2010 Sixth International Conference on Intelligent Information Hiding and Multimedia Signal Processing

Kernel Canonical Correlation with Similarity Refinement for Automatic Image Tagging

Yanhui Xiao
Institute of Information Science
Beijing Jiaotong University
Beijing, China
xiaoyanhui@gmail.com

Yao Zhao
Institute of Information Science
Beijing Jiaotong University
Beijing, China
yzhao@bjtu.edu.cn

Zhenfeng Zhu
Institute of Information Science
Beijing Jiaotong University
Beijing, China
zhfzhu@bjtu.edu.cn

***Abstract*—Automatic image tagging (AIT) is an effective technology to facilitate the process of image retrieval without requiring user to provide a retrieval instance beforehand. In this paper, we propose an AIT method based on kernel canonical correlation analysis (KCCA) with similarity refinement (KCCSR). As a statistic correlation technique, the KCCA aims at extracting some kind of hidden information shared commonly by the two random variables. Different from the previous KCCA based tagging methods, the graph based similarity refinements are first implemented by an interactive way to obtain the enhanced visual and textual representations. Subsequently, the KCCA is applied to them to mine the unique intrinsic semantic representation space, in which the AIT can be completed. The final experimental results validate the effectiveness of the proposed KCCSR.**

Keywords-CBIR; automatic image tagging; KCCA; similarity refinement

I. INTRODUCTION

Nowadays, the digital images have become widely available on World Wide Web (WWW) due to the popularity of digital cameras and online community such as Flickr and Picasa Web Album. It has brought about great challenges for managing and retrieving a large number of available images. As is well known, the content based image retrieval (CBIR) is an effective image retrieval technique. But the low level features, such as color, edge and texture, can't fully represent the exact semantic content of the images in CBIR owing to the insurmountable semantic gap between the human beings and computer vision. One solution for this problem is to manually tag each image with keywords, which assign multiple semantic contents for each image. However, manual tagging is tedious and even unrealizable in some circumstance, especially for the large image database. Therefore, automatic image tagging (AIT) has become a focus and received massive investigations on it while still keeps the advantages of semantically tagging. In recent years, many approaches have been proposed for AIT. The representative work includes: Translation Model [1], Latent Dirichlet Allocation Model (LDA) [2], Cross-Media Relevance Model (CMRM) [3], Continuous Relevance Model (CRM) [4] and Multiple-Bernoulli Relevance Model (MBRM) [5]. Despite the continuous efforts put on image tagging, the results of existing AIT methods are still unsatisfactory.

Recently, refinement based tagging technique has received some extensive focuses. Wang et al. presented Random Walk with Restarts Model (RWRM) [6] to refine the candidate tags by employing the co-occurrence of words. Moreover, Jia utilized multi-graph similarity reinforcement (MGSR) method [7] to boost the learning of word correlation by exploiting the consistency of image contents and their associated tags. But MGSR hasn't considered the local visual representation for measuring the similarity of images. In real application, it is difficult to decide which kind feature is more effective for AIT. Therefore, it is straightforward to leverage both local and global features as they can complement each other. However, it keeps a nontrivial work on how to fuse the word correlation, local and global visual similarity to improve the performance of AIT.

The KCCA can address this problem to some extent by learning the representation of image which is more closely to the latent semantics and mapping both visual and textual descriptions into the semantic space with the minimal distance. Moreover, the KCCA is effective to extract nonlinear discriminative features of input samples by using the kernel trick [11, 12]. But the previous KCCA model [8] just employed simply some low level visual features and plain textual representation from labeled training samples to seek their correlation space, without taking large number of unlabeled samples into in depth consideration. To alleviate this limitation, we propose the KCCSR algorithm for AIT. The KCCSR utilizes the global and local visual similarity propagation method to construct the image visual similarity graph for enhancing the ability of view of visual description. Meanwhile, we employ the co-occurrence of two tags in the training dataset to build word correlation graph for the representation of textual view. Furthermore, the KCCSR maps both visual and textual description into the semantic space with maximal correlation.

