An Object Co-occurrence Assisted Hierarchical Model for Scene Understanding

Xin Li
xinli@temple.edu
Yuhong Guo
yuhong@temple.edu

Computer and Information Sciences
Temple University
Philadelphia
PA 19122, USA

Abstract

Hierarchical methods have been widely explored for object recognition, which is a critical component of scene understanding. However, few existing works are able to model the contextual information (e.g., objects co-occurrence) explicitly within a single coherent framework for scene understanding. Towards this goal, in this paper we propose a novel three-level (superpixel level, object level and scene level) hierarchical model to address the scene categorization problem. Our proposed model is a coherent probabilistic graphical model that captures the object co-occurrence information for scene understanding with a probabilistic chain structure. The efficacy of the proposed model is demonstrated by conducting experiments on the LabelMe dataset.

1 Introduction

The task of scene recognition or scene understanding usually automatically labels an image with a set of semantic categories such as *office, street, coast and etc*. It is different from the object categorization problem, since the latter focuses on local information that reflects the presence and absence of objects, while the former requires global information that describes the whole image. Scene understanding is a fundamental computer vision task as it can provide contextual information to guide other processes such as object recognition [2], and has high potentials to improve the performance of computer vision application systems such as browsing (natural grouping of images based only on low-level features) and retrieval (filtering images in archives based on contents).

Historically, there is a controversy between cognitive psychology and computer vision on the task of scene recognition, the main source of which is about achieving scene recognition using low-level features to directly capture the gist of a scene versus using intermediate semantic representations [4]. Following this controversy, two main directions have been explored on this task. One attempts to use supervised classifiers that directly operate on low-level image features such as color, texture, and shape [3, 20]. The weakness is that a low-level representation is difficult to be generalized to some scene categories such as indoor scenes, as discussed in [14]. The other direction ventures to bridge the gap between low-level image properties and the semantic content of a scene using intermediate semantic representations that can be obtained by processes such as segmentation and object recognition. Most recent works have put effort on semantic models to tackle the scene recognition problem [1, 7, 9,

http://dx.doi.org/10.5244/C.26.81

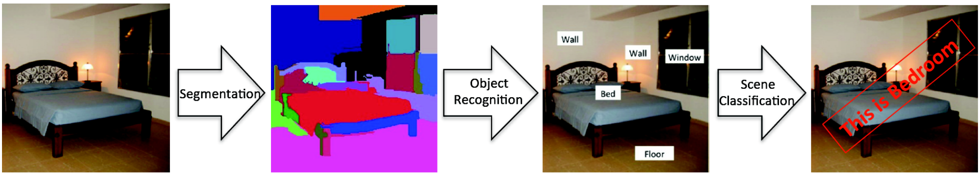


Figure 1: The work flow of our proposed system. First, one performs segmentation on the input image. Second, features are extracted from each superpixel. Third, one assigns semantic labels to each superpixel based on the feature extracted in the 2nd step. (Note that in this step, what our unsupervised model assigns to a region is an object index such as 'A' or 'B' and it is not aware what A refers to. For the sake of understanding, we replace the index to its corresponding label manually in the figure.) Finally, a scene-level label is associated to this image. In the example, *bedroom* is most likely to contain all present objects.

10]. It is a currently well accepted view that in order to understand the context of a complex scene, one needs first to recognize the objects and then in turn recognize the category of the scene [7, 10]. Moreover, L. Li *et al.* pointed out in [8] that there is a discrepancy between the image representations and the image recognition goal; with lower-level features, more work needs to be done to achieve higher-level recognition goals. Thus capturing high-level representations is important for the target scene recognition task. Although the high-level contextual information has been proved useful on numerous higher-level classification tasks in recent years, we notice that few existing scene recognition models in the literature are able to encode contextual information such as *scene layout, background class, and object co-occurrences*. We thus aim to employ the object-level contexts in a generative probabilistic model which does not require tedious object annotations over the training data.

