Mixing Hierarchical Contexts for Object Recognition

Billy Peralta and Alvaro Soto

Pontificia Universidad Católica de Chile
bmperalt@uc.cl asoto@ing.puc.cl

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-ocurrences and spatial contraints has been successfully applied to reduce the inherent uncertainty of the visual world. Recently, Choi et al. [5] propose the use of a tree-structured graphical model to capture contextual relations among objects. Under this model there is only one possible fixed contextual relation among subsets of objects. In this work we extent Choi et al. approach by using a mixture model to consider the case that contextual relations among objects depend on scene type. Our experiments highlight the advantages of our proposal, showing that the adaptive specialization of contextual relations improves object recognition and object detection performances.

1 Introduction

Humans have the remarkable ability to quickly recognize objects in images even though the objects might have different sizes, rotations, and poses. This ability is still a main challenge for artificial vision systems. In particular, several works in robotics [2],[8] and [15] have shown the relevance of using visual object recognition modules to interact with the world, but there is still a need for more robust and flexible object recognition techniques.

In the literature of object recognition, there are significant milestones, such as [16], which proposes a new feature that is invariant to rotation and scaling, and [24], which proposes a real time object detector. Machine learning techniques have also been successfully used in the computer vision area, such as [13], [11] and [12]. In general, the most recent progress in the area of object recognition has been closely related to the sinergistic combination of tool from computer vision and machine learning.

Currently, object detectors are mainly trained using images from single object categories, as we can see in datasets such as Pascal [10], Caltech [13] and MIT-CSAIL [1]. As a consequence, typical approaches do not consider contextual relations among objects and scenes. These types of relations are highly relevant to reduce some of the ambiguities of the visual world. For example, in Figure

(a) Grass and building relations

(b) Road and tree relations

Fig. 1. In real images, particular objects usually tend to co-occur and have positional relationships among them.

1, we can see cases of outdoor scenes where particular objects usually tend to co-occur and have positional relationships among them.

An interesting option to improve the performance of single object detectors is to include in the models contextual relations among objects [14], such as co-ocurrences, or mutual spatial or scale constrains. In [19] spatial context is modelled using a variant of a boosting algorithm. In a seminal work, [22] uses contextual relations based on the statistics of low-level features in terms of the global scene and the objects in it. Recently, Hoi et al. [5] model inter-object relations using a tree-structured Bayesian network.

The works mentioned above assume that there is only one possible fixed contextual relation among objects. We believe that a richer representation should include the typical variations that occur among object relationships under different scenarios. For example, if we analyze the relation between a person and a dog, this is not fixed but it changes according to different scenarios. In the case of a park scene, person and dog objects co-occurs frequently, but in an office scene they hardly co-occur.

Our idea is to learn conditional relationships among objects according to each latent scene. In particular, we present an extension to the work in [5], where we use mixture models to capture a richer set of adaptive relations among objects and scenes. This paper is organized as follows. Section 2 describes previous work. Section 3 introduces the method proposed in this paper. Section 4 presents and discusses our main results. Finally, Section 5 shows our main conclusions and some future avenues of research.

2 Background

2.1 Related work

In the case of object recognition considering context, we can divide the related work in two levels: global and local context [14]. In the case of global context,

most works exploit scene configuration as an complementary information source. This configuration is represented using a statistics of the complete image. Ulrich and Nourbakhsh introduce color histograms as the representation of an image and a k-nearest neighbors scheme for classification [23]. They apply their method to topological localization of an indoor mobile robot, but retraining is needed for each specific indoor environment. Torralba proposes an image representation based on global features that represent dimensions in a space that they call Gist feature [22]. To construct it, an image is passed through a Gabor filter bank using 4 scales and 8 orientations, then the image is divided into a 4x4 non-overlapping grid, and finally the output energy of each filter is averaged within each grid cell. Chang et al. use low-level global features that are used to estimat a belief or confidence function over the available scene labels [3]. They build one classifier for each available scene category.

In the case of local context, contextual information is derived from specific blocks or local areas around object positions. Sinha and Torralba [20] improve face detection using local contextual regions. Torralba et al. [21] 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. [19] combine layouts of textures and context to recognize objects. He uses a CRF to learn model of objects and a boosting algorithm to combine the texture information and the object model. Galleguillos et al. [14] use a CRF to maximize the true labeling of objects inside one scene constrained by co-ocurrence and location relations. Hoi et al. [5] model inter-object relations using a tree-structured Bayesian network. By using a tree, they avoid the combinatorial explosion in the number of possible relations. Another advantage of this tree-representation is the efficiency for making inference over data. Aditionally, Rabinovich et al. [18] show that textual data from Web is an useful source to estimate co-ocurrence between objects.

