IMPROVED SEMANTIC REGION LABELING BASED ON SCENE CONTEXT

Matthew R. Boutell[1], Jiebo Luo[2], and Christopher M. Brown[1]

[1]Department of Computer Science
University of Rochester
{boutell,brown}@cs.rochester.edu

[2]Research and Development Labs
Eastman Kodak Company
jiebo.luo@ kodak.com

ABSTRACT

Semantic region labeling in outdoor scenes, e.g., identifying sky, grass, foliage, water, and snow, facilitates content-based image retrieval, organization, and enhancement. A major limitation of current object detectors is the significant number of misclassifications due to the similarities in color and texture characteristics of various object types and lack of context information. Building on previous work of spatial context-aware object detection, we have developed a further improved system by modeling and enforcing spatial context constraints specific to individual scene type. In particular, the scene context, in the form of factor graphs, is obtained by learning and subsequently used via MAP *estimation to reduce misclassification by constraining the object detection beliefs to conform to the spatial context models. Experimental results show that the richer spatial context models improve the accuracy of object detection over the individual object detectors and the general outdoor scene model.*

1. INTRODUCTION

Object detection can facilitate a number of image understanding applications, such as content-based image retrieval (CBIR). For example, Naphade and Huang use semantic features, such as the presence of sky, rock, snow, and water, to index and retrieve video [1]. Intuitively, these semantic features help bridge the so-called semantic gap between pixels and the desired understanding of the image, causing many researchers to look beyond traditional low-level features, such as color, texture, and edges.

Color and texture have been the central features of existing work on natural object detection. For example, Saber *et al.* [2] used color classification to detect sky by assuming a 2-D Gaussian probability density function. More recently, location (knowing correct image orientation) has been used to boost accuracy: Smith and Li [3] assumed that a blue, extended patch at the top of an image is likely to represent clear sky, while Vailaya and Jain [4] presented an exemplar-based approach that uses a combination of color, texture, and location features to classify sub-blocks (16 × 16 pixels) in an *outdoor* scene.

However, even with all this work on object and material detection, detectors are still not perfect. How can one improve them further? An idea that we can take from humans is that they use *context*. One type of context is *spatial* context. Whereas individual detectors only use isolated patches of pixels (which is difficult even for humans, as shown in [5]), spatial context refers to the spatial relationships between objects in the scene, and is often useful to reduce ambiguity among conflicting detectors and to remove improbable spatial configurations of objects. For example, while snow and cloudy sky can be confusing without context, sky tends to occur *above* foliage, while snow occurs *below* foliage. Singhal, *et al*. [5] successfully used spatial context models to improve material detectors for natural scenes. They first combined the output of the individual object detectors to produce a belief vector for the objects potentially present in an image. They then imposed spatial context constraints, in the form of learned probability density functions (pdfs), for spatial relations. These pdfs were learned from a large set of general outdoor scenes.

However, another type of context that has been little exploited is *scene* context. Knowing what type of scene (e.g., beach, field) one is viewing lends specific evidence both toward the type of objects to expect and the spatial configuration in which they occur. Configurations of objects can vary greatly from scene to scene. For example, in beach scenes, sand regions tend to occur in the foreground of the image, leading to the relation, *sand below water*, while in open water scenes, when sand-like regions (such as land on the horizon) occur, they occur in the background, yielding *sand above water*. Hereafter, we refer to this scene-type specific spatial context as *scene context* for conciseness and to differentiate it from the general *spatial context* in [5]. An overview of our system, similar to that in [5], is shown in Figure 1, with the main difference being that the spatial context models are now scene-specific.

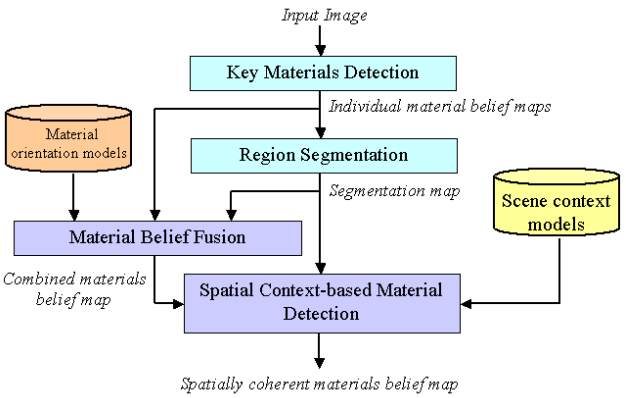

Fig. 1. Architecture of the holistic object-detection system.

Some psychophysical research has shown that humans can discern the gist of a scene very quickly (before they recognize objects) [6], so it is plausible that humans have access to scene information as well before they completely determine the identity of every object in the scene. With image understanding, semantic scene classification [8][9] may be performed prior to semantic region labeling. We envision two use scenarios. In the first, automatic algorithms are used to determine the scene classification, although scene classification errors could affect region labeling

0-7803-9332-5/05/$20.00 ©2005 IEEE

performance. In the second, the scene labels are pre-assigned. This is particularly appropriate if the photographer has already organized her images into albums and labeled them; in any case, labeling an album or each image is much easier than labeling each *region* in each image.