II. KCCSR BASED TAGGING SCHEME

Figure 1 illustrates the overview of our KCCSR method and the overall work is composed of three parts. First, the system extracts image visual and textual features of training samples. Meanwhile, global and local features are extracted to estimate a query aware visual similarity graph by the propagation of local and global similarity. And the tagged training images are utilized to explore the word correlation

978-0-7695-4222-5/10 $26.00 © 2010 IEEE
DOI 10.1109/IIHMSP.2010.145

graph based on the co-occurrence of two words. Second, the two views, which are the visual and textual descriptions of the training data, are refined by word correlation and visual similarity graphs. And the KCCA algorithm is employed to solve the two projection vectors with maximal correlation at the latent semantic space for two views. At last, given a query image q the KCCSR can output the highest scoring tags.

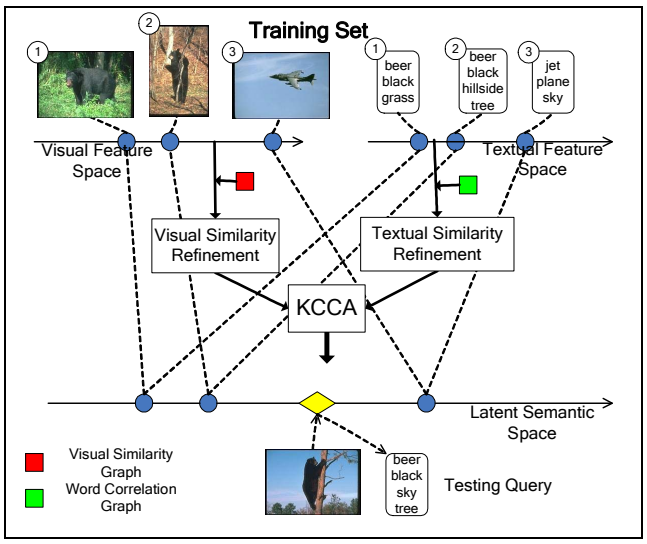

Figure 1. Overview of our KCCSR method.

A. Image Visual Similarity and Word Correlation Graphs

The image visual similarity is usually estimated with only local or global features which cannot be fully represented for all the images. The similarity graph can be constructed by incorporating the global visual similarity, local visual similarity and visual word co-occurrence and combines the local and global similarity and gets both the merits in [9]. The iterative propagation process is modeled as,

$$\begin{cases} G^{(t+1)} = \mu G^{(t)} + (1-\mu)\lambda Z K^{(t)} Z^T \\ K^{(t+1)} = \nu K^{(t)} + (1-\nu)\lambda Z^T G^{(t)} Z \end{cases}, \quad (1)$$

where G and Z are global and visual similarity graphs respectively, K is visual word co-occurrence matrix, μ and ν are trade-off parameter and λ is a decay factor. For convergence, G and K are normalized graphs for each iteration. The final visual similarity graph G is utilized to get the kernel matrix as a graph kernel for the visual views of images.

The tags of the images often have high co-occurrence with each other while the unrelated ones are often isolated. Based on this observation, the word correlation graph is built as follows [6], and which is based on the co-occurrence of the two tags in the training dataset. The correlation of the tags is directly proportional to the times of two tags co-tagging of training image. Let $W(w_i, w_j)$ be the correlation of words w_i and w_j. That is,

$$W(w_i, w_j) = \begin{cases} 0 & num(w_i, w_j) = 0 \\ \dfrac{num(w_i, w_j)}{\min(num(w_i), num(w_j))} & num(w_i, w_j) \neq 0 \end{cases} \quad (2)$$

,

where $num(w_i, w_j)$ is the number of images tagged by both words w_i and w_j, $num(w_i)$ and $num(w_j)$ are the number of images tagged with words w_i and w_j respectively. The word correlation graph W is used to refine the textual views of training images.

