CONTINUUM REGRESSION FOR CROSS-MODAL MULTIMEDIA RETRIEVAL

Yongming Chen, Liang Wang, Wei Wang, Zhang Zhang

National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

ABSTRACT

Understanding the relationship among different modalities is a challenging task. The frequently used canonical correlation analysis (CCA) and its variants have proved effective for building a common space in which the correlation between different modalities is maximized. In this paper, we show that CCA and its variants may cause information dissipation when switching the modals, and thus propose to use the continuum regression (CR) model to handle this problem. In particular, the CR model with a fixed variance coefficient of 1/2 is adopted here. We also apply the multinomial logistic regression model for further classification task. To evaluate the CR model, we perform a series of cross-modal retrieval experiments in terms of two kinds of modals, namely image and text. Compared with previous methods, experimental results show that the CR model has achieved the best retrieval precision, which demonstrates the potential of our method for real internet search applications.

Index Terms— Canonical correlation analysis, continuum regression, partial least squares regression, cross-modal retrieval

1. INTRODUCTION

Human beings are living in the era of information massive explosion. It is quite easy and convenient to obtain desired information using internet search engines, such as articles, pictures, music, and movies. Yet current search engines mainly employ the keyword-based retrieval techniques, and can seldom find the desired text from an image query and vice versa because it is very difficult for these engines to fully understand the relationship among different modals. This causes a challenging task for scientists and engineers to bridge the gap. Recently, some researchers have made great efforts to develop new algorithms in order to enhance the capabilities of search engines for cross-modal retrieval [1-5].

Canonical correlation analysis (CCA) is one popular method in the cross-modal retrieval literature since it can be used to correlate different media data. CCA uses correlation projection to reduce the dimensionalities of two blocks of data. After the joint dimensionality reduction by CCA, the projection subspaces are isomorphic and it is natural to exchange different modalities in such isomorphic subspaces using implicit identity mapping [1]. However, in this process some detailed features representing the original modalities may be ignored and the retrieval precision will accordingly be influenced. CCA-related mixture models also inevitably have such problems.

In order to avoid information dissipation in the process of different modal correlations, the reasonable relations of the CCA subspaces must be built. The partial least squares regression (PLSR) [6] is proposed to solve this problem based on the regression concept. The PLSR approach uses the least square method to correlate the subspaces of CCA. The classical and popular PLSR models are usually divided into two cases: PLS1 and PLS2. PLS1 is a special case of PLS2 containing only one response variable [6]. Generally, the idea behind PLS regression derives from the ideas of ordinary least square (OLS) and principal component regression (PCR) [7]. Recently some theoretical studies discover a continuum regression (CR) model, which unifies those previous linear regression expressions into a simplified mathematic form in terms of regularization of the variances and shrinkage properties [8, 9]. In this paper, the CR model will be used in the modal switch. Note that the performance of modal switch influenced by the variance coefficient of CR is not yet concerned here. We only use the PLS2 model to demonstrate some numerical experiments for cross-modal retrieval, which is a special application case of CR with a variance coefficient of 1/2. Current study just involves two different-modal media data, i.e., text and image. The main aims of this paper are to introduce the CR model for the application of cross-modal retrieval and demonstrate the effectiveness of the proposed method.

The remainder of this paper is organized as follows. Section 2 introduces the modal switch methods. Section 3 gives the experimental results and analysis. The conclusion is drawn in Section 4.

2. MODAL SWITCH METHODS

In this section, a novel method of CR for modal switch is introduced. Let $\mathbf{X} \in \Re^{n \times p}$ and $\mathbf{Y} \in \Re^{n \times q}$ be two blocks of data matrices, n is the number of data samples, p and q are the number of dimensionalities of $\mathbf{X}$ and $\mathbf{Y}$ in original space, respectively. For convenience, the columns are zero mean.

These two blocks of data can be correlated by using a classical linear regression model as

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E} \quad (1)$$

where $\mathbf{B} \in \Re^{p \times q}$ is the regression coefficient matrix and $\mathbf{E} \in \Re^{n \times q}$ is the residual matrix. In the application of cross-modal retrieval, it is rather convenient to transform one modality $\hat{\mathbf{X}}$ to another $\hat{\mathbf{Y}}$ by multiplying the regression coefficient **B**:

$$\hat{\mathbf{Y}} = \hat{\mathbf{X}}\mathbf{B} \quad (2)$$

The PLSR method is usually used for maximizing the correlations of the above two blocks of data matrices [6, 7]. The optimization objective of the PLSR models (including CCA mode) is to find the maximal covariance between projection subspaces **T** and **U**, but the regression relations between them are not considered. To correlate projection subspaces, a simple regression function can be employed