In this work, we propose a novel three-level (superpixel level, object level and scene level) generative hierarchical model for scene understanding, which captures the high-level contextual information expressed in form of object co-occurrences. Specifically, it encodes the correlation of object classes using a probabilistic chain structure over the object class assignment variables in each image. Different from previous works such as [10], the proposed model only requires the images as inputs and it can automatically extract information at different levels within a generative probabilistic graphical model framework. Figure 1 illustrates the work flow of our proposed system which involves three stages: (1) segmentation, (2) object recognition, and (3) scene classification. We integrate the last two stages into a coherent probabilistic graphical model. The object-level annotation under our model is accomplished automatically during training, and thus the proposed model is *unsupervised* at this level. This is based on an assumption that recognizing explicit object categories is unimportant as long as one can build the association between the objects from the same category [13]. For example, after seeing a *car* passing by for the first time, the information from this encounter is critical to determine whether a newly observed object belongs to the same *car* category even if we are not aware these objects are named as *car*. The proposed model indexes each object class with a unique integer and ensures that the appearance-wise similar objects share the same class, instead of associating each object class with a fixed human defined concept. Unsupervised learning at the object-level then is expected to automatically capture useful concepts for each object class. Moreover, an ensemble prediction strategy based on training multiple models with random restarts is employed to further improve the model performance. The efficacy of the proposed model is demonstrated by experiments conducted over the LabelMe dataset.

The remainder of the paper is organized as follows. In Section 2, we introduce the related work. In Section 3, we present the proposed hierarchical model. We report the experiments and results in Section 4 and then conclude this paper in Section 5.

2 Related Work

As we mentioned in the previous section, the literature of scene recognition can be divided into two groups, following two directions, low-level modeling and semantic modeling, respectively. A comprehensive review on this topic can be found in [2]. The low-level modeling methods assume the categories of a scene can be directly determined by the low-level features such as color and texture properties of the image. For example, horizontal edges have been frequently observed in natural scenes (mountain, coast, and forest), and vertical edges appear often in urban scenes (street and building). Some low-level modeling methods work with the features extracted from the whole image. For example, [21] proposes a hierarchical structure that discriminates many scene classes effectively merely using low-level image features. [17] presents a new type of features, Combined Multi-Visual Features, which integrate color, texture, and shape features altogether. A few other methods split the image into a set of subregions, where features are extracted independently. For example, [16] develops a framework to combine multiple SVM classifiers in a belief network, where the color and texture features are extracted from image sub-regions and classified separately.

The semantic modeling methods take scene contents such as presence or absence of objects as cues for improving the classification performance obtained using low-level features alone. Most of recent work in this direction focuses on dealing with the gap between the image representations and the image recognition goals. A few works, [6, 18, 19, 22], utilize the correlations between the statistics of low-level features across images that contain one object, or the whole object category. A high-level representation named Object Bank is presented in [7] for scene classification, which encodes the semantic and spatial information of the objects within an image. Specifically, in an object bank representation, an image is represented as a collection of scale-invariant response-maps of a large number of pre-trained generic object detectors. Although representing the state-of-the-art for the scene recognition task, this approach is a supervised method and requires object annotations to be provided to train the object detectors. [1, 9] propose some hierarchical probabilistic methods that use unsupervised techniques with bag-of-words schemes to obtain relevant intermediate representations. [10] develops a graphical model framework to perform three visual recognition tasks: annotation, segmentation, and classification. This graphical model however requires additional textual information as inputs.

Following the path of seeking a proper intermediate representation, we develop a hierarchical probabilistic graphical model in this work to model the high-level image representations for scene recognition and understanding. Different from [7], our approach is unsupervised and can avoid the expensive and time-consuming object annotation process required for training the object detectors. Unlike [10], our model does not require pre-collected relevant textual information. We assume that only the number of object categories is known pre-hand without involving human defined object concepts. In particular, our model takes the high-level contextual information in form of object co-occurrence into account, which hasn't been captured by previous probabilistic hierarchical methods developed in the literature.