B. Kernel Canonical Correlation Analysis

The CCA [10] is a statistical tool that can be used to identify the correlated projections between two views. And CCA attempts to find two sets of basis vectors w_x and w_y, one for each view, such that the correlation between the projections of these two views into the basis vectors are maximized.

Let $S = \begin{bmatrix} (x_1, y_1) \\ \vdots \\ (x_l, y_l) \end{bmatrix}$, $S_x = \begin{bmatrix} x_1 \\ \vdots \\ x_l \end{bmatrix}$ and $S_y = \begin{bmatrix} y_1 \\ \vdots \\ y_l \end{bmatrix}$ be the sample of size l where $x_i \in R^m$ and $y_i \in R^n$, $l \times m$ matrix and $l \times n$ matrix respectively, are the two views of the same instance S. For example, S can be an event, x and y can be a photograph and textual descriptions of S.

To overcome the drawback of linear CCA in extracting nonlinear correlation spaces, the kernel CCA [8], i.e. kernel canonical correlation analysis (KCCA), offers an alternate solution by first projecting the input space into a higher dimensional feature space by a nonlinear mapping function.

Let ϕ be the feature space mapping of x view and φ be the feature space mapping of y. Denote $\phi(S_x) = \{\phi(x_1), ..., \phi(x_l))\}$ the feature space mapping applied to x view of the sample and $\varphi(S_y) = \{\varphi(y_1), ..., \varphi(y_l))\}$ as the feature space mapping of y view of the sample. We can now define a kernel function $K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$ where $\langle \cdot, \cdot \rangle$ is the dot product (similarly for y). We can use $K_x = \phi(S_x)^T \phi(S_x)$ and $K_y = \varphi(S_y)^T \varphi(S_y)$ as the kernel matrices for x and y views of instances.

The projection vectors w_x and w_y can be written as $w_x = \alpha^T S_x^T$ and $w_y = \beta^T S_y^T$ ($\alpha, \beta \in R^l$). Thus, the

objective function ρ becomes

$$\rho = \arg\max_{\alpha,\beta} \frac{a^T S_x^T S_x S_y^T S_y \beta}{\sqrt{a^T S_x^T S_x S_x^T S_x \alpha \cdot \beta^T S_y^T S_y S_y^T S_y \beta}}.$$

$$w.r.t. \begin{cases} \alpha^T S_x^T S_x S_x^T S_x \alpha = 1 \\ \beta^T S_y^T S_y S_y^T S_y \beta = 1 \end{cases}$$

The inner products in two hidden spaces S_x and S_y are two corresponding kernel matrices $K_x = S_x^T S_x$ and $K_y = S_y^T S_y$ respectively as defined above. By some mathematical manipulation, the coefficient vector α can be solved from Eq.3.

$$(K_x + \kappa I)^{-1} K_y (K_y + \kappa I)^{-1} K \alpha = \lambda^2 \alpha \quad (3)$$

The matrix I is the identity matrix and κ is used for in case of the singularity of kernel matrix K_y. Then, corresponding β can be obtained by Eq.4.

$$\beta = \frac{1}{\lambda}(K_y + \kappa I)^{-1} K_x \alpha \quad (4)$$

C. The KCCSR Algorithm

The KCCSR algorithm employs visual similarity and word correlation graph to refine the semantic of images on kernel canonical correlation subspace. The algorithm of KCCSR is described in Table I.

TABLE I. THE KCCSR ALGORITHM

Input:

S_x, S_y : the training matrix of x-view and y-view

Process:

1. Build the image visual similarity and word correlation graphs G and W using Eq.1 and Eq.2.
2. Execute the KCCA algorithm with graph G and W to get weight matrices α and β by Eq.3 and Eq.4.
3. Obtain the kernelised inner products K_x^q between the query image q and the training images and refined y-view S_y^T
4. Compute the weights for the tags $w_y = \beta^T S_y^T$ and the score for KCCA using $s = w_y \alpha^T K_x^q$.

