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

Abstract

Robust category-level object recognition is currently a major goal for the computer vision community. Intra-class and pose variations, as well as, background clutter and partial occlusions are some of the main difficulties to achieve this goal. Contextual information, in the form of object co-occurrences and spatial constraints, has been successfully applied to improve object recognition performance, however, previous work considers only fixed contextual relations that do not depend of the type of scene under inspection. In this work, we present a method that learns adaptive conditional relationships that depend on the type of scene being analyzed. In particular, we propose a model based on a conditional mixture of trees that is able to capture contextual relationships among objects using global information about a scene. Our experiments show that the adaptive specialization of contextual relationships improves object recognition accuracy outperforming previous state-of-the-art approaches.

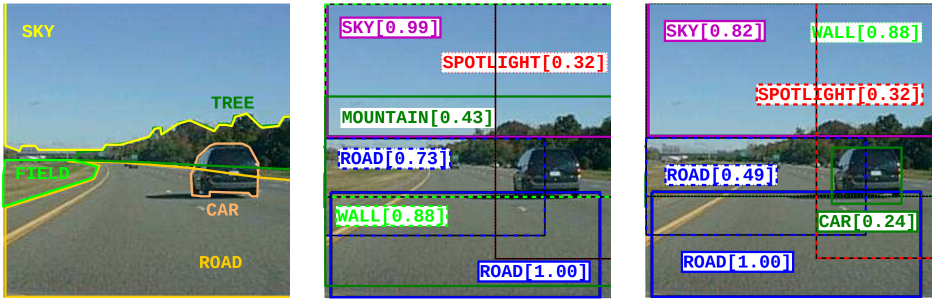
1 Introduction

Lately, the synergistic combination of computer vision and machine learning techniques have been successfully applied to the problem of automatic visual recognition [20] [9] [7] [8]. In particular, contextual information has emerged as an attractive option to boost the performance of single object detectors [10][2][5].

Context based methods can be divided into two groups: global and local context methods [10]. Regarding global or holistic context methods, most works exploit whole scene statistics to perform recognition. In [19], Ulrich and Nourbakhsh introduce color histograms as the holistic representation of an image that is used by a K-nearest neighbors scheme to classify scenes. In [17], Torralba proposes an image representation based on global features that represent dimensions in a space that they call spatial envelope. In [1], Chang et al. use low-level global features that are used to estimate a belief or confidence function over scene labels.

Regarding local context techniques, contextual information is derived from specific blocks or localized areas around object positions. Sinha and Torralba [16] improve face detection

http://dx.doi.org/10.5244/C.26.121



(a) Ground-Truth Image (b) Results with Single Tree [2] (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 [2]. We show the top six most confident detections for each model using the same underlying single object detector [8]. In the case of 1(b), the model uses fixed contextual relations that do not detect the car object and produce a wrong detection of a mountain object. In contrast, our model 1(c) uses global scene information to adaptively select a suitable component of a mixture of trees that embeds particular contextual relations that provide a correct detection of the car object and do not detect a phantom mountain object.

using local contextual regions. Torralba et al. [18] introduce a Boosting approach in combination with a Conditional Random Field (CRF) to recognize objects. They apply their method to recognize objects and structures in office and street scenes. Shotton et al. [15] combine layouts of textures and context to recognize objects. They use a CRF to learn a model of objects and a boosting algorithm to combine the texture information and the object model. Galleguillos et al. [10] present a critical review of different contextual cues and machine learning models commonly used to improve object categorization. Rabinovich et al. [14] show that textual data from the web is a useful source to estimate co-ocurrence between objects. Choi et al. [2] presents an efficient scheme to model inter-object relations using a tree-structured Bayesian network. Recently, [5] shows a technique that is able to model 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 believe 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. Moreover, in terms of a probabilistic graphical model (PGM) representation for object relations, such as [2], the relevance of the information provided by each type of object can change dramatically for different types of scenes. For example, in an office scene a computer monitor is commonly a highly informative object, therefore under a suitable PGM representation it should have strongly related children objects. In contrast, in situations like a living room scene, a monitor is usually not very informative, therefore under a suitable PGM representation its related children structure is not very relevant.