3 Proposed Method

In this section, we present a hierarchical probabilistic graphical model that employs the contextual information to assist scene classification. This model, shown in Figure 2, integrates both low-level features and high-level representations for scene understanding. In our setting, the total number of object classes for the whole image set is assumed to be known. But as an unsupervised model, it does not require the object annotations to be provided. Instead, object annotation will be accomplished implicitly as an intermediate result in our approach. In this model, object classes are not pre-associated with fixed human defined concepts (*e.g.* desk, computer, and sky), but are simply represented using consecutive index integers from 1 to the number of object classes, which is 30 in our experiments.

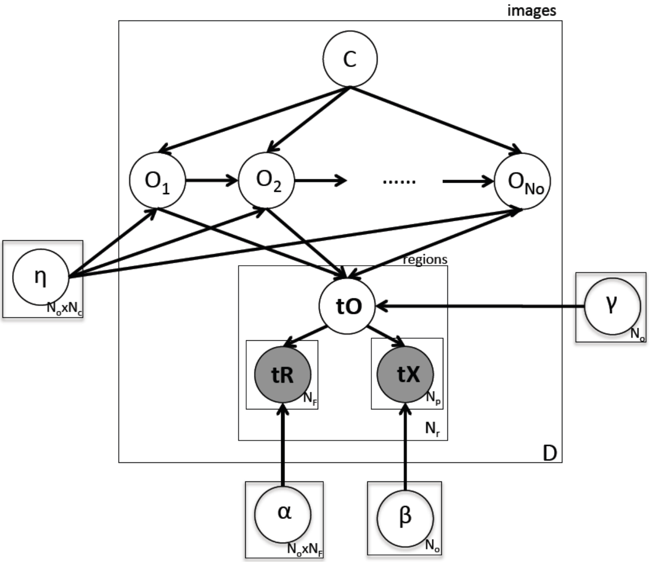


Figure 2: The proposed model. Nodes denote random variables and edges indicate dependencies. The variables at the bottom-right corner of each box denote the numbers of replications. The box indexed by D represents a single image in the image set of size D. The box indexed by N_r denotes the visual information of the image. N_c, N_o, N_F, and N_p denote the number of different scenes, objects, region features, and patches respectively. $\alpha, \beta, \gamma, \eta$ are the parameters of the distributions associated with the variables. We omitted the distribution hyperparameters for clarity's sake.

3.1 Model Components

The proposed model integrates both low-level representations and intermediate semantic modeling to explain an image from three different levels: the superpixel level, the object level and the scene level.

Low-Level Representation The bag of words methodology is adopted to represent an image at a low level. First, we segment the input image into multiple superpixels. The segmentation method used in this study is from [5]. Second, local descriptors over these superpixels are computed. Similar to [10], for each region we extract $N_F = 4$ types of features, where F={shape, color, location, texture}, which are represented as the **tR** nodes in our model. We use the shape and location features described in [12]. The color features are simply computed from histograms and the texture features are the average responses of filter-banks in each superpixel/region. Moreover, a set of patches can be obtained by dividing the image into blocks. SIFT [11] features are extracted from these patches and the **tX** nodes in the model denote the vector quantized SIFT features of the patch locating in a particular region. Third, the produced descriptors are then quantized to form the visual vocabulary. Codebooks of the *shape, color, location* and *texture* features have the sizes of 100, 30, 50, and 150, respectively. The codebook of the SIFT feature has 500 code words. Finally, we count the occurrences of each specific code word in the vocabulary in order to build the histograms of the code words.

