OBJECT RECOGNITION BASED ON DEPENDENT PACHINKO ALLOCATION MODEL

Yuanning Li[1], Weiqiang Wang[2], Wen Gao[1]

[1]Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, 100080
[2]Graduate School, Chinese Academy of Sciences, Beijing, China, 100039
{ynli, wqwang, wgao}@jdl.ac.cn

ABSTRACT

Recently the "bag of words" model becomes popular in the approaches to object recognition. These approaches model an image as a collection of local patches called "visual words", and recognize objects in the image through inferring latent topics associated with the set of visual words. In this paper, we apply an extension version of Pachinko Allocation Model (PAM) [5] to object recognition. Our PAM based approach models the correlation-ship of latent topics explicitly in a hierarchical structure. To relax the independent assumption for visual words and refine the topic inferring, we incorporate the prior knowledge of co-occurrence dependence among visual words into PAM. Highly competitive recognition results on both Caltech4 and Caltech101 datasets show the proposed approach is more expressive and discriminative than most existing methods of object recognition.

Index Terms— *object recognition, "bag of words"*

1. INTRODUCTION

Object recognition is a challenging issue in computer vision. Even rigid objects in the same category may take on very different appearances, due to variable lighting, affine or projective transformation, occlusion, and clutter, etc. In recent years, approaches to object recognition based on the "bag of words" model become very popular. The related approaches follow a same general framework: firstly, a stable keypoint detector is used to identify informative local patches; secondly, discriminative descriptors are calculated for these local patches; thirdly, each descriptor is quantized into a discrete visual word; finally, an effective learning technique is exploited to model the mapping between object categories and the set of visual words.

The "bag of words" representation of an image makes the researchers of object recognition easily benefit from successful paradigms in natural language processing (NLP). Recently some important works [1,2] in object recognition motivate its popularity. Sivic et.al. [1] apply the model Probabilistic Latent Semantic Analysis (pLSA) used in the statistical NLP to discover object categories. Li Feifei et.al. [2] present a Bayesian hierarchical model based on Latent Dirichlet Allocation (LDA) to classify natural scenes. Both methods assume: each visual word in an image arises from a mixture of topics; topics are shared by all images in a collection; topic proportions are image-specific and randomly drawn from a certain distribution. Probabilities of visual words as well as the latent topics are learned in a statistical manner. Inter-relationships among the visual words are ignored and topics are assumed to be independent with each other.

Some researchers model object categories by adding correlation information among visual words. Fergus et.al.[3] present a new model TSI-pLSA that integrates relative location information into the pLSA model to learn object categories. Gang Wang et.al. [4] extend Hierarchical Dirichlet Process (HDP) by taking account of co-occurrence of visual words. The HDP captures topic correlation defined by nested data structure, but it can not automatically discover correlations from un-structured data. These methods above discover the topic distribution by capturing correlations among visual words, but they fail to directly model correlation of topics.

Correlated topics are common for visual data in real world (e.g. cars and streets). Ignoring topic correlations may hamper coherent topics discovery. In this paper, we propose a new approach to object recognition based on dependent PAM (DPAM), which captures topics correlation explicitly from dependent visual words.

2. OBJECT RECOGNITION BASED ON DPAM

Our approach also uses the "bag of words" representation of images and follows the general framework for object recognition. In the four-level PAM as depicted in Fig.1, a sub-topic is assigned to each visual word of an image, and a super-topic which explicitly models the correlation among sub-topics is assigned to each sub-topic. All the super-topics share a same root. This multi-level directed acyclic graph (DAG) structure is called Pachinko Allocation Model(PAM) [5] in text processing. Inspired by [4], we incorporate some prior knowledge of co-occurrence dependence among visual words into the PAM to better model a generating process of an image. In the following subsection, we will introduce the

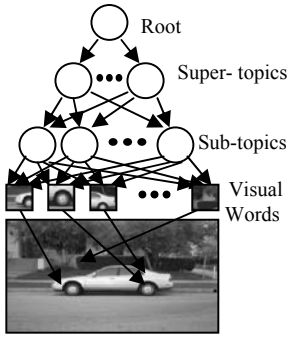

Fig1. Hierarchical topic structure.