In the previous cases, the flexibility of a mixture model able to provide adaptive contextual relations among objects can be a useful tool to boost object recognition performance.

Consequently, in this work, we present a method that learns adaptive conditional relationships among objects according to the scene information. In particular, we achieve this by introducing a PGM based on a conditional mixture of trees [11]. Next, we provide first background information relevant to our model. Afterwards, we describe the details of our approach. Finally, we present the results of our experiments showing the advantages of our approach.

2 Hierarchical context

In this section, we summarize the work by Choi et al. [2] that is the baseline algorithm considered in our work. The model in [2] is composed of a prior and a measurement model. Next we provide details about prior and measurement models. Note that we use b_i to refer to a specific object class i, while we use B to refer to a generic object class.

2.1 Prior Model

The prior model uses a binary tree structured PGM to represent co-ocurrence and spatial relationships among object categories. Nodes in this tree are given by variables indicating object presence, as well as, its location and scale (see [2] for details). Specifically:

1. **Object presence**: $b_i \in \{0,1\}$ corresponds to the presence of an object of class i.

2. **Object location and scale**: L_i corresponds to the location and scale of object instance b_i. $L = \{L_i, \ldots, L_N\}$ resumes all object classes. L is modelled as dependant of the presence of objects b: $p(L|b) = p(L_{root}|b_{root}) \prod_i p(L_i|L_{pa(i)}, b_i, b_{pa(i)})$, where L_i is the median of the location and scale for all instances of object i and is composed by $(L_y^i, \log L_z^i)$. L_y^i is the median of vertical positions for object i and L_z^i is the median of scales for object i. Medians are computed using all training images. The use of a logarithm for scales and the omission of horizontal positions is justified in [2].

2.2 Measurement Model

The measurement model predicts the presence of an object category b_i in an image by using global gist features and outputs of object detectors. Figure 2(a) shows the PGM that relates the variables considered in the measurement model. Specifically:

1. **Correct detections**: $c_{ik} \in \{0,1\}$ represents the k-th detection of instances of object category i, being 1 if the detection is a true positive and 0 otherwise. Correct detections depend on object presence, where $p(c_{ik} = 1|b_i = 1)$ corresponds to the frequency of correct detections in the training set and $p(c_{ik} = 1|b_i = 0) = 0$.

2. **Classifier scores**: $s_{ik} \in \Re$ represents classifiers scores, which according to Figure 1 depends on correct detections c_{ik}. Using Bayes rule, $p(s_{ik}|c_{ik}) = p(c_{ik}|s_{ik})p(s_{ik})/p(c_{ik})$. Here a logistic regression is used to model $p(c_{ik}|s_{ik})$.

3. **Detection window location**: $w_{ik} = (L_y^{ik}, \log L_z^{ik})$ represents the location of a detection window, where L_y^{ik} and L_z^{ik} are vertical location and scale of the window corresponding to the k-th detection of an instance of object category i. Location is modeled

as a Gaussian distribution and it depends on c_{ik} and the median location L_i of instances of object class i. If a window is a correct detection then w_{ik} is modeled as $p(w_{ik}|c_{ik} = 1, L_i) = Gaussian(w_{ik}; L_i, \Lambda_i)$, where Λ_i is the covariance around the predicted location. If a window is a false positive then w_{ik} does not depend on L_i and it is modeled with a uniform distribution.

4. **Gist**: Gist features g_L [17] are used to related global image features to object presence by estimating $(g_L|b_i$. To deal with the high dimensionality of the gist vector g_L, a logistic regression is used to estimate $p(b_i|g)$, then likelihoods $p(g_L|b_i)$ are estimated indirectly using $p(g|b_i) = p(b_i|g)p(g)/p(b_i)$.