Output: highest scoring tags for query q

First of all, we need to build the word correlation graph and image visual graph G and W using Eq.1 and Eq.2 with the training images. While the visual similarity graph G is utilized to get the kernel matrix K_x as a graph kernel and the word correlation graph W is used to refine the y-view of training matrix to get the updated view $S_y = S_y \times W$. Subsequently, the regularized KCCA algorithm is proposed to get the projection coefficients α and β by Eq.3 and Eq.4. Furthermore, given a query q, we use image visual similarity graph G to get the kernelised inner products between the query image q and the images in the training K_x^q and compute the weights for the tags $w_y = \beta^T S_y^T$ and the score for KCCA using $s = w_y \alpha^T K_x^q$. Finally, the tags are transferred for each query image from the training set while each tag is assigned a score, and the highest scoring tags are provided as the output.

III. EXPERIMENTAL RESULTS

A. Dataset and Performance Evaluation

We test the proposed KCCSR scheme on the Corel dataset obtained from Barnard et al. [1]. The experimental dataset comprises 5,000 images, of which 4,500 images are used as training set and the remaining 500 images as testing set. Each image is annotated with 1 to 5 tags, and totally 374 tags have been used in the tagging.

We extract the global features for each image, including color histograms of RGB and LAB with 16 bins in each color channel and 512-dimensional Gist features which can be used to characterize the spatial arrangement of image. For the local visual representation, we use SIFT as the local descriptor with each local feature descriptor quantized using k-means on samples from the training set. Thus, a 'bag of words' of 1000-dimensional histogram is built for each image.

The AIT performance is evaluated by the average recall, average precision and F-measure. They are denoted by precision=B/A and recall=B/C, where A is the number of images automatically annotated with a given keyword; B is the total of images correctly annotated with that keyword; and C is the number of images having that keyword in ground truth annotation. Following the definition in [7], F-measure is used for evaluating the performance and given by F-measure = 2×Precision×Recall / (Precision + Recall).

B. Tagging Performance

We tag each test image with the 5 most relevant tags. To evaluate the quality of the proposed KCCSR scheme, we compare it with other tagging models: MBRM [5], MGSR [7] and KCCA [8] model (without using the similarity refinement). The performances of them are given in Table II.

As we can see from Table II, the proposed KCCSR scheme achieves competitive results. KCCSR based tagging method takes the best performance on all metrics. Specially, compared with the previous better result (MBRM), KCCA and KCCSR have more words with positive recall. The reason is that KCCA can learn the representation of image which is more closely to the latent semantics and map both visual and textual descriptors into the semantic space with the maximal correlation.

TABLE II. COMPARISON WITH OTHER RELATED ANNOTATION MODELS

Model	MBRM	MGSR	KCCA	KCCSR
N+	122	—	151	157
Results on all 374 keywords				
Average Precision	0.24	—	0.24	0.25
Average Recall	0.25	—	0.29	0.32
F-measure	0.245	0.268	0.267	0.281

—The author didn't provide the data for this measure. N+ denotes the number of tags with non-zero recall value.

				
Ground Truth	beach palm people tree	flowers sky tree tulip	herd plane tree zebra	foals horses mare tree
KCCSR	palm beach tree people sand	flowers sky tree tulip trail	tree herd zebra plane grass	foals horses mare plants tree
				
Ground Truth	mountain sky tree water	cars formula tracks wall	iguana lizard marine rocks	bear grass polar tundra
KCCSR	tree mountain sky water buildings	cars formula wall tracks beer	iguana marine lizard sand rocks	bear grass tundra polar detail

Figure 2. Good examples for image tagging.

Compared with KCCA model, KCCSR is on the increase for the performance. That can be analyzed that KCCSR leveraging the word correlation and image visual similarity can obtain the better visual and textual views for the projection. Figure 2 presents some good examples of the predicted tags and ground truth. It is clear to find the images tagged well, and KCCSR can enhance the effective tags for the ground truth. For example, the first image in Figure 2 adds the tag 'sand' which is ignored by ground truth.