3 Our model

We learn scene types using a classical clustering algorithm: *K-Means*. The use of an unsupervised method to find scene types is due to the absence of scenes labels. We execute the clustering on the space of the global feature Gist G_G [5]. We use the clusters provided by K-Means to build a Gaussian Mixture Model with variances and weigths of components equal to one. Accordingly, we modify the graphical model proposed by Choi et al. [5] by adding a mixture element, as shown in Figure 2. We describe next each of the elements of this model.

3.1 Specification of the model

In what follows, the subindex i refers to object class i and the subindex k refers to the ranking of the detection. This ranking comes from the order of the scores of the multiple detections of the object detector of Felzenszwalb [12].

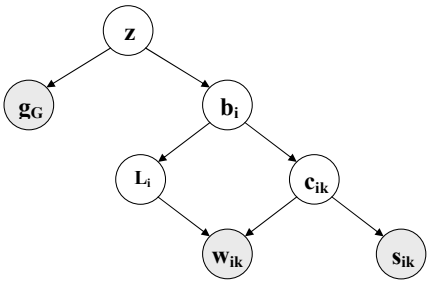

Fig. 2. Proposed graphical model.

Location of window (W) : The location of a detection window $w_{ik} = (L_y, \log L_z)$, where L_y is the vertical location of the window and L_z is the scale of the window. We model the location distribution as a Gaussian. It is conditional on c_{ik}, a binary variable that indicates if a detection is correct, and also the expected mean of the location of objects L_i; where we consider the true appearances, L_i, and the false appearances, $\bar{L}_i$. The covariance for the true positive cases is given by Λ_i and for the false positives by $\bar{\Lambda}_i$. We define **(i)** $p(w_{ik}|c_{ik} = 1, L_i) = Normal(w_{ik}; L_i, \Lambda_i)$ and **(ii)** $p(w_{ik}|c_{ik} = 0, L_i) = Normal(w_{ik}; \bar{L}_i, \bar{\Lambda}_i)$. In case that there is not object, we assume a Uniform distribution.

Score (S) : The score of classifier $s_{ik} \in \Re$. We model the score as a distribution that depends on whether the window is a correct detection or a false positive. Using Bayes rule, we have $p(s_{ik}|c_{ik}) = p(c_{ik}|s_{ik})p(s_{ik})|p(c_{ik})$. The logistic regressors are used to model $p(c_{ik}|s_{ik})$. This allow us to increase robustness, since there are few samples with $c_{ik} = 1$.

In the case that there is not a positive case in the partition, we add an artificial detection with score slightly greater than the maximum value of the current scores. This is because the logistic regresor requires at least two classes.

Correct detection (C) : The correct detections $c_{ik} \in 0, 1$, where 0 means a false positive and 1, a true positive. We model the correct detection as depending on the presence of objects b: **(i)** $p(c_{ik} = 1|b_i = 1)$ equal to frequencies from training set when object i appears; and **(ii)** $p(c_{ik} = 1|b_i = 0)$ equal to zero.

Parameter of location (L) : We model the distribution of locations as depending of 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 all instances of object i, and it is composed by $(L_y, \log L_z)$. Its components are L_y, the vertical position in the image, and L_z, the scale of detection. The use of logarithm has been shown suitable in previous work [5].

In the case of conditional components, we use Gaussians for the location of object i, using the following expressions for the respectives cases: **(i)** if $b_i = 1$

and $b_{pa(i)} = 1$, we use the location of object $pa(i)$; **(ii)** if $b_i = 1$ and $b_{pa(i)} = 0$, we use the location of object i; and finally **(iii)** if $b_i = 0$, we use the location of all objects in all images. In case of objects that do not appear in a partition, again we use a Uniform distribution.

Presence of object (B) : The presence of object $b_i \in 0, 1$. We model object presence as a function of the learned tree model and the partition z. We have: $p(b|z) = p(b_{root}|z) \prod_i p(b_i|b_{pa(i)}, z)$.

We learn the tree from data using the Chow-Liu algorithm on each partition [6]. There is a restriction for the case of objects that do not appear in a particular partition. For these objects we consider a mutual information equal to zero, and we add them as children of the last added node to the tree. In this way, we diminish the influence of these variables.