   Following the notation in [2], from here on we use variables without subindexes to denote the set of variables related to individual object class detections in a image. For example, $= b\{b_1, \ldots, b_N\}$ denotes the binary values of all variables related to the detection of the D possible object classes. In the same way, W resumes all candidate detection windows variables w_{ik} in a given image.

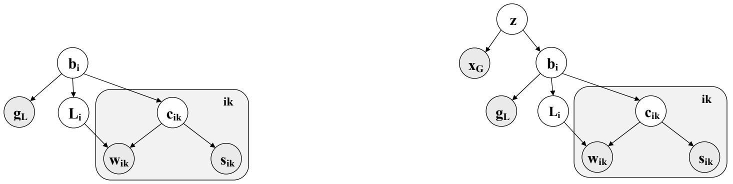
3 Our model

As mentioned earlier, an important limitation of the method by Choi et al. [2] is that it assumes a fixed contextual relationship among objects. In this work, we avoid this limitation by incorporating in the model adaptive contextual relationships between objects that depend on a estimation of the current scene type. Our main intuition is that contextual object-to-object co-occurrences strongly depend on the underlying scene, as shown in the person and dog example mentioned before. In particular, we propose to modify the fixed single tree co-occurrence model by Choi et al., using a model based on a mixture of trees. This mixture of trees incorporates scene information to adaptively represent different possible contextual relations between objects. In terms of the original PGM considered [2], our main modification is the incorporation of a latent variable representing the underlying scene type. This latent variable depends on the output of a Gist feature [17]. Figure 2(b) shows the resulting modified PGM after adding the new latent variable. Also global features x_G are added to the PGM, as observation to infer the scene type.

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 the Gist feature). 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 networks. We stress that our main contribution is the joint modeling of the dependence between the ensemble of trees and a global representation of the image data. Next, we provide details of the proposed conditional mixture of trees model and how we conduct estimation and inference with this model.

3.1 Conditional mixture of trees

In order to incorporate global scene information in our model, we model object presence as dependent on scene type. In our current implementation, we infer scene type using a Gist feature vector x_G. In this sense, in our PGM $x_G = g_L$, however, other global features can



(a) A single tree model

(b) Conditional mixture of trees model

Figure 2: Modification of contextual objects relationships: (a) The model by Choi et al. (b) Proposed model that incorporates global scene information as a root element that influences object-to-object relationships.

be used to estimate scene type, therefore we consider x_G and g_L as separate variables in the PGM shown in Figure 2(b).

For a training set of N images, we can construct N instance-label pairs (x_G, b), where, as stated before, $b \in \{0,1\}^D$ represents the potential presence of the D possible object categories in a given image. 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 and using a Gaussian Kernel for the weighting of experts, we can model the conditional density $p(b|x_G)$ [21] as:

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

Here, we have K contexts represented by z_i, with $i = \{1, ..., K\}$, where each context has its own class-conditional probability function.

We specify the two components of the mixture model given by Equation 1 as follows:

- ***Context gate:*** given by $p(z_i|x_G)$, represents the influence of each local context. It represents an estimate of the likelihood of selecting each of the K experts for the input x_G. The gate function has K components, one for each expert.

- ***Tree experts:*** given by $p(b|z_i)$ represents the class-conditional local models. It represents an estimate of the probability of appearances b given the expert z_i for input x_G. There are K context functions. In this case, we use a Bayesian Network model following Choi et al [2].

The proposed model is similar to the mixture of trees model presented by Meila and Jordan [11]. A mixture of trees model represents the distribution of a convex sum of K tree components over a random variable x as: $Q(x) = \sum_{k=1}^{K} \lambda_k T^k(x)$ with $\lambda_k \geq 0$ and $\sum_{k=1}^{K} \lambda_k = 1$. Tree distributions $T^k(x)$ are the mixture components and coefficients λ_k are the mixture proportions. This model can be viewed as containing a latent variable z that with probability λ_k selects mixture component k. Therefore, conditioned on the value of z, the distribution of mixture Q is represented by a single tree. An advantage of this model is its flexibility because the trees may have different structures and parameters. Nonetheless, the main difference with our work is that while in [11] they assume the weight of each component as fixed, we model these weights as variable. In particular, in the proposed model these weights depends on the global representation of a given image using the context gates.