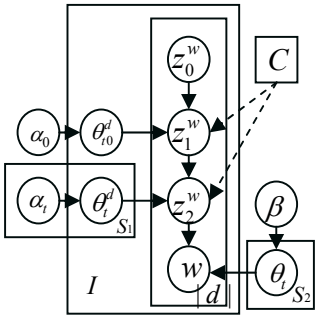

Fig.2. Graphical model for DPAM

details of our model DPAM, the parameter estimation algorithm and the object recognition algorithm.

2.1. Dependent Pachinko Allocation Model

In our model, each visual word is denoted by an integer $w \in \{1, 2, ..., W\}$, which corresponds to the index of the visual word in a visual vocabulary *V*. An image *d* is a collection of N_d visual words, i.e. $d = (w_1, w_2, ...w_{Nd})$, where w_i is the *i*th visual word of the image *d*. A corpus *D* is a collection of *I* images of the same object class, denoted as $D = \{d_1, d_2, ...d_I\}$.

Dependent Pachinko Allocation Model (DPAM) is a four level hierarchical structure composed of one root topic, s_1 super-topics at the second level, s_2 sub-topics at the third level and visual words as leaves. Parent nodes in the upper level are fully connected with child nodes in the lower level. It means that sub-topics are shared among different super-topics, and visual words are shared among different sub-topics. Fig. 2 shows the graphical model of DPAM. The multinomial distributions of the root topic and super-topics are sampled individually for each image from a single Dirichlet distribution $g_i(\alpha_i)$, where α_i is a vector with the same dimension as the number of children in root topic or super-topic. For simplification, the multinomial distributions for sub-topics are sampled once for the whole corpus from a single Dirichlet distribution $g(\beta)$, where β is a vector of *W* dimensions. $g(\beta)$ can be thought as a mixture model for all visual words, while $g_i(\alpha_i)$ acts as a mixture model for all sub-topics in different images. According to the standard PAM [5], an image *d* is generated by the following process:

1. Sample $\theta^d_{t0}, \theta^d_{t1}, ..., \theta^d_{ts} (s = 1 + s_1 + s_2)$ from $g_0(\alpha_0), g_1(\alpha_1), ..., g_{s1}(\alpha_{s1}), g(\beta)$, where θ^d_{ti} is a multinomial distribution of topic t_i over its children.
2. For each visual word *w* in an image *d,*
   a) Sample a topic path $z^w = < z^w_0, z^w_1, z^w_2 >$. z^w_0 is always the root topic t_0. z^w_1 and z^w_2 correspond to a super-topic and a sub-topic respectively. z^w_i, as a child of z^w_{i-1}, is sampled according to the multinomial distribution $\theta^d_{z^w_{i-1}}$
   b) Sample the word *w* from $\theta^d_{z^w_2}$.

Following this process, the joint probability of generating an image *d*, the topic assignments z^d, and the multinomial distribution θ^d can be calculated by:

$$P(d, z^d, \theta^d \mid \alpha, \beta) = \prod_{j=0}^{s1} P(\theta^d_{tj} \mid \alpha_j) \prod_{j=s1+1}^{s} P(\theta^d_{tj} \mid \beta) \times \prod_{i=1}^{Nd} \left(\prod_{k=0}^{1} P(z^{wi}_{k+1} \mid \theta^d_{z^{wi}_k}) P(w_i \mid \theta^d_{z^{wi}_2}) \right) \quad (1)$$

Integrating out θ^d and summing over z^d, we can get the marginal probability of a document $P(d \mid \alpha, \beta)$. The probability of generating the corpus *D* is calculated by:

$$P(D \mid \alpha, \beta) = \prod_{i=1}^{I} P(d_i \mid \alpha, \beta) \quad (2)$$

