AUTOMATIC IMAGE ANNOTATION VIA LOCAL SPARSE CODING

Wenbo Zhang[1,2],Dongping Tian[1,2],Hong Hu[1],Xiaofei Zhao[1] and Zhongzhi Shi[1]

1 Key Laboratory of Intelligent Information Processing,
Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190, China
2 University of the Chinese Academy of Sciences, Beijing, 100049, China
{zhangwb,tiandp,huhong,zhaoxf,shizz}@ ics.ict.ac.cn

ABSTRACT

Sparse coding is an active research topic in machine learning and signal processing community. In this paper, we propose a novel local sparse model for multi-label image annotation. Existing feature descriptors and extraction algorithms pay less attention to semantic information and extracted feature dimension usually is high, which leads to heavy computation. Noise and redundant information often reduce the performance of sparse model. To address these issues, we combine label and visual information for feature selection while most previous work only utilizes labels and ignores visual information itself. First of all, we make use of label sets to seek images neighbor relations and generate the Gaussian kernel matrix over these neighbor images, then use LLP(Local Learning Projection) algorithm to get minimal local estimation error. After that, for each query image, we find its K nearest neighbors in the transformed space and use these neighbors to reconstruct it via sparse coding. Moreover, during coding, we penalize the corresponding reconstruction coefficients to implicitly reflect the neighbor relations. Finally, propagating tags from training data to test data. Image annotation experiments on the Corel5k dataset show the performance of our approach is comparable to several state-of-the-art algorithms.

Index Terms— image annotation, feature selection, KNN, local, sparse coding

1. INTRODUCTION

Automatic image annotation is a hot research topic in computer vision community. The goal of image annotation is to assign a few relevant text keywords to the given input images which reflect their visual content. It's a typical multi-label learning problem in nature, where each image contains multiple objects and therefore be associated with a set of labels.

Image annotation is a difficult task since the well-known *semantic gap* problem. Furthermore, the lack of correspondence between the labels and image visual words makes the task more complex. The image annotation problem has been studied for more than a decade. The popular algorithms can be roughly categorized into three scenarios: classification-based method, probabilistic modeling-based method and nearest neighbor method. The classification-based method treats each label as a class and for each class by constructing a classifier to represent it. The probabilistic modeling-based method attempts to infer the correlations or joint probabilities between images and labels. The nearest neighbor method regards image annotation as image retrieval problem and uses a greedy label transfer mechanism[1]. Non-parameter nearest neighbor method has been found to be quite successful for image annotation.

In recent years, sparse coding method is popular in computer vision research field. Sparse coding assumes that a signal can be efficiently represented by a sparse linear combination of atoms from a given or learnt dictionary.It has been successfully applied to face recognition[2].Sparse model has also been introduced into image annotation recently[3, 4].

In this paper, we present a local sparse model to solve image annotation. Our work is related to [3], but differs from it. In their work, they utilized label information for feature selection whereas ignored visual information itself. We combine the label and visual information for feature selection. Most previous work[3, 4] applied all training data with equal weight to reconstruct testing data. Here, we assume that an image can be sparsely represented by its local images with the weight which exponential delay with distance. Our contributions in this work are the following:

- Combining the label and visual information for feature selection in image annotation while most of existing work just utilizes label information. The motivation of this combination is to decrease the impacts of polysemy and synonymy.
- Reconstructing an unlabeled image just use those images which have neighbor relations with it, instead of all the training images. Classic sparse coding regards all training data as dictionary and reconstruct test data

This work was supported by the National Natural Science Foundation of China(No.61072085,60933004,61035003,60903141),National Program on Key Basic Research Project(973 Program)(No.2013CB329502).

even if it has a large amount of redundant dictionary elements. In image annotation, the large number of training images makes the computation complexity expensive.

- Penalizing the corresponding reconstruction coefficients to embed the local information during coding.

The rest of this paper is organized as follows. Related work is briefly reviewed in section 2. We introduce our method in details in section 3. In section 4, we show experimental results on Corel5k dataset. Finally,we conclude the paper in the last section.

2. RELATED WORK

In this section we give a brief review of the models for automatic image annotation.