				
Ground Truth	fox ice river water	cat grass tiger water	jet plane sky smoke	frozen ice snow
KCCSR	water river fox ice shadows	grass tiger cat sand water	sky snow plane flag	ice bear snow polar tree
	(a)		(b)	

Figure 3. Bad examples for image tagging.

However, in Figure 3.a the highest scored tags are inconsistence with our visual about the content of the images. The reason is that the KCCSR lacks of weighting in the visual feature of humans interests. For instance, we prefer to 'fox' in the first image and 'tiger' in the other one rather than 'water' and 'grass' in the first image of Figure 3.a. Furthermore, the tagging results in Figure 3.b are not satisfactory to us. It is because that there is no significant difference in the aspect of low level visual representation, which causes the ambiguity in understanding of high level semantic concept such as the 'bear' and 'snow'.

IV. CONCLUSIONS

In this paper, we proposed a KCCSR method to refine the tags of image. The advantages of the proposed KCCSR lie on two aspects. On the one hand, the correlation between the two views with the minimal distance is used to learn a direct map from image descriptors to tags. On the other hand, the word correlation and image visual similarity can refine the representation of images in kernel canonical correlation subspace. The experiments conducted on Corel dataset have demonstrated the effectiveness of the proposed KCCSR algorithm.

ACKNOWLEDGMENTS

This work was supported in part by National Natural Science Foundation of China (No.60776794), PCSIRT (No.IRT707), Sino-Singapore JRP (No. 2010DFA11010), Fundamental Research Funds for Central Universities (No.2009JBZ006) and Open Foundation of NLPR.

REFERENCES

[1] Duygulu, P. and Barnard, K. "Object recognition as machine translation: learning a lexicon for a fixed image vocabulary", In Proceedings of ECCV, 2002.

[2] Blei, D. M. and Jordan, M. I. Modeling annotated data. In Proc. SIGIR, Toronto, July. 2003.

[3] Jeon, J., Lavrenko, V., and Manmatha, R., "Automatic Image Annotation and Retrieval Using Cross-media Relevance Models", In Proceeding of SIGIR, Toronto, July 2003.

[4] Lavrenko, V., Manmatha, R., and Jeon, J. A Model for Learning the Semantics of Pictures. In Proc. NIPS, 2003.

[5] Feng, S., Manmatha, R., and Laverenko, V. "Multiple Bernoulli Relevance Models for Image and Video Annotation", CVPR, pages.1002-1009, 2004.

[6] Wang, F. Jing, L. Zhang, H.J. Zhang. "Image Annotation Refinement using Random Walk with Restarts". In Proceedings of ACM Multimedia, 2006.

[7] J. Jia, N. Yu, X. Rui, and M. Li. "Multi-graph similarity reinforcement for image annotation refinement". In 15th IEEE International Conference on Image Processing. ICIP, pages 993–996, 2008.

[8] D. Hardoon, C. Saunders, S. Szedmak, and J. Shawe-Taylor. "A correlation approach for automatic image annotation", Proc. of Second International Conference on Advanced Data Mining and Applications, ADMA, pages 681–692, China, 2006.

[9] Wang, L.J. Yang, X.M. Tian, "Query aware visual similarity propagation for image search reranking". ACM Multimedia, 725-728, 2009.

[10] H. Hottelling. "Relations between two sets of variates". Biometrika, 8:321–377, 1936.

[11] Jeng-Shyang Pan, Junbao Li and Zheming Lu, "Adaptive Quasiconformal Kernel Discriminant Analysis", Neurocomputing, Vol. 71, No. 13-15, pp. 2754-2760, 2008

[12] Junbao Li, Jeng-Shyang Pan and Shu-Chuan Chu, "Kernel Class-wise Locality Preserving Projection", Information Sciences, vol. 178, no. 7, pp. 1825-1835, 2008