Intermediate Semantic Modeling In order to bridge the gap between the low-level representation and the high-level classification goal, we design an object-level structure over the superpixel-layer features. As Figure 2 shows, each region is assigned an object label tO that comes from one of the object classes indexed from O_1 to O_{N_o}. Moreover, objects appearing in a scene are not independent to each other, but correlated. We encode the object correlation information using a chain structure over the O_i nodes, $i = 1, 2, \dots, O_{N_o}$, in Figure 2. The rationale behind this design is to capture the correlation information between object categories without inducing more complicated inference problems. As demonstrated in the figure, the resulting joint distribution of a given scene with class C, the appearances of object classes, O_1, O_2, ..., O_n, the objects **tO**, the region features **tR**, and the image patch features **tX** can be expressed as:

$$
\begin{aligned}
P(C, O_1, O_2, \dots, O_n, \mathbf{tO}, \mathbf{tR}, \mathbf{tX} | \alpha, \beta, \gamma, \eta) = P(C) \cdot P(O_1 | C, \eta_1) \cdot \prod_{i=2}^{N_o} P(O_i | O_{i-1}, C, \eta_i) \\
\times \prod_{l=1}^{Nr} (P(tO_l | O_1, O_2, \dots, O_n, \gamma) \cdot \prod_{k=1}^{N_F} p(tR_{lk} | tO_l, \alpha_k) \cdot \prod_{m=1}^{N_p} P(tX_{lm} | tO_l, \beta))
\end{aligned}
\tag{1}
$$

The top-down generative process of this model can be outlined in the following way. Given the scene class C, the probability of an object class indicator variable is governed by a binomial distribution. Specifically, for each image, we sample the object class indicator variable as $O_i \sim Bino(\eta_i | C, O_{i-1})$. Then, given the object configuration of the current image, $O_1, O_2, \dots, O_{N_o}$, the probability of the object variable tO of each image region has a multinomial distribution, $tO \sim Mult(\gamma | O_1, O_2, \dots, O_n)$. Next for each image region, we sample its image appearance features from a multinomial distribution $tR_i \sim Mult(\alpha_i | tO)$ for $i \in F$, and sample its patches in a similar way, $tX \sim Mult(\beta | tO)$. We use hyperparameters $\{\pi_{f_i}\}, \pi_x, \pi_o$ to define the Dirichlet distributions of the model parameters $\{\alpha_i\}, \beta, \gamma$ respectively and use $\{\theta_i\}$ to define the Beta distributions of $\{\eta_i\}$.

3.2 Automatic Model Learning

To learn the model parameters automatically, we derive a collapsed Gibbs sampling algorithm. For each image we sample the latent variables $O_j \in \{0, 1\}, \forall j$ and tOs.

For the dth image, given all other variables, the conditional distribution of each latent variable O_{dj} can be computed as

$$P(O_{dj}|C_d,O_{d1},O_{d2},\ldots,O_{dN_o},\mathbf{tO}_d) \propto P(\mathbf{tO}_d|O_{d1},\ldots,O_{dN_o})\cdot P(O_{d1}|C_d)\cdot\prod_{i=2}^{N_o}P(O_{di}|C_d,O_{di-1}). \tag{2}$$

The first term of this equation can be calculated by:

$$P(\mathbf{tO}_d|O_{d1},\ldots,O_{dN_o}) = \prod_{n=1}^{N_r} P(tO_{dn}|O_{d1},\ldots,O_{dN_o}) \tag{3}$$

where tO_{dn} denotes the object class assignment for the nth region in the dth image. The consecutive terms can be obtained by

$$P(O_{dj}=t_j|O_{dj-1}=t_{j-1},C_d=c) = \frac{n_{ct_jt_{j-1},-d}+\theta_j^{t_j}}{\sum_{t_j} n_{ct_jt_{j-1},-d}+\theta_j} \tag{4}$$

where t_j (or t_{j-1}) takes an indicator value of 0 or 1. $t_j = 1$ indicates the jth object class presents in the dth image and $t_j = 0$ indicates its absence. The value $n_{ct_jt_{j-1},-d}$ denotes the number of times that the setting $\{C=c, O_j=t_j, O_{j-1}=t_{j-1}\}$ appears in the whole image set excluding the dth image. θ_j are the hyperparameters for $\eta_j \sim Beta(\theta_j^1,\theta_j^0)$ where $\theta_j = \theta_j^0+\theta_j^1$. We integrate the η_j out according to the conjugacy of the beta distribution and the binomial distribution.