Since the visual words with high dependence may share the same semantic meaning in the real world. To relax the independent assumption for visual words in standard PAM and refine the topic inferring, we extend PAM through introducing a prior knowledge of the co-occurrence dependence among visual words. A dependence structure *C* is introduced at the stage of topic sampling, so that the co-occurrence visual words in an image tend to share the same topic. $P(z^w_k \mid \theta^d_{z^w_{k-1}})$ (k = 1, 2) in (1) is replaced by $P(z^w_k \mid \theta^d_{z^w_{k-1}}, C_w)$, and we have

$$P(z^w_k \mid \theta^d_{z^w_{k-1}}, C_w) = P(z^w_k \mid \theta^d_{z^w_{k-1}}) \prod_{w' \in A(z^w_k), w' \neq w} (1 + C(w, w')) \quad (3)$$

where C_w represents the dependence between the visual word *w* and the others in the vocabulary *V*. $A(z^w_k)$ is the set of visual words that have been assigned with the topic z^w_k. The dependence coefficient $C(w, w')$ which captures the co-occurrence of two visual words *w* and w' is defined as

$$C(w, w') = \frac{2Frq(w, w')}{Frq(w) + Frq(w')} - \frac{Frq(w, w')}{Frq(w)^{\lambda} + Frq(w')} - \frac{Frq(w, w')}{Frq(w) + Frq(w')^{\lambda}} \quad (4)$$

$Frq(w, w')$ is the number of times that *w* and w' appear in a same image. $Frq(w)$ denotes the number of times that *w* appears in the corpus. λ is a constant slightly bigger than 1.0. The last two terms in (4) are penalty factors.

2.2. Parameter estimation with Gibbs sampling

Now we show how to train a DPAM for an object category. Our goal is to obtain the probabilities $P(w \mid z)$ of visual

words conditioned on different sub-topics and the sub-topics distribution $P(z)$ for each object class. Given a corpus of images from the same class, for each visual word w in an image d, the joint probability of a super-topic and sub-topic is estimated by

$$P(z_1^w = t_k, z_2^w = t_p \mid d, z_{-w}, \alpha, \beta, C_w) \propto$$
$$\Big(\prod_{w' \neq w, w' \in A_{t_k}} (1 + C(w, w'))\Big) \times \frac{n_{0k}^d + \alpha_{0k}}{n_0^d + \sum_{j=1}^{s1} \alpha_{0j}} \times \tag{5}$$
$$\Big(\prod_{w' \neq w, w' \in A_{t_p}} (1 + C(w, w'))\Big) \times \frac{n_{kp}^d + \alpha_{kp}}{n_k^d + \sum_{j=1}^{s2} \alpha_{kj}} \times \frac{n_{pw} + \beta_w}{n_p + \sum_{j=1}^{W} \beta_j}$$

where t_k and t_p correspond to the super-topic and sub-topic assignments respectively for w, and z_{-w} is the topic assignments for all the visual words except w. Excluding the current visual word w, n_x^d is the number of occurrences of a topic t_x in the image d; n_{xy}^d is the number of occurrences that a topic t_y is sampled from its parent t_x in d. n_x is the number of occurrences of sub-topic t_x for the whole corpus; n_{xw} is the number of occurrences of the word w in the topic t_x for the whole corpus. α_{xy} is the yth component of α_x, and β_w is the wth component of β. The first and the second ratio in (5) express the probabilities of super-topic t_k and sub-topic t_p in d respectively. The last ratio expresses the probability of w under sub-topic t_p. The first and the third term are dependent correction factors. For simplicity, we update α_x and α_{xy} after each iteration by moment matching, and the related details refer to [5].

2.3. Object recognition Based on DPAM

Given an image $d' = (w_1', w_2', ..., w_{Nd'}')$, the probability of an object class c is

$$P(c \mid d') \propto P(d' \mid c) P(c) \propto P(d' \mid c) \tag{6}$$

We use equal probability for $P(c)$ here, since we have no knowledge about the distribution of object classes. Through learning, we obtain a DPAM for each object class c. Based on the DPAM for c, $P(w \mid z, c)$ and $P(z \mid c)$ can be evaluated, where z denotes sub-topic. Then we have:

$$P(d' \mid c) = \prod_{i=1}^{Nd} P(w_i' \mid c) = \prod_{i=1}^{Nd} \Big(\sum_{l=1}^{S2} P(w_i' \mid z_l, c) P(z_l \mid c)\Big) \tag{7}$$

The object class with the highest likelihood $P(d' \mid c)$ is regarded as the final recognition result.