Regarding context gate function, we use normalized Gaussian Kernels [21]. This function can be interpreted as a simple mixture model. In this case, $p(z_i|x) = \frac{\alpha_i P_i(x)}{\sum_j \alpha_j P_j(x)}$, where each P_i is a Gaussian probability density functions with weights α_i, $\sum_j \alpha_j = 1$ and $\alpha_i \geq 0$.

In the case of the tree expert function, we use a Bayesian Network function similar to Meila and Jordan [11]. Following their work, we parameterize a tree with a graph $G = (V;E)$, where V is the vertex set and E is the set of edges. Assuming a set of K trees, V_i represents the vertex set of tree i, where $i \in \{1,\dots,K\}$. In the case of our adaptive contextual model, the probability distribution of variable b conditioned on the context variable z_i, $p(b|z_i)$, is represented by $T^i(b)$. Each tree models this component as $T^i(b) = \prod_{v \in V^i} T_{v|pa(v)}(b_v|b_{pa(v)})$, where $T_{v|pa(v)}(b_v|b_{pa(v)})$ is an arbitrary conditional distribution. Variable $pa(v)$ represents the parents of variable v inside the tree.

In order to find optimal parameter values for tree experts and the gate function, we use the Expectation Maximization (EM) algorithm [4]. To obtain a one-pass solution for the gating function, instead of the conditional log-likelihood for the complete data, we use the joint log-likelihood [21]. Assuming that the posterior probabilities of context gates or *responsabilities* R_{in} for each expert i and training instance n are known, we can apply the EM algorithm over the expected log-likelihood:

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

The expectation step of EM is given by the calculation of the posterior probability of context gate i, which is given by:

$$R_{in} = p(z_i|x_n, b_n) = \frac{p(b_n|z_i)\, p(z_i|x_n)}{\sum_{j=1}^{K} p(b_n|z_j)\, p(z_j|x_n)} \tag{3}$$

The maximization step of EM is given by the maximization with respect to each parameter. We can observe two decoupled components in the expected log-likelihood:

$$E_{expert} = \sum_{n=1}^{N} \sum_{i=1}^{K} R_{in} \left[log\, T^i(b_n) \right],\; E_{gate} = \sum_{n=1}^{N} \sum_{i=1}^{K} R_{in} \left[log\,(\alpha_i P_i(x_n)) \right] \tag{4}$$

In the case of the tree expert component, we must minimize the negative cross-entropy between R and T. Following the work of Meila and Jordan [11], this problem is solved using a weighted version of the Chow-Liu algorithm. This component requires K runs of the Chow-Liu algorithm, where R_{in} is the normalized posterior probability obtained in the E-step.

In order to find the parameters of the gating function, α_i, μ_i and σ_i (being μ_i and σ_i the parameters of the Gaussian representation), we use Equation 2, obtaining:

$$\alpha_i = \frac{1}{N}\sum_{n=1}^{N} R_{in}, \quad \mu_i = \frac{\sum_{n=1}^{N} R_{in}\, x_n}{\sum_{n=1}^{N} R_{in}}, \quad \sigma_i = \frac{1}{d}\frac{\sum_{n=1}^{N} R_{in}\, \|x_n - \mu_i\|^2}{\sum_{n=1}^{N} R_{in}} \tag{5}$$

Algorithm 1 Conditional Mixture of Context Trees

```
while Not convergence do
    Compute responsabilities $R$ according to Equation 3
    for $i = 1 \to K$ do
        Estimate $\alpha$,$\mu$ and $\sigma$ according to Equation 5 and $T$ with a weighted Chow Liu algorithm according to [11].
    end for
end while
```

The operation of the EM algorithm for a conditional mixture of trees is summarized in Algorithm 1. We initialize the gates using a variant of the K-means algorithm that clusters variables b over the training set using Hamming distance. The resulting conditional mixture of context model follows the intuition that general context is naturally divided into many component contexts, thus, we can make inference on each tree and then combine the outputs using the gating function.