As mentioned above, the majority of popular models for automatic image annotation can be roughly divided into three categories: classification-based, probabilities model-based and nearest neighbor methods. The first category regards each visual concept (i.e. label) as a class and trains a classifier for each class. Typical models include Support Vector Machine (SVM)[5, 6], Hidden Markov Model (HMM)[7] and Gaussian Mixture Model (GMM)[3, 8, 9], etc. This family of methods deals with each label independently and less considers the relations between labels. Since the diversity of labels, many labels are so rare that there aren't enough positive samples to train a reliable classifier. Although these models are difficult to learn label distributions accurately, they are usually used with other methods together. For instance, in SML[9], a Gaussian mixture hierarchy was proposed to model the class distribution; in MSC[3], universal Gaussian Mixture Model was presented to encode each image into supervector.

The second category attempts to learn the joint probabilities between images and annotations. Representative work includes Machine Translation (MT)[10], Cross-Media Relevance Model (CMRM)[11], Continuous Relevance Model (CRM)[12], Multiple Bernoulli Relevance Model (MBRM)[13], probabilistic latent semantic analysis (PLSA)[14] and hierarchical Dirichlet process[15], etc. Due to lack of mature and universal image segmentation algorithm to obtain the correspondence between labels and image regions, so these models often use "bag of visual words" model, which inevitably reduces the annotation performance.

Nearest neighbor method treats image annotation as image retrieval problem. Makadia et al. [1] introduce a new baseline technique for image annotation that treats annotation as a retrieval problem. They utilized low-level image features and a simple combination called Joint Equal Contribution (JEC) of basic distances to find nearest neighbors of a given image and used a greedy label transfer mechanism to annotate image. Subsequently, Guillaumin et al. [16] proposed TagProp, a discriminatively trained nearest neighbor model. TagProp allows the integration of metric learning by directly maximizing the log-likelihood of tag predictions in the training set. This kind of methods has good scalability in the number of labels of interest and can achieve very competitive annotation performance.

3. OUR APPROACH

In this section, we present our approach for image auto-annotation in detail. Nearest neighbor methods show that labels of query image are mainly determined by its local neighbors according to the image-to-image distance measure. However, image feature usually has noise and inevitably reduces the performance of annotation. For another, images with more similar labels should be more similar in feature space while original feature does not always. We make use of label and visual information to select features.

3.1. Feature Selection

The goal of feature selection is to learn a linear transformation matrix $P \in R^{d\times p}(p < d)$ to transform data from the original space to a lower-dimensional space, in which the semantic relations can be retained, i.e. $y_i = P^T x_i$. $X = [x_1, x_2, ..., x_n] \in R^{d\times n}$ denotes training images' features and $Y = [y_1, y_2, ..., y_n] \in R^{p\times n}$ indicates the corresponding transformed features. In the multi-label context, there are some samples with less similar label sets are even more similar in original feature space. The projection learning should decrease or eliminate this difference.

As existing samples with label sets similarity inconsistent with visual feature similarity, we adopt label sparse coding by l_1-$minimization$ to construct semantic graph as[3]. The weight matrix W^1, calculated from Algorithm 1, describes the semantic relations between each image and the rest ones. The semantic relations should be retained in low dimensional feature space, hence,

$$\min_P \frac{1}{2}\sum_i \|P^T x_i - \sum_j W^1_{ij} P^T x_j\|_2^2, s.t. P^T P = I \quad (1)$$

Using the label vectors of other images in the training set to sparsely reconstruct the label vector of each image instead of calculating the similarity between two label vectors isn't valid to deal with some polysemous words. For example, an image has labels "apple" and "table". Label "apple" might mean fruit, apple computer while they are visually different. So we combine visual features and labels together to obtain desired low-dimensional feature space.

As we have stated above, local neighbors information, which is help for annotation, should be kept in projection learning. In order to find the projection matrix with the minimal local estimation error, we can solve the following optimization problem[17]:

Algorithm 1: Semantic Graph W^1 Construction

Input: The label matrix of training images, represented by $C = [c_1, c_2, ..., c_n] \in R^{m \times n}$;

Output: The semantic graph W^1 with all diagonal elements being zero;

Set $C/c_i = [c_1, ..., c_{i-1}, c_{i+1}, ..., c_n]$;

for $i = 1; i \leq n; i = i + 1$ **do**

Set $D = C/c_i$;

