Adaptive hierarchical contexts for object recognition with conditional mixture of trees

Billy Peralta
bmperalt@uc.cl
Pablo Espinace
pespinac@uc.cl
Alvaro Soto
asoto@ing.puc.cl

Computer Science Department
Pontifica Universidad Catolica de Chile
Santiago, Chile

Contextual information has emerged as an attractive option to boost the performance of single object category detectors [1][2]. Regarding to these techniques, Choi et al. [1] presents an efficient scheme to model inter-object relations using a tree-structured Bayesian network. Recently, [2] uses contextual cueing, spatial co-ocurrence, and inhibitory intra-class constraints among objects using a max-margin approach. In all these cases, contextual relations among objects are fixed and do not depend of the type of scene being analyzed. We propose that using an adaptive scheme to model contextual relations among objects can boost the performance of current object recognition techniques. We can illustrate this idea by the following example. Consider the case of the contextual relation between the presence of a person and a dog objects. Under a park scene, person and dog objects co-occur frequently, but in an office scene, they hardly co-occur, therefore modeling such relation with a fixed contextual constraint limits the flexibility of the model to fit real data.

In this work, we present a method that learns adaptive conditional relationships among objects according to the underlying scene information. To achieve this goal, we use a probabilistic model based on a conditional mixture of trees [4]. Our work is based on an extension of the scheme proposed by Choi et al. [1]. In that work, the authors use a single tree graphical model to represent dependencies among objects. In contrast, our mixture of tree allow us to model different contextual relations among object categories.

Our mixture of trees context model is built by a conditional mixture of tree-structured Bayesian Networks, each of which is an expert in some partition of the set of images. These networks have a weight that depends on the global scene information given by a Gist feature vector x_G. The model can be seen as a mixture of experts where the gate function is given by a function of the global representation, and each expert function is given by an individual Bayesian network.

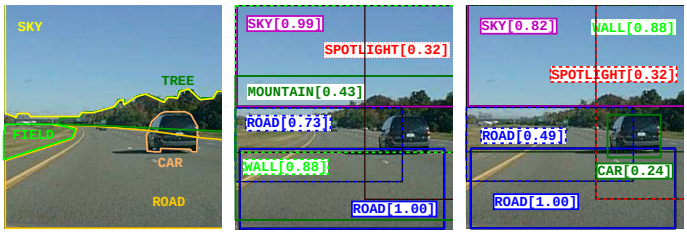
In order to incorporate global scene information in our model, we model object presence as dependent on scene type. In particular, for a training set of N images, we can construct N instance-label pairs (x_G, b), where $b \in \{0,1\}^D$ represents presence in a given image of instances of D possible object categories. Our goal is to use instances (x_G, b) to include in our model different types of contextual relations between object classes. We achieve this goal by introducing latent variable z which simplifies the analysis of the model. We refer to this latent variable as the *context variable*. We assume that there are K possible values for z, i.e., we assume the existence of K *contexts* for object classes. This is similar to a mixture of experts model with the exception that in our case b is conditionally independent of x_G given z. The context variable is assumed as a winner-take-all variable, i.e., each object class detection occurs under a specific contextual scenario. Considering K contexts, we can model the conditional density as :

$$p(b|x_G) = \sum_{i=1}^{K} p(b, z_i|x_G) = \sum_{i=1}^{K} p(b|z_i)\, p(z_i|x_G) \tag{1}$$

Here, we have the K contexts represented by z_i with $i = \{1,...,K\}$, where each context has its own class-conditional probability function. The two components of the mixture model given by Equation 1 are the context gate function, given by $p(z_i|x)$, and the tree experts function, given by $p(b|z_i)$. In particular, $p(z_i|x)$ is modelled as a normalized Gaussian kernel. On the other hand, $p(b|z_i)$ is modelled as a tree-structured Bayesian Network [4].

Assuming that the posterior probabilities of context gates R_{in} (*responsabilities*) for each expert i and training instance n are known, we apply the EM algorithm over the expected log-likelihood $\langle L_c \rangle$ in order to obtain the parameters:

$$\langle L_c \rangle = \sum_{n=1}^{N} \sum_{i=1}^{K} R_{in}\, log\,(\, p(b_n|z_i)\, P_i(x_n)\, \alpha_i) \tag{2}$$



(a) Ground-Truth Image (b) Results according to [1](c) Results with Mixture of trees

Figure 1: An example of the results of our method with respect to a state-of-the-art model [1]. Our model 1(c) adaptively selects a suitable component of a mixture of trees for providing a correct detection of the car object and do not detect a phantom mountain object as in [1].

Inference is straightforward, as we separate each tree in its own partition. We make inference using message passing algorithms for each tree $(p(b,c,L/g,W,s,z))$ in an iterative fashion as in [1]. Then we obtain the final score by combining the scores of each component with its respective parameters, similar to the work of Meila and Jordan [4].

We perform an empirical evaluation of the proposed approach considering two real public datasets: (i) OUTDOOR dataset created by Oliva and Torralba [5], and (ii) SUN09 dataset publicated by Choi [1]. We employ the object detector proposed by Felzenszwalb et al. [3]. Our work use average precision-recall (APR) as a performance metric for our model. This metric corresponds to the area under the precision-recall curve.

In relation to our experiments, we achieve the best performance using 6 trees. In this case, the relative improvements in terms of APR with respect to [1] are 5.5% and 5.7% for OUTDOOR and SUN09 datasets, respectively. In terms of individual classes, for the case of 6 trees, we notice that with respect to [1] APR increases for 10 and 53 objects and decreases for 6 and 34 objects for the OUTDOOR and SUN09 datasets, respectively.

As a main conclusion, our experiments using standard object datasets indicate that the proposed model improves object recognition performance with respect to a single tree model. This validates our main hypothesis indicating the relevance of including adaptive contextual relations to boost the performance of object category detectors.

[1] M. Choi, J. Lim, A. Torralba, and A. Willsky. Exploiting hierarchical context on a large database of object categories. In *Proc. CVPR*, pages 129–136, 2010.

[2] C. Desai, D. Ramanan, and C. Fowlkes. Discriminative models for multi-class object layout. *IJCV*, 95:1, 2011.

[3] P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *Proc. CVPR*, 2008.

[4] M. Meila and M. Jordan. Learning with mixtures of trees. *JML*, 1: 1–48, September 2001.

[5] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 42:145–175, May 2001. ISSN 0920-5691.