3.2 Inference

Inference is straightforward, as we separate each tree in its own partition. Similarly to [2], we make inference using message passing algorithms for each tree $(p(b,c,L/g,W,s,z))$ [13]. Afterwards, similarly to [11], we obtain the final score by combining the scores of each component with its respective parameters.

$$\hat{b},\hat{c},\hat{L} \quad = \quad argmax_{b,c,L}\sum_{z} p(z) * p(b,c,L/g,W,s,z) \tag{6}$$

Following Choi et al. [2], we use an iterative procedure. First, we make inference without considering the locations ($\hat{b_0},\hat{c_0} \propto p(b,c|g,s)$), then we infer the locations ($\hat{L} \propto argmax_L p(L|\hat{b_0},\hat{c_0},W)$), and finally we infer the object presence ($\hat{b},\hat{c} \propto p(b,c|s,g,\hat{L},W)$) considering the previous inferred location. The last step is equivalent to sampling from a binary tree with node and edge potentials modified by $p(\hat{L},W/b,c)$.

4 Experiments

In this section, we perform an empirical evaluation of the proposed approach considering two real datasets: (i) OUTDOOR dataset created by Oliva and Torralba [12], and (ii) SUN09 dataset created by Choi [2]. OUTDOOR dataset has 2600 images and includes 8 outdoor scene categories, such as coast, mountain, forest, etc. We randomly divide the dataset into two sets of approximately equal size, one for training and one for testing. Similarly to

Choi et al. [2], we prune object categories by considering only those that have at least 3 true detections in the training set. As a result, for OUTDOOR dataset we have 21 object categories with equal-sized training and test sets. In the case of SUN09 dataset, as in [2], we prune the dataset considering only object categories with at least 4 true detections in the training dataset. As a result, for SUN09 dataset we have 111 object categories, 4367 training images, and 4317 test images.

In general, in both datasets object detections are highly challenging, including a variety of poses, scales, rotations, and scene types. We use the object detector proposed by Felzenszwalb et al. [8], which is based on the mixture of multi-scale deformable parts model and a latent SVM approach. We use the same object detector models for both datasets. In average, this detector outputs approximately 5 detections per category in each image. In both datasets, for each image we consider the top 10 detections. We use the average precision-recall (APR) [3] as a performance metric for our model. This metric corresponds to the area under the precision-recall curve.

Table 1 shows APR for both datasets: OUTDOOR and SUN09. We show the resulting APR for: i) Direct object detections provided by the underlying object class detector [8] (*Object detector*), ii) Choi et al. [2] method based on hierarchical context (*Single Tree*), iii) Our proposed method using different number of trees (*MixTree-X*, where *X* is the number of used trees). Relative improvement in APR with respect to Choi et al. is shown in parenthesis.

Method	**OUTDOOR**	**SUN09**
Object detector	14.02 (-6.5%)	6.82 (-13.2%)
Single Tree	15.00 (0.0%)	7.87 (0.0%)
MixTree-2	15.07 (0.5%)	7.98 (1.5%)
MixTree-3	14.87 (-0.9%)	8.09 (2.9%)
MixTree-4	15.12 (0.8%)	8.06 (2.5%)
MixTree-5	15.25 (1.7%)	8.03 (2.2%)
MixTree-6	**15.83 (5.5%)**	**8.31 (5.7%)**
MixTree-7	14.84 (-1.1%)	7.88 (0.3%)

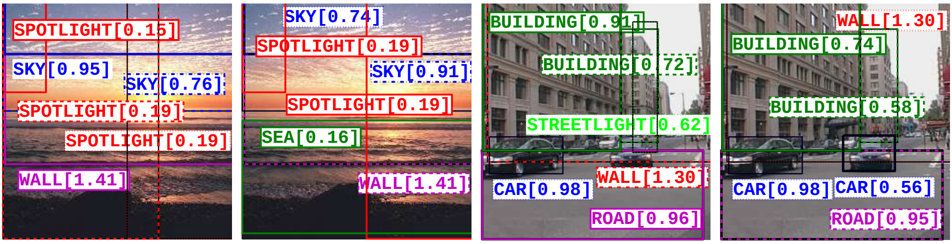
Table 1: APR for OUTDOOR and SUN09 databases provided by the tested methods. Relative improvement with respect to Choi et al. is shown in parenthesis.