Get sparse representation for label vector c_i by solving the optimization problem:

$$\min_{\alpha} \frac{1}{2}\|c_i - D\alpha\|_2^2 + \lambda\|\alpha\|_1 \qquad (2)$$

for $j = 1; j \leq i - 1; j = j + 1$ **do**

$W_{ij}^1 = \alpha_j$;

for $j = i + 1; j \leq n; j = j + 1$ **do**

$W_{ij}^1 = \alpha_{j-1}$;

$W_{ii}^1 = 0$;

$$\min_{P \in R^{d \times p}} trace(P^T XTX^T P), s.t. P^T P = I \qquad (3)$$

where $T = (I - W^2)^T (I - W^2)$. In order to construct the matrix W^2,we make use of label sets to seek images neighbor relations and generate the Gaussian kernel matrix over these neighbor images. The matrix is constructed as follows:

1. For $\forall x_i$ and $\forall x_j$, compute label sets similarity $S_{ij} = \frac{c_i^T c_j}{\min(|c_i|, |c_j|)}$. If $S_{ij} > \delta$ (δ is a constant), then we think x_i and x_j are neighbors (i.e. $x_j \in N_i$ and $x_i \in N_j$).
2. For $\forall x_i$, compute $\alpha_i^T = k_i^T (K_i + \lambda I)^{-1}$, $K_i \in R^{n_i \times n_i}$ is kernel matrix over $x_j \in N_i$, where $K(x, x_i) = \exp(-\frac{\|x - x_i\|^2}{\gamma})$, k_i denotes the vector $[K(x_i, x_j)]^T$ for $x_j \in N_i$, $\alpha_i \in R^{n_i}$.
3. $W^2 = [w_{ij}]^{n \times n}$, if $x_j \in N_i$, then w_{ij} equals the corresponding elements of α_i, otherwise w_{ij} equals 0.

By combining Eqs.(1) and (3),the projection matrix P can be formulated by solving the following optimization problem:

$$\min_{P \in R^{d \times p}} trace(P^T X \Delta X^T P), s.t. P^T P = I \qquad (4)$$

where $\Delta = \beta(I - W^1)^T (I - W^1) + (I - W^2)^T (I - W^2)$, β is a constant, which balances objective (1) and (3). The solution for formulation (4) can be obtained with the eigenvalue decomposition method,

$$X \Delta X^T p_k = \lambda_k p_k \qquad (5)$$

where p_k is the eigenvector corresponding to the k-th smallest eigenvalue λ_k of $X \Delta X^T$ and also the k-th column vector of the matrix P.

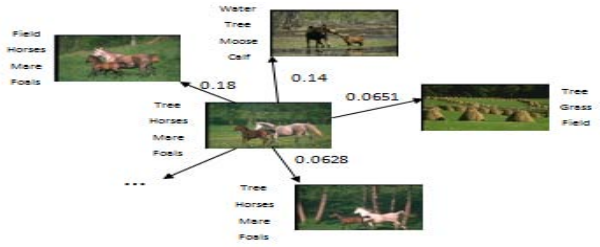

Fig. 1. An example of classic sparse reconstruction

3.2. Local Sparse Model

Based on the above feature selection approach, we can transform the original feature space into a low-dimensional feature space. The task of multi-label image annotation is to assign a set of labels to the query image. We assume that the more similar label sets among images are more similar in transformed lower-dimensional feature space, since feature selection here trys to solve or decrease the inconsistent between label and visual similarity. An image has many components, in this context, even though two images are very close, they have some different components. Since we don't segment these images, it's difficult to tag an image using traditional one-to-one mode. Hence, we apply sparse coding to solve this problem while [3, 4] also address the image tagging problem via sparse coding. However, they treat every element in the given dictionary equally for each unlabeled image. An example of the classic sparse coding is illustrated in Fig.1.Since treating all the training data as dictionary and all the dictionary elements are assigned with equal weight, it makes the reconstruction coefficient smaller even the image has more similar labels, as shown in Fig.1, and it loses the local information that is beneficial to annotate image. In order to drop off it and improve the performance, we propose the following local sparse coding procedure to reconstruct a query image.