$$\mathbf{U} = \mathbf{TD} + \mathbf{F} \quad (3)$$

where $\mathbf{D} \in \Re^{m \times m}$ is the subspace regression coefficient matrix and $\mathbf{F} \in \Re^{n \times m}$ is the subspace residual matrix. Subspace regression relation can be mapped to the original space as follows:

$$\mathbf{Y} = \mathbf{XB}_{\text{PLS}} + \mathbf{E} = \mathbf{XX}^T\mathbf{U}\left(\mathbf{T}^T\mathbf{XX}^T\mathbf{U}\right)^{-1}\mathbf{T}^T\mathbf{Y} + \mathbf{E} \quad (4)$$

where $\mathbf{B}_{\text{PLS}} = \mathbf{X}^T\mathbf{U}\left(\mathbf{T}^T\mathbf{XX}^T\mathbf{U}\right)^{-1}\mathbf{T}^T\mathbf{Y}$ is the regression coefficient matrix of original spaces.

Recently, the PLSR model has proven to be a special case of the CR model. The CR model [8, 9] finds the projection matrix by solving the following problem

$$\max_{\|\mathbf{p}\|=1}\left\{\text{Cov}\left(\mathbf{Xp}, \mathbf{y}\right)^2 \cdot \text{Var}\left(\mathbf{Xp}\right)^{\frac{\sigma}{1-\sigma}-1}\right\} \quad (5)$$

where **p** is the projection vector and $0 \le \sigma \le 1$. Then the OLS method is performed for regression:

$$\mathbf{y} = \mathbf{XB}_{\text{CR}} + \mathbf{E} = \mathbf{XP}\left(\mathbf{T}^T\mathbf{T}\right)^{-1}\mathbf{T}^T\mathbf{y} + \mathbf{E} \quad (6)$$

where **P** is the projection coefficients of the subspace **T**, i.e., **T**=**XP**. This concept is the same as the PCR techniques. The CR model has proven equivalent to OLS, PLSR and PCR for σ = 0, 1/2 and 1, respectively. The CR model can be also extended to deal with the case of multivariate **Y** by the joint continuum regression and the continuum power regression techniques [8, 9].

More models beyond the OLS, PLSR and PCR can be obtained by tuning parameter σ in the CR algorithm, which is more flexible in model fitting and prediction. It should be noted that, how to choose an optimal σ for cross-modal correlation and the influence of correlation performance for different σ are not considered in this work. The aims of this paper are to introduce the CR model to bridge the gap between different modalities and use a special case of the CR model with σ =1/2 (which is equivalent to the PLSR) to demonstrate both applicability and effectiveness of the CR model for the task of cross-modal retrieval.

3. EXPERIMENTS

In this section we carry out a series of experiments in terms of both uni-modal and cross-modal retrieval tasks and make comparison of our method with the state-of-the-art methods.

3.1. Evaluation dataset

To evaluate our method, we choose the dataset used in [1] which consists of 2866 Wikipedia documents with paired text and images (see Fig. 1 for examples). These documents are classified into ten categories, namely art & architecture, biology, geography & places, history, literature & theatre, media, music, royalty & nobility, sport & recreation, and warfare. These documents are randomly divided into the training and test set consisting of 2173 and 693 documents, respectively.

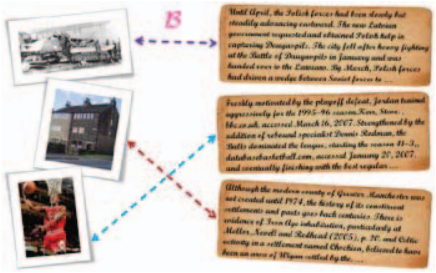


Fig. 1 An instance showing several paired text and images in the evaluation database.

3.2. Experimental settings

We adopt the latent Dirichlet allocation (LDA) and scale-invariant feature transform (SIFT) to extract effective features for representing text and image, respectively [1]. In order to evaluate the uni-modal retrieval performance we apply the 10-topic LDA text representation for the text query and the 128-codeword SIFT image representation for the image query. The multinomial logistic regression algorithm [10] is used for both modeling and prediction. The logistic regression algorithm generates 10-dimensional probability simplexes for each observation, and the category is specified according to the maximal probability.

We evaluate the cross-modal retrieval performance using the proposed CR model. This method uses the LDA feature representation for the image query and the SIFT feature representation for the text query. Then the logistic regression algorithm is used for the classification purpose.

We also compare our method with the uni-modal and other previous cross-modal methods including correlation matching (CM), semantic matching (SM), and semantic

correlation matching (SCM), all of which are proposed in [1]. To describe our cross-modal method conveniently, we name it original matching (OM).

