels. Similarly, taking texts as the input variables and labels as the output variables, $KPLS_{ZL}$ constructs the relation between texts and labels:

$$KPLS_{XL}: \begin{cases} \boldsymbol{X} = \boldsymbol{V}_x\boldsymbol{P}_x^T + \boldsymbol{\varepsilon}_x \\ \boldsymbol{L} = \boldsymbol{U}_x\boldsymbol{Q}_x^T + \boldsymbol{\varepsilon}_{Lx} \end{cases} \tag{3}$$

$$KPLS_{ZL}: \begin{cases} \boldsymbol{Z} = \boldsymbol{V}_z\boldsymbol{P}_z^T + \boldsymbol{\varepsilon}_z \\ \boldsymbol{L} = \boldsymbol{U}_z\boldsymbol{Q}_z^T + \boldsymbol{\varepsilon}_{Lz} \end{cases} \tag{4}$$

where $\boldsymbol{V}_x$, $\boldsymbol{U}_x$, $\boldsymbol{V}_z$ and $\boldsymbol{U}_z$ are the latent variables; $\boldsymbol{P}_x$, $\boldsymbol{Q}_x$, $\boldsymbol{P}_z$ and $\boldsymbol{Q}_z$ represent the loading matrices; $\boldsymbol{\varepsilon}_x$, $\boldsymbol{\varepsilon}_{Lx}$, $\boldsymbol{\varepsilon}_z$ and $\boldsymbol{\varepsilon}_{Lz}$ are the residuals. Eq. 3 and Eq. 4 can be solved with NIPALS algorithm. According to Eq. 2, the regression coefficient matrix $\boldsymbol{B}_x$ in $KPLS_{XL}$ and $\boldsymbol{B}_z$ in $KPLS_{ZL}$ can be computed as follows:

$$KPLS_{XL}: \begin{cases} \boldsymbol{B}_x = \boldsymbol{X}^T\boldsymbol{U}_x(\boldsymbol{V}_x^T\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{U}_x)^{-1}\boldsymbol{V}_x^T\boldsymbol{L} \\ \boldsymbol{L} = \boldsymbol{X}\boldsymbol{B}_x^T + \boldsymbol{\varepsilon}_{Bx} \end{cases} \tag{5}$$

$$KPLS_{ZL}: \begin{cases} \boldsymbol{B}_z = \boldsymbol{Z}^T\boldsymbol{U}_z(\boldsymbol{V}_z^T\boldsymbol{Z}\boldsymbol{Z}^T\boldsymbol{U}_z)^{-1}\boldsymbol{V}_z^T\boldsymbol{L} \\ \boldsymbol{L} = \boldsymbol{Z}\boldsymbol{B}_z^T + \boldsymbol{\varepsilon}_{Bz} \end{cases} \tag{6}$$

Finally, $\boldsymbol{B}_x$ and $\boldsymbol{B}_z$ map the input variables which are heterogenous features $\boldsymbol{X}$ and $\boldsymbol{Z}$ into the target feature space constructed by label information.

3.2 Real Labels

In $KPLS_{XL}$ and $KPLS_{ZL}$, the similarity of different modalities are computed in the label space. In the previous works, the class indicators often are the binary matrix. In other words, all the elements are zeros except for the corresponding classes [9]. Specially, if sample $\boldsymbol{a}$ is assigned to the kth class, its indicator vector $\boldsymbol{l}$ is constructed by $l_k = 1$, and $l_j = 0$ for $j \neq k$. For all the samples belonging to the same class, their class indicator vectors are the same. In fact, different samples exist in the different regions in the high dimensional space. The same labels mean that the local structure of samples is ignored.

In RL-PLS, we construct the class indicator matrix using real values to reserve the local structure. We assume that data points in each class can be modeled by several Gaussian distributions. The values in the class indicator matrix are set to the probabilities that samples belong to the clustering center. So, for the samples which near the class center, they have the bigger probability belonging to the class and their values are close to 1. For the samples which far away the class center, they have the bigger probability to be the outliers and their values are close to 0. In other words, by using the probability belonging to the class, we pay more attentions on the samples which near the clustering center and reduce the influence of the outliers.

Specially, there are five steps to compute class indicator matrix $\boldsymbol{L}$ in RL-PLS. Firstly, for each class in the modality $\boldsymbol{X}$, we use r Gaussian distributions to model it. So, for the total q classes in the training set, we can gain $s(= r \times q)$ Gaussian distributions $\{G_1, G_2, \cdots, G_s\}$. For the jth Gaussian distribution G_j, the class of its corresponding samples is c_j^g. The clustering center vector and covariance matrix of G_j is $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_j$, respectively.