Let $\mathbf{tR}_{dn}$ and $\mathbf{tX}_{dn}$ represent the sets of region features and patches of the nth region in the dth image. Following the Markov property of variables $\mathbf{tO}$, we analytically integrate out parameters α, β, and γ. Then the posterior over the object variable tO_{dn} can be described as:

$$\begin{aligned} P(tO_{dn}=o|\overline{tO}_{dn},O_{d1},\ldots,O_{dN_o},\mathbf{tR}_{dn},\mathbf{tX}_{dn}) &\propto P(tO_{dn}=o|\overline{tO}_{dn},O_{d1},\ldots,O_{dN_o}) \\ &\times P(\mathbf{tR}_{dn}|\overline{\mathbf{tR}}_{dn},tO_{dn})\cdot P(\mathbf{tX}_{dn}|\overline{\mathbf{tX}}_{dn},tO_{dn}) \end{aligned} \tag{5}$$

where $\overline{tO}_{dn}$ denotes all other $\mathbf{tO}$ variables except the tO_{dn}. Similar explanations can be applied on the notations $\overline{\mathbf{tR}}_{dn}$ and $\overline{\mathbf{tX}}_{dn}$. The first term of this product can be easily calculated in the following way:

$$\begin{aligned} P(tO_{dn}=o|\overline{tO}_{dn},O_{d1},\ldots,O_{dN_o}) &= \frac{P(tO_{dn}=o,\overline{tO}_{dn}|O_{d1},\ldots,O_{dN_o})}{P(\overline{tO}_{dn}|O_{d1},\ldots,O_{dN_o})} \\ &= \frac{n_{oO_{d1}\ldots O_{dN_o},-dn}+\pi_o}{\sum_{o'} n_{o'O_{d1}\ldots O_{dN_o},-dn}+N_o\pi_o} \end{aligned} \tag{6}$$

where the value $n_{oO_{d1}\ldots O_{dN_o},-dn}$ denotes the number of appearances of the setting $\{tO = o, O_1 = O_{d1},\ldots,O_{No} = O_{dNo}\}$ excluding the nth region of the dth image. π_o is the hyperparameter for $\gamma \sim Dir(\pi_o)$. Using standard Dirichlet integral formulation, we can obtain the last two terms:

$$\begin{aligned} P(\mathbf{tR}_{dn}|\overline{\mathbf{tR}}_{dn},tO_{dn}=o) &= \prod_{i=1}^{N_F} P(tR_{dni}=f_i|\overline{\mathbf{tR}}_{dni},\mathbf{tO}_{dn}=o) \\ &= \prod_{i=1}^{N_F} \frac{n_{of_i,-dn}+\pi_{f_i}}{\sum_{f_i'} n_{of_i',-dn}+N_{f_i}\pi_{f_i}} \end{aligned} \tag{7}$$

where π_{f_i} is the hyperparameter for the symmetric Dirichlet distribution of α_i; the value $n_{of_i,-dn}$ defines the number of occurrences for f_i with o excluding the instances related to dn.

$$P(\mathbf{tX}_{dn}|\overline{\mathbf{tX}}_{dn}, tO_{dn} = o) = \frac{\Gamma(\sum_{x'} n_{ox',-dn} + N_x\pi_x)}{\Pi_{x'}\Gamma(n_{ox',-dn} + \pi_x)} \cdot \frac{\Pi_{x'}\Gamma(n_{ox'} + \pi_x)}{\Gamma(\sum_{x'} n_{ox'} + N_x\pi_x)} \tag{8}$$

where π_x is the hyperparameter for β such that $\beta \sim Dir(\pi_x)$. The value $n_{ox,-dn}$ indicates the number of occurrences for x with o excluding the instances related to dn.