Global Gist (g_G) : We model the global Gist feature [22] as depending on partition z, so we have $p(g_G = g|z) = Normal(g; u_z, \Sigma_z)$. We consider the variance and weights equal to one. It helps to simplify the distant metric.

In the model of Choi et al. [5], g_G is used as a direct prior of the individual objects. However, it is independent of the context-hierarchical model, as we want to evaluate the goodness of the context model, we do not use this information for both techniques in order to make a fair comparison.

Partition (z) : The partition $z \in 0, |Z|$, where $|Z|$ is the number of scences. This partition represents the latent scenes for the database. In this case, we obtain its value for a particular image according to the model of clustering.

It is important to mention that due to the computational complexity, we do not apply a joint optimization over partitions and local models. Instead, we first apply a clustering of images, and afterwards we learn the local models.

3.2 Inference

The inference is straightforward because we separate each tree in its own partition. We make an inference using message passing algorithms for each tree ($p(b, c, L/g, W, s, z)$) [17]. Then we obtain the final score by combining the scores of each component with its respective parameters.

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

Similarly to [5], we use the following iterative procedure to detect objects: first we make an inference without consider 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 presence of each object ($\hat{b}, \hat{c} \propto p(b, c|s, g, \hat{L}, W)$).

4 Experiments

In our experiments, we use the dataset created by Choi et al. [4]. This dataset has 111 classes, 4.367 training images, and 4.317 test images. In general, the detection of the objects in these images is highly challenging, including a variety of poses, scales, rotations, and scene types. As a baseline technique, we use the object detector proposed by Felzenszwalb et al. [12], which is based on the mixture of multiscale deformable part models and a variant of SVM. In average, this detector provides approximately 500 detections for each image. We use as a performance metric of our model the average precision-recall metric (APR) [7]. This metric corresponds to the area under the precision-recall curve.

In order to test the method, we define the detection and recognition tasks. The detection for an object Γ is defined as the procedure where we determine if the object Γ appears or not in the entire image; in this case, we only use the detection for object Γ with top likelihood. In the recognition task for object Γ, we check if each detection of object Γ inside the image is a true positive.

Table 1 shows the results using the baseline technique [12] (*BaseLine*), the method proposed by Choi et al. [5] based on a single tree (*Tree*), and our proposed method based on a mixture model (MT).

Table 1. APR for Choi-database

	Baseline	**Tree**	**MT**					
#Trees	-	1	2	3	4	5	6	7
Recognition	6.82	7.08	7.28	7.39	**7.59**	7.30	7.47	7.53
Detection	13.31	17.74	18.08	18.16	**18.28**	18.14	18.11	17.94

In table 1, we show the average of the APR for the 111 object categories. We note that our method improves performance for both tasks, recognition and detection. The best number of trees in this dataset is 4. When we compare this result to [5], we find a favourable difference of 0.51 for APR recognition and 0.54 for APR detection.

Figure 3 shows the six most confident detections provided by the proposed method for some test images [1]. As an example, we can see in the first figure of the second row how our method correctly recognizes three cars, one person, one tree, and the sky.

5 Conclusions

In this work, we present an extension of the model of Choi et al. based on a mixture of trees that combines conditional contextual relations among objects. Our experiments using a standard dataset indicate that the proposed model

[1] We suggest to see these images in color

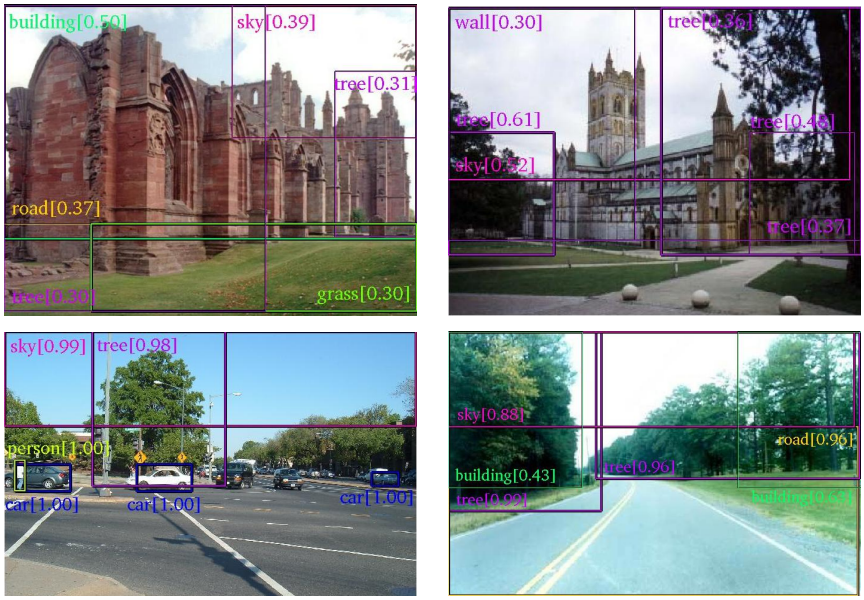

Fig. 3. Sample detections for the proposed method.

improves the results of a state-of-art technique in terms of object detection and recognition according to the APR metric. These improvements provide evidence that an adaptive modelling of the interactions among object help recognition.