Secondly, we compute the Mahalanobis distance d_{ij} between $\boldsymbol{x}_i$ and $\boldsymbol{\mu}_j$ under the covariance matrix $\boldsymbol{\Sigma}_j$:

$$d_{ij} = (\boldsymbol{x}_i - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_j^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_j) \tag{7}$$

Then, we compute the probability w_{ij} that $\boldsymbol{x_i}$ belongs to the j-th Gaussian distribution:

$$w_{ij} = \exp(-\frac{d_{ij}}{\sigma^2}) \tag{8}$$

where σ is the standard deviation of $\boldsymbol{D} = [\boldsymbol{d}_1, \boldsymbol{d}_2, \cdots, \boldsymbol{d}_s]^T$.

Fourthly, p_{ij} is designed to show the relation between the sample's class and the Gaussian distribution. Suppose c_i is the class label of sample x_i, the p_{ij} is set to 1 when c_i is equal to c_j^g. Otherwise, it is set to -1.

$$p_{ij} = \begin{cases} +1, & if \quad c_i = c_j^g \\ -1, & otherwise. \end{cases} \tag{9}$$

Finally, l_{ij} is computed as:

$$l_{ij} = p_{ij} \times w_{ij} \tag{10}$$

Then, after the computation of the probabilistic class indicator vector of all the samples in the training set, the indicator matrix $\boldsymbol{L}$ can be gained by: $\boldsymbol{L} = [\boldsymbol{l}_1, \boldsymbol{l}_2, \cdots, \boldsymbol{l}_n]^T$ and $\boldsymbol{l}_i = (l_{i1}, l_{i2}, \cdots, l_{is})$.

From Eq. 10, we can know that the class indicator vector can be divided into two parts: character and value. For the samples whose class label is same to the label of the Gaussian distribution, the real values of class indicator can keep the local structure of the within-class. For the samples which has the different class label with the Gaussian distribution but it has the large probability w_{ij}, the negative value of the class indicator can increase the margin of the inter-classes. So, by the combination of the character and value, the discriminate ability of $\boldsymbol{L}$ is improved greatly.

It is must be noted that for images and texts, their class indicator matrices are different according to the computation process of $\boldsymbol{L}$. But considering that we try to find a common projection subspace for texts and images while the distributions of texts are not as complicated as images', we just construct $\boldsymbol{L}$ based on the texts modality.

3.3 Analysis of Computational Complexity

Lastly, we briefly analyze the computational complexity of RL-PLS, which involves h dimensions of latent variables, and each dimension is solved by NIPALS algorithm, in which the max iterations is set as g. Set n as the number of sample pairs in the training set, thus image feature and text feature are n dimensions as a result of RBF kernel. And set q as the number of total classes. The computational complexity of RL-PLS is $O(hgn^3 + n^2h^5q)$.

4. EXPERIMENTS

In this section, we test RL-PLS on two popular databases to show its performance on cross-modal retrieval.

4.1 Experimental Databases

Wiki [13] is constructed from Wikipedia including 2866 image-text pairs with 10 different classes. We randomly select 2173 image-text pairs for the training set and 693 image-text pairs for the testing set. The texts in Wiki are represented by the 10-dimensional Latent Dirichlet Allocation (LDA) features. For images, we use two different features to show the effectiveness of RL-PLS. On one hand, the images are represented by the traditional 128-dimensional vectors based on the SIFT descriptors; on the other hand, considering the success of deep learning, we also perform experiments using the Convolutional Neural Network (CNN). So the image features are also represented by the 4096-dimensional vectors output from the 'fc7' layer of convolutional neural network(CNN) based on central crop (no mirroring).

Pascal VOC2007 [3] consists of 5011/4952(training/testing) image-tag pairs with 20 classes. Taking account that some pairs of the database are multi-label, we choose single-label pairs in our experiment. As a result, training set includes 2808 image-tag pairs and testing set includes 2841 image-tag pairs. For feature representation, we use the 512-dimensional GIST features for images, and the 399-dimensional word frequency features for text.

4.2 Experimental Setting

We compare RL-PLS with PLS, KPLS, SCM, GMMFA, GMLDA and LGCFL in two retrieval tasks. PLS and KPLS are unsupervised learning methods while the other methods are supervised learning methods. In PLS and KPLS, we set the image features as input variables, and the text feature as output variables. As for SCM, it uses CCA to learn two maximally correlated subspaces, and then learns the logistic regressors in each subspaces. Both GMMFA and GMLDA use the framework of GMA[17]. LGCFL uses the valuable class information to jointly learn the consistent features and achieves the state-of-the-art performance.