3. EXPERIMENTS

We evaluate our method on two popular datasets Caltech 4 and Caltech 101. Two keypoint detectors, saliency [6] and DoG[7] are used to locate interesting regions. Each region is characterized by a 72 dimensional SIFT descriptor. Through the K-means algorithm, we obtain a codebook of visual words with 800 entries. In sampling procedure, λ is set to 1.05. Each entry of α_0 is set to 0.1, and each entry of β is set to 0.01.

3.1. Exp.1: Caltech 4

For the Caltech 4 dataset, we randomly select 100 images from each object class for training and test. Four super-topics and ten sub-topics are chosen in both PAM and DPAM. The related experimental results are summarized in Table 1. The experiment shows DPAM has a better overall performance 97.75% than PAM (96.5%). Introducing the dependence of visual words is effective. The competitive performance obtained in our comparison experiments with [2, 4, 8] (see in Fig.3) also demonstrates DPAM is an effective modeling approach for object recognition.

Table1. The confusion matrix of PAM and DPAM for Caltech 4 dataset.

	Airplane	Face	Leopard	Motorbike
a.	95.0/97.0	4.0/3.0	0/0	1.0/0
f.	0/0	99.0/99.0	0/0	1.0/1.0
l.	1.0/0	3.0/1.0	93.0/96.0	3.0/3.0
m.	1.0/1.0	0/0	0/0	99.0/99.0

The rows denote the ground-truth category labels. The columns denote classification results. For a/b in each grid, a: the performance of PAM, b: the performance of DPAM.

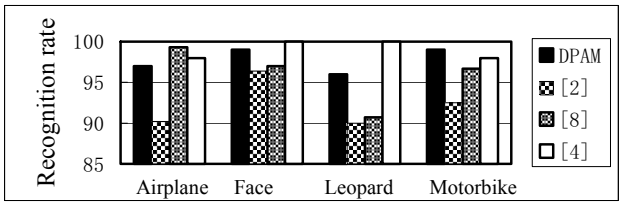

Fig.3. Performance comparison between DPAM and [2, 4, 8].

3.2. Exp.2: Caltech 101

We further evaluate our model on more object classes based on the Caltech 101 dataset. 30 images are randomly selected from each class for training. Other settings are the same as Exp.1. An illustration is given in Fig.4 to show the power of DPAM in capturing the correlation-ship among multiple sub-topics, where each circle corresponds to a super-topic learned in car side category, and each box corresponds to a sub-topic. The number on the edge corresponds to α_{xy} for super-topic x and sub-topic y.

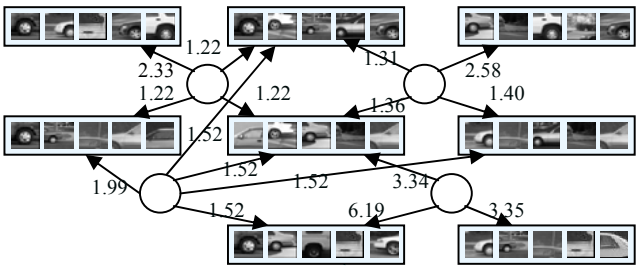

Fig.4. An illustration of correlated sub-topics captured by DPAM.

For each super-topic x, we only show the sub-topics whose Dirichlet parameter α_{xy} is bigger than 1.0. For each sub-topic, top five visual words with biggest $P(w|z)$ weight are shown in the box. We can see that all super-topics share the same sub-topic in the middle, which is prominent in car side category.

We also compare the performances of our approach with other methods recently published on Caltech 101 dataset in Fig.5. It shows the result of our approach is one of the best reported results on the Caltech 101 dataset. Compared with the methods [2, 4] based on latent topics and "bag of words" model, our method achieves significant improvement. It shows that modeling correlation of topics and dependence of visual words jointly is helpful for latent topic analysis in object recognition. Compared with other methods, the performance of DPAM is improved sharply as the number of training images increases. One possible explanation is that topics correlation and dependence of visual words can be modeled more accurately if more training images are used, and it helps to better model object categories. This phenomena has also been noticed in [4]. When the number of training images for each object class reaches 30, DPAM obtains a performance of 64.8%.