1. For a query image q, find its K nearest neighbors in training data over transformed feature space, which is denoted by N_q
2. Treat all $y_i \in N_q$ as dictionary elements , denoted by D_q and optimize the formulation :

$$\min_{\alpha_q} \frac{1}{2} \parallel q - D_q \alpha_q \parallel_2^2 + \lambda \parallel diag(w_q)\alpha_q \|_1 \qquad (6)$$

where $w_q \in R^K$ and $w_{qi} = \exp(\frac{\|y_i - q\|_2}{\delta})$, $y_i \in N_q$. We adopt a weighted version of $LARS$[18, 19] to solve formulation (6).

3. Expand the dimensionality of α_q to n, represented by ${\alpha_q}'$. If $y_i \in N_q$, then ${\alpha_{qi}}'$ equals the corresponding elements of α_{qi}; otherwise, ${\alpha_{qi}}'$ equals 0.

method	MT	CMRM	CRM	CRM-Rect	MBRM	SML	LASSO	MSC*	Our method
P	0.04	0.10	0.16	0.22	0.24	0.23	0.24	0.23	0.24
R	0.06	0.09	0.19	0.23	0.25	0.29	0.29	0.24	0.25
N_+	49	66	107	119	122	137	127	159	180

Table 1. Performance comparison of different automatic image annotation methods on Corel5k dataset. MSC* refers to our implementation of [3] using our features.

3.3. Tag Propagate

Our goal is to assign a few of related labels to the unlabeled image. As above, we have reconstructed these query images, obtained the coefficient matrix $\alpha' = [\alpha_1', \alpha_2', ..., \alpha_t']$. In tag propagation process, the semantic relations are transformed from feature space to label space. We can get annotation matrix as follows:

$$C^q = C\alpha' \tag{7}$$

where $C = [c_1, c_2, ..., c_n]$ is the label matrix of the training images. The top labels with the largest values in every column c_i^q are considered as the final tagging results of the query images.

4. EXPERIMENTS

In this section, we evaluate the effectiveness of proposed approach for automatic image annotation task by comparing it with several existing state-of-the-art algorithms on Corel5k dataset.

Feature Extraction. Corel5k dataset has become an important benchmark for keyword based image retrieval and image annotation. It contains around 5000 images and manually annotated with 1 to 5 words. The set splits into 4500 training and 500 test examples. Here, we extract five different types of features, namely Grid Color Moment, Local Binary Pattern, Gabor Wavelets Texture, Edge[20] and Gist[21].

Evaluation Measures. We evaluate our approach with standard performance measures as previous work, that evaluate annotation performance per keyword, and then average over keywords. Precision and recall of each keyword are used as the performance measures. Precision of a word is defined as the number of images correctly annotated with this word (r) divided by the total number of images annotated (n), i.e.$P = \frac{r}{n}$; Recall of a word is defined as the number of images correctly annotated with it divided by the number of images that have this word in the ground-truth annotation (N), i.e.$R = \frac{r}{N}$. N_+ is used to denote the number of words with nonzero recall value, which indicates how many words the system has effectively learned. As previous work, each image is forced to be annotated with five words, even if the image has fewer or more words in the ground-truth. With more words annotated, recall will be increased while precision is decreased.

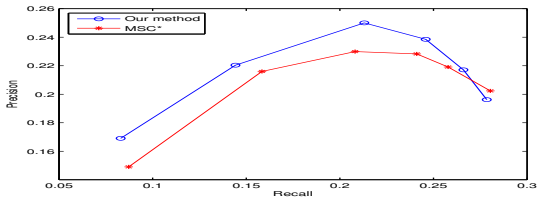

Fig. 2. Precision-Recall curves of our method and MSC* with the number of annotations from 2 to 7.

Results on Corel5k. Table 1 lists the comparison results of automatic image annotation on Corel5k dataset.Several state-of-the-art methods are compared.Results are reported for 260 words in the testing set.The parameters of Local Sparse Coding approach, proposed in this work, are selected corresponding to the best F1 value.In our experiment, we find 720 nearest neighbors to reconstruct unlabeled image.The results in Table 1 show that our approach is effective for image annotation. From the results in Table 1, we can make several observations. First, even though we achieve the same precision and recall as MBRM, which is one of the most popular and effective algorithms in image annotation field, we can obtain more words with positive recall.Second, although the recall of our method is lower than that of SML, we can get higher P and N_+.Third, compared to MSC method using the same features, we obtain improvements in precision and recall, and count 21 more words with positive recall. In addition, from Fig.2 we can see that our method consistently outperforms MSC*. In a word, our method is superior or highly competitive to several state-of-the-art approaches.