3.3 Inference

With the trained model, we predict the most likely scene class for an image from the new test image set. We use the visual components of the proposed model to compute the posteriori probability of each scene class by integrating out the latent object variables, Os and tOs:

$$\begin{aligned} P(C = c|\mathbf{tR}, \mathbf{tX}) &= \frac{P(C = c, \mathbf{tR}, \mathbf{tX})}{P(\mathbf{tR}, \mathbf{tX})} \\ &\propto \prod_{n=1}^{N_r} \sum_{\{O_1, O_2, \dots, O_{N_o}\}} \big(P(O_1, O_2, \dots, O_{N_o}|C = c) \cdot \\ &\sum_{o} P(\mathbf{tR}_n|tO_n = o) \cdot P(\mathbf{tX}_n|tO_n = o) \cdot P(tO_n = o|O_1, \dots, O_{N_o})\big) \end{aligned} \tag{9}$$

where the first summation is marginalizing over all possible configurations of objects under the scene class c and the second one is marginalizing over all possible object assignments under a particular configuration. The most likely scene class can then be determined as:

$$c^* = \arg\max_{c \in \mathcal{C}} P(C = c|\mathbf{tR}, \mathbf{tX}) \tag{10}$$

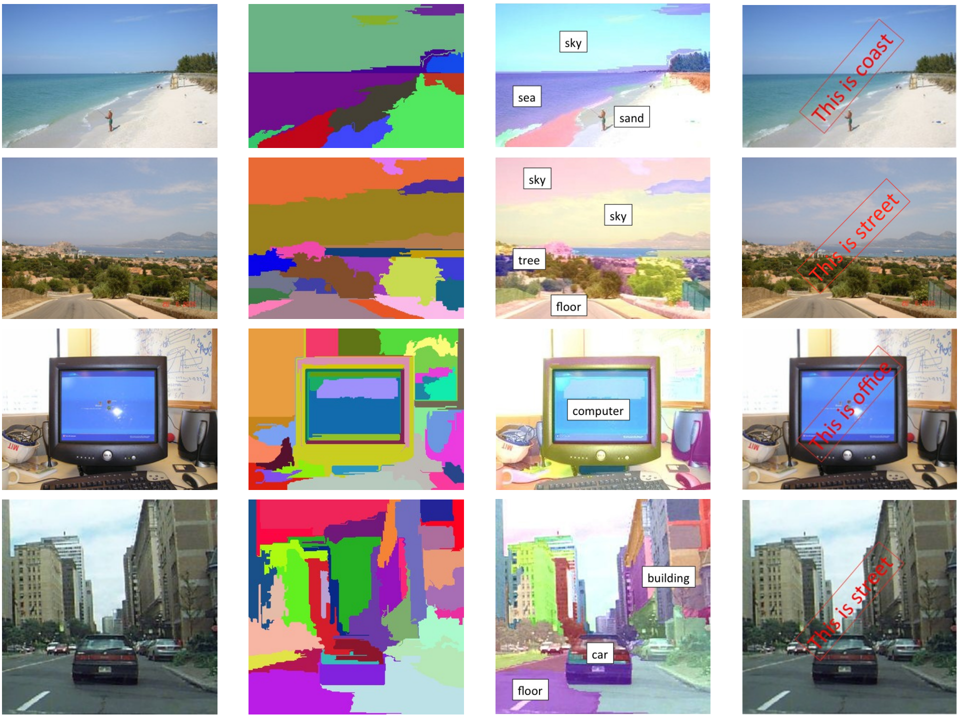
3.4 Ensemble Prediction

Same as many probabilistic graphical models developed for image modeling in the literature, the presented model above can only be trained to reach local optimal solutions due to the existence of many latent variables. In order to make the predictions more robust, instead of training only one model, we train multiple models (we used 5 in our experiments) by random restarts and *ensemble* them together in the inference phase. With ensemble prediction, the predicted scene value for a test image is determined as the scene label voted by most models.