Analyzing Table 1, in OUTDOOR database we note that performance improves as the number of trees grows up to 6. After that, APR decays. In general, results improve the performance of Choi et al. [2]. Considering the best number of trees in this dataset (six), relative improvement is 5.5%. In the case of SUN09 database, the improvement with respect to Choi et al. in terms of APR is 5.7% for the best number of trees (also six). It is important to note that in this case adding additional trees does not necessarily improve performance. In terms of individual classes, we found for the case of six trees that in OUTDOOR and SUN09 dataset, the APR increases for 10 and 53 object classes and decreases for 6 and 34 classes objects, respectively.

Figure 3 shows example detections of the top six most confident detectors from our model and Choi et al. [2]. For example in Figure 3(b) the adaptive scheme correctly detect a sea object that is not detected by the single tree model in Figure 3(a). In Figure 3(d) a car object is recovered and a streetlight object is discarded in relation to Figure 3(c).

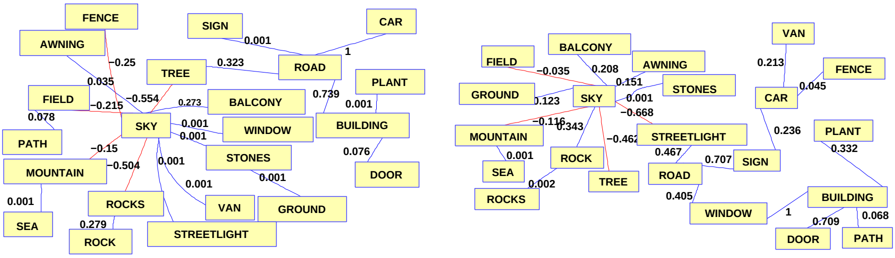
Figure 4 presents examples of resulting trees for OUTDOOR database. The value over each edge represents the strength of dependency relation between each pair of object classes. Following [2], these dependencies are calculated using the magnitude of the mutual information between each pair of object classes, while the sign is positive if $p(b_i = 1, b_j = 1 > p(b_i = 1)p(b_j = 1)$. This scheme is used in [2]. Tree A shows relationships mainly for mountain and rural highways scenes, while tree B shows relationships associated to street scenes. As an example of variable relationships between objects, we observe the correlation between the objects road and sign. Both objects are connected in both trees A and B, however, in tree B the dependence of both variables (0.71) is considerably higher than in tree A (0.01). This reflects the fact that these two objects relationship are more important in streets scenes than in other cases.

Figure 5 shows in a grid the dependences between objects for the single tree model and for two trees of a mixture of six trees. Considering space restrictions, we only show the values related to the eleven most frequent objects. For example, we can evaluate an image of rural highway scene. If we inspect the tree of the single tree model 5(a), we see that the objects tree and road are weakly correlated. On contrast, in tree A of the conditional mixture tree model 5(b), both objects appear as strongly correlated.



(a) Image **A**:S. Tree (b) Image **A**:MixTree-6 (c) Image **B**:S. Tree (d) Image **B**:MixTree-6

Figure 3: Some detections considering a single tree model [2] and a conditional mixture of trees. Conditional mixture of trees usually provides better detections than a single tree model.