The precision-recall (PR) curves, average precision (AP) scores and confusion matrices are respectively used to rank the retrieval performance. All the following illustrations for the classification results refer to the highest AP scores.

3.3. Uni-modal retrieval results

The uni-modal retrieval results can provide a baseline for the cross-modal retrieval. The uni-modal PR curves and confusion matrices are shown in Fig. 2 (a) and (b) and Fig. 3(e1) and (e2), respectively. The AP scores computed by the logistic regression method on both training and test sets are listed in Table 1. Note that the classification accuracy of using the LDA-based text features is much higher than that of the SIFT-based image features. It is not surprising that, even for a person, classification of images is much harder than that of text.

Table 1. Uni-modal AP scores

Dataset	Image query	Text query	Average
Training set	0.366	0.890	0.628
Test set	0.320	0.674	0.497

3.4. Cross-modal retrieval results

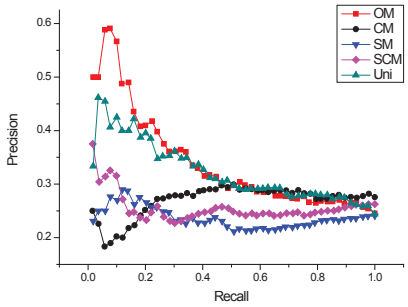
This subsection shows the performance of our cross-modal retrieval method. The PR curves and confusion matrices are shown in Fig. 2 (a) and (b) and Fig. 3 (a1) and (a2), respectively. The AP scores computed by the logistic regression method on the test set are also listed in Table 2. Similar to the uni-modal retrieval results, the classification accuracy of using the LDA features is much higher than that of using the SIFT features.

Table 2. Cross-modal AP scores for the test set

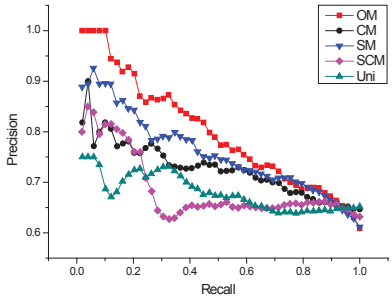
Matching method	Image query	Text query	Average
OM	**0.334**	**0.782**	**0.558**
CM	0.265	0.711	0.488
SM	0.235	0.743	0.489
SCM	0.253	0.673	0.463

3.5. Comparison

In this part, the retrieval performance of various methods is compared and discussed. Fig. 2 shows the PR curves of all the retrieval approaches for both image and text queries. Comparing the PR curves of the uni-modal approach with the cross-modal (CM and SM) approaches for the text query (Fig. 2(b)), it is obvious that the retrieval precision of CM and SM is better than that of the uni-modal, which means CM and SM can be used to switch modalities of the text to the image. However, for the image query (Fig. 2(a)), it can be observed that the retrieval precisions of the cross-modal approaches (CM, SM and SCM) are much lower than that of the uni-modal, suggesting these approaches not very robust. Comparing the PR curves of our OM approach with the other four, it is clear that our retrieval results are very encouraging, where the OM obtains the highest retrieval precision for both image and text queries. This superior retrieval performance indicates that our OM approach can be used to effectively and reasonably interpret the relationship between different multimedia modalities.



(a)



(b)

Fig. 2 PR curves of uni-modal and cross-modal approaches for the image (a) and text (b) queries, respectively

The AP scores computed by four cross-modal retrieval approaches are also summarized in Table 2. Looking at the experimental results in Table 1 and 2, the average AP score of the uni-modal retrieval is 0.497, which is 1.7% higher than the best CCA-based cross-modal approach's 0.489. It means some useful information is missing in the process of modal transformation. The retrieval performance of our OM is greatly improved compared with previous approaches. The average AP score reaches 0.558, which is 10.9% better than the uni-modal. These indicate that the subspaces with the same dimensions projected by CCA-based approaches cannot be directly associated through simple identity mapping (see [1]), since there are certain latent relations between them. The CR approach can prevent information

loss by using the regression technique and performs best in finding the correlation of the two blocks of variables.