4 Experiments and Results

We evaluated our approach on the widely used LabelMe dataset [15], out of which we picked 10 scene categories that contain both outdoor and indoor scenes: *bathroom, bedroom, airport, coast, corridor, livingroom, office, park, speech* and *street*. Each category contains $70 \sim 200$ images. We selected 50 images from each category to form the training set and kept the remaining images in the test set. We set the number of object categories as 30, i.e., $N_o = 30$, in our experiments.

Figure 3 presents a few examples of our experimental results on the test data. The results suggest that the automatic object annotation achieved in our model is helpful for scene



(a) (b) (c) (d)

Figure 3: The scene recognition results. Column (a) are the original images; Column (b) are the segmentation results; Column (c) are the object-level intermediate results; Column (d) are the scene classification results.

classification even in the case that the image context can not be completely captured. For example, in the third row of Figure 3, though only one object "computer" (reflected as a class index in our model) is recognized, our system is able to correctly tell the image is "office" since computers often appear in the "office" scene instead of other scenes. On the other hand, the missing of some critical objects may lead to wrong scene classification. For example, in the second row of Figure 3, the "sea" object is missed from the object level, and the predicted scene label is different from the ground truth "coast", even the objects *sky, tree,* and *floor*, are correctly detected. This however is a hard case even for human being. Nevertheless, overall our automatic object recognition still positively contributes to the final scene classification and leads to good prediction results.

To measure our prediction results, we compared our proposed model with three baseline methods: (1) The state-of-the-art scene classification method using SVM Classifier and Object Bank features (SVM+OB) [7]. The OB features are obtained by applying pretrained object detectors on input images, and then SVM classifier is used for scene classification with these features. From the set of well established object detectors [7], we manually picked 30 object detectors that are most closely related to the 10 scene categories of our dataset. Note the object detectors used in this method are supervised trained on multiple datasets including ESP, LabelMe, ImageNet, Flickr, etc. This method therefore obviously benefits from both the large amount of labeled data and the prior knowledge on the LabelMe dataset, since accurate object detections are critical for scene classification. We can take this method as

Method	Proposed	Proposed w/o Ensemble	Model w/o Chain	OB+SVM
bathroom	0.765	0.604	0.565	0.938
bedroom	0.704	0.673	0.573	0.568
airport	0.676	0.638	0.584	0.459
coast	0.920	0.875	0.607	0.534
corridor	0.786	0.757	0.550	0.964
livingroom	0.471	0.464	0.447	0.765
office	0.938	0.822	0.675	0.938
park	0.660	0.749	0.630	0.849
speech	0.769	0.592	0.438	0.846
street	0.718	0.688	0.425	0.875
Average	0.741	0.686	0.549	0.774

Table 1: Comparison results of scene classification in term of test accuracy.

our *golden* standard. (2) The proposed model without ensemble. We randomly pick one out of the trained multiple models to perform classification so as to evaluate the effect of the simple ensemble technique on the performance of our system. (3) The proposed ensemble model without the object chain structure. We simply remove the chain structure from the model in Figure 2 to analyze the benefit gained by capturing object co-occurrence contextual information.

The test prediction accuracies of all four methods are presented in Table 1. We can see that the average performance of our approach is almost as good as the golden standard method OB+SVM, which used supervised object detectors. In three out of the ten classes, our unsupervised model even greatly outperforms OB+SVM. We have also tested an unsupervised setting for OB+SVM where we randomly picked 30 object detectors to generate OB representations for images and keep other procedures same as before. Its performance drops dramatically so that we do not report those numbers in this paper. Comparing to the other two baselines, the models without either the ensemble strategy or the chain structure, our proposed model produces consistent superior performances across almost all scene classes. The model without the chain structure produces the worst performance among all four methods. These results suggest the object co-occurrence information is valuable for scene classification, while the ensemble strategy can effectively increase the model robustness.