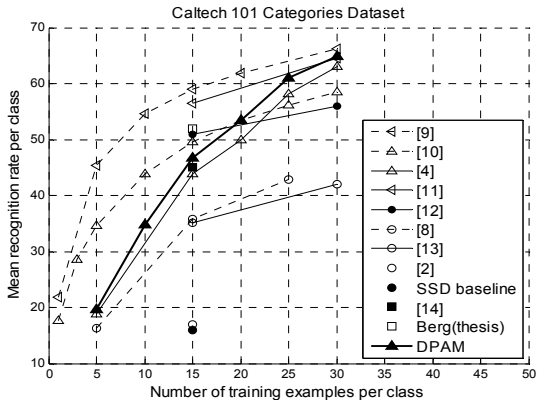

Fig.5. Performances of DPAM and other recent methods

4. CONCLUSION

In this paper, we present a new approach to modeling object categories based on an extension of PAM model, i.e., DPAM. Our approach adopts more realistic assumptions for generating images and models the correlation of latent topics in images explicitly. Our experiments demonstrate the DPAM has a better modeling power to the issue of object recognition through considering the dependence among visual words. The competitive experimental results obtained on the Caltech 4 and Caltech 101 datasets show that our object recognition approach based on DPAM is very effective, and it is expected that DPAM can model object categories more accurately when more training images are available for each object category.

5. REFERENCES

[1] J. Sivic, B. Russell, A.A.Efros, and A. Zisserman, "Discovering Objects and Their Location in Images," *IEEE ICCV*, pp. 370-377, vol.2, 2005.

[2] Li Fei-Fei, Pietro Perona, "A Bayesian Hierarchical Model for Learning Natural Scene Categories," *IEEE CVPR* pp. 524-531, vol3, 2005.

[3] R.Fergus, L.FeiFei, P. Perona, A. Zisserman, "Learning Object Categories from Google's Image Search," *IEEE ICCV*, pp. 1816-1823, vol.2, 2005.

[4] G. Wang, Y. Zhang, and L. Fei-Fei, "Using Dependent Regions for Object Categorization in a Generative Framework," *IEEE CVPR,* pp. 1597-1604, 2006.

[5] Wei Li and Andrew McCallum, "Pachinko Allocation: DAG-Structured Mixture Models of Topic Correlations," *ICML*, 2006.

[6] T. Kadir and M. Brady, "Scale Saliency and Image Description," *IEEE IJCV*, pp. 83-105, 2001.

[7] D. Lowe, "Object Recognition from Local Scale Invariant Feature," *IEEE ICCV,* pp. 1150-1157, vol.2, 1999.

[8] A. Holub and P.Perona, "A Discriminative Framework for Modeling Object Class," *IEEE CVPR,* pp. 664 - 671 vol. 1,2005.

[9] Hao Zhang, Alexander C. Berg. Michal Maire, and Jitendra Malik, "Svm-knn: Discriminative Nearest Neighbor Classification for Visual Category Recognition," *IEEE CVPR*, pp. 2126-2136, 2006.

[10] Kristen Grauman and Trevor Darrell, "Pyramid Match Kernels: Discriminative Classification with Sets of Images Features," Technical Report CSAIL-TR-2006-020, MIT, 2006.

[11] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories," *IEEE CVPR*, pp. 2169-2178, 2006.

[12] Jim Mutch and Davie Lowe, "Multiclass Object Recognition with Sparse, Localized Features," *IEEE CVPR*, pp. 11-18, 2006.

[13] T.Serre, L. wolf , and T.Poggio, "Object Recognition with Features Inspired by Visual Cortex," *IEEE CVPR,* pp. 994-1000, vol.1, 2005.

[14] A.C. Berg, T.L. Berg, and J. Malik, "Shape matching and Object Recognition Using Low Distortion Correspondence," *IEEE CVPR*, pp. 26-33, vol.1, 2005.