5. CONCLUSIONS

In this paper, we present a novel local sparse coding approach for automatic image annotation.During reconstruction process, we penalize the corresponding reconstruction coefficients to embed local information. We reported experimental results on Corel5k by using P, R and N+ performance measures. Experiments show the performance of our approach is comparable to several state-of-the-art algorithms. In the future, we would like to investigate other sparse methods such as structured sparse methods and use label semantic relations and local information to further guide the annotation task.

6. REFERENCES

[1] A. Makadia, V. Pavlovic, and S. Kumar, "A new baseline for image annotation," *Computer Vision–ECCV 2008*, pp. 316–329, 2008.

[2] M. Yang, L. Zhang, J. Yang, and D. Zhang, "Robust sparse coding for face recognition," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 625–632.

[3] C. Wang, S. Yan, L. Zhang, and H.J. Zhang, "Multi-label sparse coding for automatic image annotation," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 1643–1650.

[4] S. Zhang, J. Huang, H. Li, and D.N. Metaxas, "Automatic image annotation and retrieval using group sparsity," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 42, no. 3, pp. 838–849, 2012.

[5] Y. Gao, J. Fan, X. Xue, and R. Jain, "Automatic image annotation by incorporating feature hierarchy and boosting to scale up svm classifiers," in *International Multimedia Conference: Proceedings of the 14 th annual ACM international conference on Multimedia*, 2006, vol. 23, pp. 901–910.

[6] C. Yang, M. Dong, and J. Hua, "Region-based image annotation using asymmetrical support vector machine-based multiple-instance learning," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*. IEEE, 2006, vol. 2, pp. 2057–2063.

[7] J. Li and J.Z. Wang, "Automatic linguistic indexing of pictures by a statistical modeling approach," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 25, no. 9, pp. 1075–1088, 2003.

[8] K. Barnard and D. Forsyth, "Learning the semantics of words and pictures," in *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*. IEEE, 2001, vol. 2, pp. 408–415.

[9] G. Carneiro, A.B. Chan, P.J. Moreno, and N. Vasconcelos, "Supervised learning of semantic classes for image annotation and retrieval," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 29, no. 3, pp. 394–410, 2007.

[10] P. Duygulu, K. Barnard, J. De Freitas, and D. Forsyth, "Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary," *Computer VisionłECCV 2002*, pp. 349–354, 2006.

[11] J. Jeon, V. Lavrenko, and R. Manmatha, "Automatic image annotation and retrieval using cross-media relevance models," in *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*. ACM, 2003, pp. 119–126.

[12] V. Lavrenko, R. Manmatha, and J. Jeon, "A model for learning the semantics of pictures," NIPS, 2003.

[13] SL Feng, R. Manmatha, and V. Lavrenko, "Multiple bernoulli relevance models for image and video annotation," in *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*. IEEE, 2004, vol. 2, pp. II–1002.

[14] F. Monay and D. Gatica-Perez, "Plsa-based image auto-annotation: constraining the latent space," in *Proceedings of the 12th annual ACM international conference on Multimedia*. ACM, 2004, pp. 348–351.

[15] O. Yakhnenko and V. Honavar, "Annotating images and image objects using a hierarchical dirichlet process model," in *Proceedings of the 9th International Workshop on Multimedia Data Mining: held in conjunction with the ACM SIGKDD 2008*. ACM, 2008, pp. 1–7.

[16] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid, "Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation," in *Computer Vision, 2009 IEEE 12th International Conference on*. IEEE, 2009, pp. 309–316.

[17] M. Wu, K. Yu, S. Yu, and B. Schölkopf, "Local learning projections," in *Proceedings of the 24th international conference on Machine learning*. ACM, 2007, pp. 1039–1046.

[18] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *The Annals of statistics*, vol. 32, no. 2, pp. 407–499, 2004.

[19] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *The Journal of Machine Learning Research*, vol. 11, pp. 19–60, 2010.

[20] J. Zhu, S.C.H. Hoi, M.R. Lyu, and S. Yan, "Near-duplicate keyframe retrieval by nonrigid image matching," in *Proceedings of the 16th ACM international conference on Multimedia*. ACM, 2008, pp. 41–50.

[21] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 42, no. 3, pp. 145–175, 2001.