Table 1: MAP results on the Wiki database. Images are represented by SIFT features.

Methods \ Tasks	im2txt	txt2im	Average
PLS [16]	0.207	0.192	0.199
KPLS [15]	0.260	0.201	0.231
SCM [13]	0.277	0.226	0.252
GMMFA [17]	0.264	0.231	0.248
GMLDA [17]	0.272	0.232	0.252
LGCFL [9]	0.277	0.229	0.253
RL-PLS	**0.316**	**0.238**	**0.277**

Table 2: MAP results on the Wiki database. Images are represented by CNN features.

Methods \ Tasks	im2txt	txt2im	Average
PLS [16]	0.302	0.272	0.287
KPLS [15]	0.374	0.322	0.348
SCM [13]	0.381	0.331	0.356
GMMFA [17]	0.373	0.340	0.357
GMLDA [17]	0.372	0.338	0.355
LGCFL [9]	0.399	0.351	0.375
RL-PLS	**0.425**	**0.363**	**0.394**

In our experiments, we use the Mean Average Precision (MAP) [9] [13] metric to evaluate the performance of the different methods. For Wiki, 5 Gaussian distributions are set for each class. But for Pascal VOC2007, since the number of samples are very limited for some classes, we set the number of Gaussian distributions to 3.

4.3 Results on the Wiki database

The MAPs of the different methods on the Wiki database where images are represented by SIFT features are shown in Tab. 1. From the table, we can find the following scenes:

Firstly, the performance of the supervised methods (SCM, GMMFA, GMLDA and LGCFL) outperforms those of the unsupervised methods (PLS and KPLS) by at least 7.4%. This indicates that the label information can provide the available information to improve the performance.

Secondly, the average MAP of KPLS outperforms 16.1% than PLS, which indicates that mapping original feature into a high dimensional feature space via kernel function can get better performance. The advantage of KPLS also motivates us to construct RL-PLS based on KPLS rather than PLS.

Thirdly, compared with KPLS, RL-PLS obtains 20.0% higher average MAP, which validates the effect of setting the label information as the output variables in RL-PLS. At the same time, the performance of RL-PLS outperforms the state-of-art method LGCFL by 9.5% higher average MAP. Because in RL-PLS, we break up the binary constraints in traditional class indicator matrix used in LGCFL and reserve the local structure and label information in RL-PLS. The results show the validity of RL-PLS.

Furthermore, we also repeat the experiments using CNN image features to verify our methods' high quality performance. The results are shown in Tab. 2. From the table, we can know that compared with using SIFT features, the performance of the different methods is improved when using CNN features. The experiments show the effectiveness of the CNN features. Besides, the results are consistent with the conclusion obtained form Tab. 1. It is mentionable that our methods achieve 5.1% higher average MAP than state-of-art method LGCFL, indicating that RL-PLS is also effective and competitive based on CNN features.

Table 3: MAP results on the PASCAL VOC2007 database.

Methods \ Tasks	im2txt	txt2im	Average
PLS [16]	0.320	0.251	0.286
KPLS [15]	0.347	0.266	0.307
SCM [13]	0.392	0.280	0.336
GMMFA [17]	0.343	0.280	0.311
GMLDA [17]	0.342	0.278	0.310
LGCFL [9]	0.401	0.320	0.360
RL-PLS	**0.429**	**0.358**	**0.393**

4.4 Results on the Pascal VOC2007 database

Tab. 3 shows the results on the Pascal database, from which, we can find as follow:

Firstly, it is remarkable that the average MAP of RL-PLS achieves 39.3%, which outperforms the state-of-art method, the LGCFL, for more than 9.2%. This verifies the effectiveness of RL-PLS again.

Secondly, the texts in the Pascal database have tags information, just using several words rather than paragraphs used by Wiki database. Under the condition, RL-PLS also gains the competitive results on the Pascal database, which indicates that RL-PLS is also viable on the texts constructed just by several words.

To sum up, we outperform above related methods and achieve competitive performance on two databases. Because of the advantages of RL-PLS, we draw the conclusion that RL-PLS can improve the performance on cross-modal retrieval via reducing the diversity in common space, through reserving the class information and local structure.

