An Object Co-occurrence Assisted Hierarchical Model for Scene Understanding

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

Computer and Information Science
Temple University
Philadelphia
PA 19122, USA

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 [1]. 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 [4]. 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 [2, 3]. In this work, we propose a novel three-level (superpixel level, object level and scene level) generative hierarchical model for scene understanding, which does not require tedious object annotations over the training data.

The proposed hierarchical probabilistic graphical model, shown in Figure 1, 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. It captures the high-level contextual information expressed in form of object co-occurrences using a probabilistic chain structure over the object class assignment variables in each image. The rationale behind this design is to capture the possible object category correlation information without inducing more complicated inference problems.

In the model setting, the total number of different objects for the whole image set is assumed to be known. But as an unsupervised model at the object level, 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 particular, 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. Unsupervised learning at the object-level is expected to automatically capture useful concepts for each object class.

Following the proposed model, the resulting joint distribution of a given scene with class C, the appearances of object classes, O_1, O_2, …, O_n, the objects $\mathbf{tO}$, the region features $\mathbf{tR}$, and the image patch features $\mathbf{tX}$ can be expressed as:

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

To learn the model parameters automatically, we derive a collapsed Gibbs sampling algorithm. Details are discussed in the paper.

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 variable Os and tOs:

$$P(C = c | \mathbf{tR}, \mathbf{tX}) = \frac{P(C = c, \mathbf{tR}, \mathbf{tX})}{P(\mathbf{tR}, \mathbf{tX})} \propto \prod_{n=1}^{N_r} \sum_{\{O_1, O_2, \ldots, O_{N_o}\}} \big(P(O_1, O_2, \ldots, O_{N_o} | C = c) \cdot \sum_{o} P(\mathbf{tR}_n | tO_n = o) \cdot P(\mathbf{tX}_n | tO_n = o) \cdot P(tO_n = o | O_1, \ldots, O_{N_o})\big) \quad (2)$$

The most likely scene class can then be determined as:

$$c^* = \arg\max_{c \in \mathcal{C}} P(C = c | \mathbf{tR}, \mathbf{tX}) \quad (3)$$

Moreover, we exploit an ensemble prediction strategy by training multiple models and take the majority vote of the multiple models as the final scene label of the test image.

The proposed model is evaluated on the LabelMe dataset, comparing to a golden standard method OB+SVM, which used supervised object detectors, and two other baselines, the models without either the ensemble strategy or the chain structure. The test accuracy results are presented in Table 1, which show the proposed approach is almost as good as the OB+SVM, and produces consistent superior performances over the other two baseline methods.

Method	Proposed	w/o Ensemble	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.

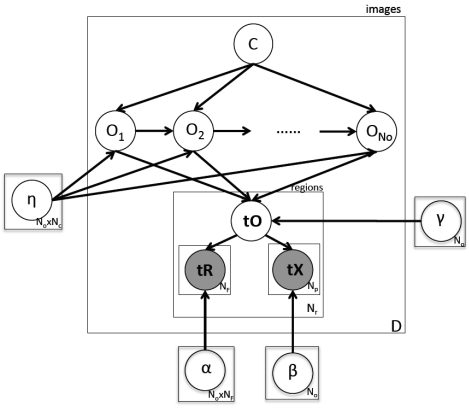

Figure 1: The proposed model. Nodes denote random variables and edges indicate dependencies. The variables at the right lower corner of each box denote the numbers of replications. The box indexed by D represents a single image in the image set of size D. The box indexed by N_r denotes the visual information 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.

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

[2] 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.

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

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