5 Conclusion

We have presented a hierarchical probabilistic graphical model to perform scene classification. The proposed model can achieve automatic and implicit object annotation during the training phase so as to save human effort of image annotation. Moreover, the contextual information in form of object co-occurrence is explicitly represented by a probabilistic chain structure in our model, and the issue of local optima is addressed using a simple ensemble strategy. Our experimental results on the LabelMe dataset suggest that accurate object annotations are important for scene classification, and the object co-occurrence information captured in our proposed model contributes to great improvements over the test performance. Overall, our proposed model demonstrated effective empirical performance, even comparing to the state-of-the-art OB+SVM method that exploits supervised object detectors.

References

[1] A. Bosch, A. Zisserman, and X. Munoz. Scene Classification via pLSA. In *Proceedings of ECCV*, 2006.

[2] A. Bosch, X. Muñoz, and R. Martí. Which is the best way to organize/classify images by content? *Image and Vision Computing*, 25(6):778 – 791, 2007.

[3] E. Chang, K. Goh, G. Sychay, and G. Wu. Cbsa: Content-based soft annotation for multimodal image retrieval using bayes point machines. *IEEE Transactions on Circuits and Systems for Video Technology*, 13:26–38, 2003.

[4] P. Espinace, T. Kollar, A. Soto, and N. Roy. Indoor scene recognition through object detection. In *Proceedings of ICRA*, 2010.

[5] P. Felzenszwalb and D. Huttenlocher. Efficient graph-based image segmentation. *Int. J. Comput. Vision*, 59(2):167–181, September 2004.

[6] X. He, R. Zemel, and M. Carreira-Perpinan. Multiscale conditional random fields for image labeling. In *Proceedings of CVPR*, 2004.

[7] E. Xing L. Li, H. Su and F. Li. Object bank: A high-level image representation for scene classification & semantic feature sparsification. In *Proceedings of NIPS*, 2010.

[8] Y. Lim L. Li, H. Su and F. Li. Objects as attributes for scene classification. In *Proceedings of ECCV*, 2010.

[9] F. Li and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *Proceedings of CVPR*, 2005.

[10] L. Li, R. Socher, and F. Li. Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. In *Proceedings of CVPR*, 2009.

[11] D. Lowe. Object recognition from local scale-invariant features. In *Proceedings of ICCV*, 1999.

[12] T. Malisiewicz and A. Efros. Recognition by association via learning per-exemplar distances. In *Proceedings of CVPR*, 2008.

[13] T. Malisiewicz and A. Efros. Beyond Categories: The Visual Memex Model for Reasoning About Object Relationships. In *Proceedings of NIPS*, 2009.

[14] A. Quattoni and A. Torralba. Recognizing indoor scenes. In *Proceedings of CVPR*, 2009.

[15] B. Russell, K. Torralba, A.and Murphy, and W. Freeman. Labelme: A database and web-based tool for image annotation. *Int. J. Comput. Vision*, 77(1-3):157–173, May 2008. ISSN 0920-5691.

[16] N. Serrano, A. Savakis, and J. Luo. Improved scene classification using efficient low-level features and semantic cues. *Pattern Recognition*, 37(9):1773–1784, 2004.

[17] J. Shen, J. Shepherd, and A. Ngu. Semantic-sensitive classification for large image libraries. In *Proceedings of MMM*, 2005.

[18] A. Torralba. Contextual priming for object detection. *Int. J. Comput. Vision*, 53(2): 169–191, July 2003. ISSN 0920-5691.

[19] A. Torralba, K. Murphy, and W. Freeman. Using the forest to see the trees: exploiting context for visual object detection and localization. *Commun. ACM*, 53(3):107–114, 2010.

[20] I. Ulrich and I. Nourbakhsh. Appearance-based place recognition for topological localization. In *Proceedings of ICRA*, 2000.

[21] A. Vailaya, M. Figueiredo, A. Jain, and H. Zhang. Image classification for content-based indexing. *Image Processing, IEEE Transactions on*, 10(1):117 –130, January 2001.

[22] L. Wolf and S. Bileschi. A critical view of context. *Int. J. Comput. Vision*, 69(2): 251–261, August 2006.