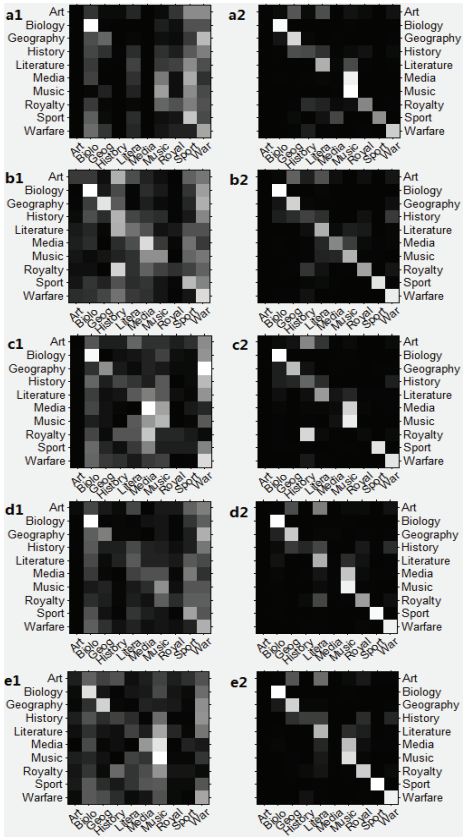


Fig. 3 Confusion matrices of cross-modal (OM, CM, SM and SCM) and uni-modal retrieval approaches for the image (a1-e1) and text (a2-e2) queries.

Confusion matrices (shown in Fig. 3) are also applied to visually compare the retrieval performance. The confusion matrix plots of the image and the text queries are listed on the left and right shown in Fig. 3, respectively. The confusion matrix plots of the former are more confusing than that of the latter. In addition, the confusion matrix plots of cross-modal approaches (OM, CM, SM and SCM) are similar to those of the uni-modal approach for the image and text queries. These illustrate that the classification precision of the model depends on the modality itself. Thereby, one important cross-modal retrieval mission is to prevent information dissipation in the modal switch. For text queries, it is hard to select a best cross-modal approach from only the confusion matrix plots (Fig. 3(a2)-(d2)), since they are basically similar to each other. However, for image queries, the confusion matrix plot of OM (Fig. 3(a1)) is cleaner than that of CM, SM and SCM (Fig. 3(b1)-(d1)). It indicates that the amount of information dissipation of OM in the modal switch is less than that of CM, SM and SCM. It can be inferred that the CR approach establishes more reasonable and reliable relations between two different modalities than the CCA-based approaches.

4. CONCLUSION

In this paper, we have proposed to use the CR model for enhancing cross-modal retrieval. The experimental results have shown that the proposed method obtains the best retrieval performance, indicating its potential application for Internet searching engines.

ACKNOWLEDGMENTS

This work is jointly funded by the National Basic Research Program of China (2012CB316300), National Key Technology R&D Program (2011BAH11B01), Hundred Talents Program of CAS, the Strategic Priority Research Program of the Chinese Academy of Sciences (XDA06030300), and National Natural Science Foundation of China (61135002 and 61175003)

5. REFERENCES

[1] 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," *Proceedings of the 18th ACM Conference on Multimedia*, 2010.

[2] H. Ohkushi, T. Ogawa, and M. Haseyama, "Kernel CCA-based Music Recommendation According to Human Motion Robust to Temporal Expansion," *Proceedings of International Symposium on Communications and Information Technologies*, pp.1030-1034, 2010.

[3] Y. Yang, F. Wu, D. Xu, Y.T. Zhuang, and L.T. Chia, "Cross-Media Retrieval Using Query Dependent Search Methods," *Pattern Recognition*, vol. 43, pp. 2927–2936, 2010.

[4] Y.T. Zhuang, Y. Yang, and F. Wu, "Mining Semantic Correlation of Heterogeneous Multimedia Data for Cross-Media Retrieval," *IEEE Transactions on Multimedia*, vol. 10, no. 2, pp. 221-229, 2008.

[5] Z.C. Li, J. Liu, and H.Q. Lu, "Sparse Constraint Nearest Neighbour Selection in Cross-Media Retrieval," *Proceedings of the 17th International Conference on Image Processing*, pp. 1465-1468, 2010.

[6] J.A. Wegelin, "A Survey of Partial Least Squares (PLS) Methods, with Emphasison the Two-Block Case," *Technical Report, Department of Statistics, University of Washington, Seattle,* pp. 1-44, 2000.

[7] R. Rosipal, and N. Krämer, "Overview and Recent Advances in Partial Least Squares," *Lecture Notes in Computer Science*, vol. 3940, pp. 34-51, 2006.

[8] S. Serneelsa, P. Filzmoserb, C. Crouxc, and P.J. Van Espe, "Robust Continuum Regression," *Chemometrics and Intelligent Laboratory Systems*, vol. 76, pp. 197-204, 2005.

[9] S. Bougearda, M. Hanafi, and E.M. Qannari, "Continuum Redundancy–PLS Regression: A Simple Continuum Approach," *Computational Statistics and Data Analysis*, vol. 52, pp. 3686–3696, 2008.

[10] R.E. Fan, K.W. Chang, C.J. Hsieh, X.R. Wang, and C.J. Lin, "LIBLINEAR: A Library for Large Linear Classification," *Journal of Machine Learning Research,* vol. 9, pp. 1871-1874, 2008.