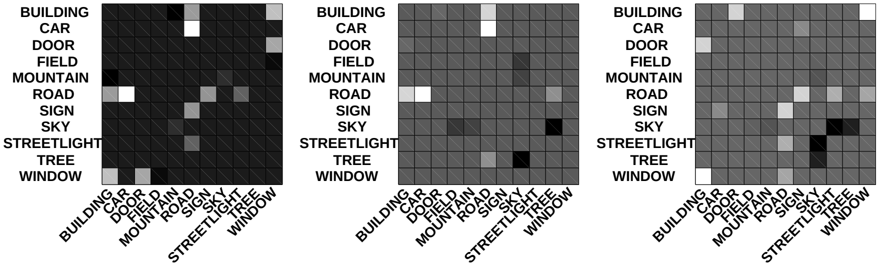


(a) Tree **A**: related to mountain scenes. (b) Tree **B**: related to street scenes.

Figure 4: Examples of component trees for a mixture of six trees in the OUTDOOR dataset. Positive and negative correlations are indicated respectively with blue and red lines.

5 Conclusions

In this work, we propose an adaptive context model that uses a conditional mixture of trees to overcome relevant limitations of a fixed tree context model. Our experiments using standard



(a) Single Tree (b) Tree **A** of Mixture of 6 Trees (c) Tree **B** of Mixture of 6 Trees

Figure 5: Dependence of the top 11 more frequent objects. Figure 5(a) shows the relationships for a single tree. Figures 5(b) and 5(c) show two relationships in a mixture of 6 trees.

object datasets indicate that the proposed model improves object recognition performance with respect to a single tree model, as it considers underlying scene information that influences object-to-object relationships. As future work, we plan to enhance our model using more powerful features for the gating function. Finally, we also plan to include adaptive policies to control the execution of object classifiers, similar to the method proposed in [6].

References

[1] 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.

[2] M. Choi, J. Lim, A. Torralba, and A. Willsky. Exploiting hierarchical context on a large database of object categories. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 129–136, 2010.

[3] J. Davis and M.Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the International Conference on Machine Learning*, pages 233–240. ACM Press, 2006.

[4] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, 39:1–38, 1977.

[5] C. Desai, D. Ramanan, and C. Fowlkes. Discriminative models for multi-class object layout. *International Journal of Computer Vision (IJCV)*, 95:1, 2011.

[6] P. Espinace, T. Kollar, A. Soto, and N. Roy. Indoor scene recognition through object detection using adaptive objects search. In *Procedings of European Conference on Computer Vision (ECCV), Workshop on Robotics for Cognitive Tasks*, 2010.

[7] L. Fei-Fei. A bayesian hierarchical model for learning natural scene categories. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 524–531, 2005.

[8] P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[9] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 264–271, 2003.

[10] C. Galleguillos and S. Belongie. Context based object categorization: A critical survey. *Computer Vision and Image Understanding*, 114(6):712 – 722, 2010.

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

[12] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision (IJCV)*, 42:145–175, May 2001. ISSN 0920-5691.

[13] J. Pearl. Reverend Bayes on inference engines: A distributed hierarchical approach. In *Proceedings of the American Association of Artificial Intelligence National Conference on AI*, pages 133–136, Pittsburgh, PA, 1982.

[14] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie. Objects in context. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 1–8, 2007.

[15] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *International Journal of Computer Vision (IJCV)*, 81:2 – 23, 2007.

[16] P. Sinha and A. Torralba. Detecting faces in impoverished images. *Journal of Vision*, 2 (7):601, 2002.

[17] A. Torralba. Contextual priming for object detection. *Proceedings of International Journal of Computer Vision (IJCV)*, 53:169–191, 2003.

[18] A. Torralba, K. Murphy, and W. Freeman. Contextual models for object detection using boosted random fields. In *Advances in Neural Information Processing Systems(NIPS)*, pages 1401–1408, 2005.

[19] I. Ulrich and I. Nourbakhsh. Appearance-based place recognition for topological localization. In *Proceedings of IEEE International Conference on Robotics and Automation (ICRA)*, volume 2, pages 1023–1029, 2000.

[20] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1:511–518, 2001.

[21] L. Xu, M.Jordan, and G. Hinton. An alternative model for mixtures of experts. In *Advances in Neural Information Processing Systems (NIPS)*, pages 633–640. MIT Press, 1994.