5. CONCLUSION

This paper proposes a novel method for the task of cross-modal retrieval. In our approach, we regard the label information as the assistant modality and construct two KPLSs. Furthermore, we design a novel class indicator matrix with real values which combines the class information and the local structure. Experiments carried out on two public databases show that our proposed method performs against the existing competitive methods.

There are several aspects to be further studied in the future. For example, we will look for more effective class indicator matrix to keep the local structure more well. We also can extend the proposed method to handle the multi-label and unpaired setting cross-modal retrieval task.

6. ACKNOWLEDGMENTS

This work was supported in part by National Basic Research Program of China (973 Program): 2012CB316400 and 2015CB351802, 863 Program of China: 2014AA015202, and Natural Science Foundation of China (NSFC): 61332016, 61572465 and 61303160. The authors also thank the suggestion and help from Liang Zhang and Chunfeng Wang.

7. REFERENCES

[1] J. Costa P, E. Coviello, G. Doyle, N. Rasiwasia, G. Lanckriet, R. Levy, and N. Vasconcelos. On the role of correlation and abstraction in cross-modal multimedia retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3):521–535, 2014.

[2] B. Ding and R. Gentleman. Classification using generalized partial least squares. *Journal of Computational and Graphical Statistics*, pages 280–298, 2012.

[3] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge 2007 (voc 2007) results. 2008.

[4] Y. Gong and S. Lazebnik. Iterative quantization: A procrustean approach to learning binary codes. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 817–824, 2011.

[5] G. Guo and X. Wang. A study on human age estimation under facial expression changes. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2547–2553, 2012.

[6] M. Haj, J. Gonzalez, and L. Davis. On partial least squares in head pose estimation: How to simultaneously deal with misalignment. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 2602–2609, 2012.

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

[8] Y. Jia, M. Salzmann, and T. Darrell. Learning cross-modality similarity for multinomial data. *Proc. IEEE International Conference on Computer Vision*, pages 2407–2414, 2011.

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

[10] A. Li, S. Shan, X. Chen, and W. Gao. Maximizing intra-individual correlations for face recognition across pose differences. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 605–611, 2009.

[11] A. Li, S. Shan, X. Chen, and W. Gao. Cross-pose face recognition based on partial least squares. *Pattern Recognition Letters*, 32(15):1948–1955, 2011.

[12] X. Mao, B. Lin, D. Cai, X. He, and J. Pei. Parallel field alignment for cross media retrieval. *ACM Multimedia*, pages 897–906, 2013.

[13] N. Rasiwasia, J. C. Pereira, E. Coviello, G. Doyle, G. R. G. Lanckriet, R. Levy, and N. Vasconcelos. A new approach to cross-modal multimedia retrieval. *ACM Multimedia*, pages 251–260, 2010.

[14] R. Rosipal and N. Krämer. Overview and recent advances in partial least squares. *Subspace, latent structure and feature selection*, pages 34–51, 2006.

[15] R. Rosipal and L. J. Trejo. Kernel partial least squares regression in reproducing kernel hilbert space. *The Journal of Machine Learning Research*, 2:97–123, 2002.

[16] A. Sharma and D. W. Jacobs. Bypassing synthesis: Pls for face recognition with pose, low-resolution and sketch. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 593–600, 2011.

[17] A. Sharma, A. Kumar, H. D. III, and D. W. Jacobs. Generalized multiview analysis: A discriminative latent space. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 2160–2167, 2012.

[18] J. Tang, H. Wang, and Y. Yan. Learning hough regression models via bridge partial least squares for object detection. *Neurocomputing*, 152:236–249, 2015.

[19] J. B. Tenenbaum and W. T. Freeman. Separating style and content with bilinear models. *Neural computation*, 12(6):1247–1283, 2000.

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

[21] J. Wang, S. Kumar, and S. Chang. Semi-supervised hashing for large-scale search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(12):2393–2406, 2012.

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

[23] X. Wang, V. Ly, G. Guo, and C. Kambhamettu. A new approach for 2d-3d heterogeneous face recognition. *IEEE International Symposium on Multimedia*, pages 301–304, 2013.

[24] L. Xie, P. Pan, and Y. Lu. A semantic model for cross-modal and multi-modal retrieval. *ACM International Conference on Multimedia Retrieval*, pages 175–182, 2013.

[25] Y. Zhuang, Y. F. Wang, F. Wu, Y. Zhang, and W. Lu. Supervised coupled dictionary learning with group structures for multi-modal retrieval. *AAAI Conference on Artificial Intelligence*, 2013.