As future work, we plan to enhance our model using a more informative clustering process, such as including explicit latent models or a discriminative clustering technique. We also plan to include adaptive policies to control the execution of object classifiers, such as the method proposed in [9].

Acknowledgements This work was partially funded by FONDECYT grant 1095140.

References

1. K. Murphy A. Torralba and W.T. Freeman. The mit-csail database of objects and scenes. http://web.mit.edu/torralba/www/database.html, 2010.
2. M. S. Bartlett, G. Littlewort, I. Fasel, J. Chenu, H. Ishiguro, and J. R. Movellan. Towards social robots: Automatic evaluation of human-robot interaction by face detection and expression classification. In *In Advances in*, page 2003. MIT Press, 2003.
3. Edward Chang, Kingshy Goh, Gerard Sychay, and Gang 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. M. Choi. Large database of object categories. http://web.mit.edu/ myungjin/www/HContext.html, 2010.

5. Myung Jin Choi, Joseph J. Lim, Antonio Torralba, and Alan S. Willsky. Exploiting hierarchical context on a large database of object categories. In *IEEE Conference on Computer VIsion and Pattern Recognition (CVPR)*, 2010.
6. C. I. Chow and C. N. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 14:462–467, 1968.
7. Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *In ICML 06: Proceedings of the 23rd international conference on Machine learning*, pages 233–240. ACM Press, 2006.
8. Staffan Ekvall, Danica Kragic, and Patric Jensfelt. Object detection and mapping for service robot tasks. *Robotica*, 25(2):175–187, 2007.
9. P. Espinace, T. Kollar, A. Soto, and N. Roy. Indoor scene recognition through object detection. In *Proc. of IEEE Int. Conf. on Robotics and Automation (ICRA)*, 2010.
10. M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.
11. Li Fei-Fei. A bayesian hierarchical model for learning natural scene categories. In *In CVPR*, volume 2, pages 524–531, 2005.
12. Pedro F. Felzenszwalb, David A. McAllester, and Deva Ramanan. A discriminatively trained, multiscale, deformable part model. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2008.
13. R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *In CVPR*, pages 264–271, 2003.
14. Carolina Galleguillos and Serge Belongie. Context based object categorization: A critical survey. *Computer Vision and Image Understanding*, 114(6):712 – 722, 2010. Special Issue on Multi-Camera and Multi-Modal Sensor Fusion.
15. Kai Huebner, Mårten Björkman, Babak Rasolzadeh, Martina Schmidt, and Danica Kragic. Integration of visual and shape attributes for object action complexes. In *ICVS'08: Proceedings of the 6th international conference on Computer vision systems*, pages 13–22, Berlin, Heidelberg, 2008. Springer-Verlag.
16. David Lowe. Object recognition from local scale-invariant features. pages 1150–1157, 1999.
17. 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.
18. Andrew Rabinovich, Andrea Vedaldi, Carolina Galleguillos, Eric Wiewiora, and Serge Belongie. Objects in context. In *ICCV'07*, pages 1–8, 2007.
19. Jamie Shotton, John Winn, Carsten Rother, and Antonio Criminisi. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context, 2007.
20. Pawan Sinha and Antonio Torralba. Detecting faces in impoverished images. *Journal of Vision*, 2(7):601, 2002.
21. A. Torralba, K. P. Murphy, and W. T. Freeman. Contextual models for object detection using boosted random fields. In *Advances in Neural Information Processing Systems 17 (NIPS)*, pages 1401–1408, 2005.
22. Antonio Torralba. Contextual priming for object detection. *IJCV*, 53:2003, 2003.
23. I. Ulrich and I. Nourbakhsh. Appearance-based place recognition for topological localization. In *Robotics and Automation, 2000. Proceedings. ICRA '00. IEEE International Conference on*, volume 2, pages 1023 –1029 vol.2, 2000.
24. Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. pages 511–518, 2001.