In this paper, we extend the system introduced in [5] in three ways. First, we augment the spatial context module shown in Figure 1 with *scene context*. Second, we double the number of semantic materials detected to *ten*. Third, we use a single, integrated model, a *factor graph*, as a means to propagate object beliefs between image regions.

2. SCENE-CONTEXT MODEL

We now discuss the details of our scene context model and the spatial relations it encompasses, and compare it with other models using general, non-scene-specific spatial context only.

2.1. Probabilistic model

There is existing work on using high-level scene models for spatial context-based object detection. Batlle *et al*. [13] provide a comprehensive review of most of the early work related to building scene models for specific image types. They describe techniques where spatial models (e.g., rules) can be constructed for scene types, such as a house scene, a road scene, and an urban scene. In each of these scene types, there is a strong expectation regarding the occurrence and location of various object types in the image. Lipson *et al*. [14] present a spatial context modeling approach, called configuration-based scene modeling, for content-based indexing and retrieval applications. They model the qualitative and photometric relationships between various objects in a scene in a spatial sense and use these relationships to extract other scenes with semantically similar content. The scene models are extremely specific to the layout of a scene, e.g., ocean on top of sand is different from ocean beside sand.

One principled method of using spatial context is within the framework of a probabilistic graphical model, such as a Bayesian network (BN) or a Markov Random Field (MRF) [1][5][6]. These systems have the following advantages (vs. heuristic-based ones, such as [7]). First, they are built on a strong theoretical framework. Second, they are highly modular; the detectors are decoupled from each other and from the context model, which is an important concern in a large system with many components, and each expected to improve over time.

We use the factor graph shown in Figure 2 as a context model, because it is more general than BN or MRF. The graph has a single *scene* variable, S, which can take on a scene type, e.g. *beach* or *field*, and a *region* variable, R_i, for each region in the scene, which takes on the values of the region labels, e.g., *sky, water,* or *foliage*. We learn pairwise spatial relationships for each scene type. These spatial relationships are encoded in the factors, F_{ij} connecting the scene node with each pair of region nodes. The detector factors, D_i, encode the detector evidence, and provide the likelihood of each region label, given the belief with which each detector fires; these factors are set at run-time. We could have detector variables for each detector for each region, but these can be absorbed in the detector factor. The structure of the scene is also determined at run-time because the number of regions, n, is unknown *a priori*. Finally, the prior factor, P_i, allows one to specify the prior probability in each scene type, although we do not make use of it in this study (i.e., all scene types are assumed equally probable).

Our factor graph encodes two independence assumptions. First, the scene is independent of the detector output, given the true label of each region. Second, a detector's output on a region depends only on the object present in that region and not on other objects or the class of the scene. Each detector's characteristics $P(D_i|R_i)$ can be learned by counting detection frequencies on a training set of regions or specified using domain knowledge.

However, there is one assumption that we cannot make. At this coarse segmentation, even distant regions (in the underlying image) may be strongly correlated, e.g., sky and pavement in urban scenes. Thus, we cannot factorize the scene structure (as could be done in low-level vision problems) and instead assume a fully-connected, pairwise scene structure with ${}_nC_2$ relation factors. However, for the types of materials presently of interest, n is generally small ($n<7$).

We can find the most likely labels for each region by setting the values of the detector factors, fixing the value of the scene variable with the known scene and using loopy belief propagation for inference. After convergence, each region node will have a vectors of beliefs (one entry per label); taking the argmax yields the most likely label, given the known scene and the detector output.

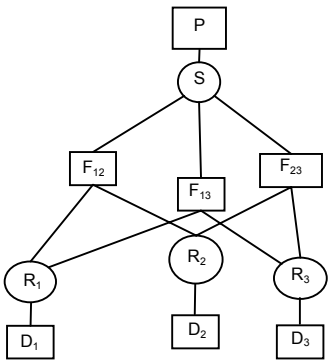

Fig. 2. Factor graph encoding scene- and spatial-context, shown with 3 regions. P = prior factor, S= scene variable, F_{ij} = spatial factor for regions *i* and *j*, R_i = region variable *1<=i<=n*, D_i = detector factor for region $_i$. See text for discussion.

2.2. Spatial relations

Pairwise spatial relations in our model are encoded as probability density functions of the two regions and the scene. In [5], the seven spatial relations *above, far above, beside, enclosing, enclosed, below*, and *far below* were shown to be effective for spatial context-aware material detection within outdoor scenes. A threshold on the distance between the nearest pixels of two regions is used to discriminate between *above* and *far above* (and *below* and *far below*).

In this study, we adopt the same spatial relations and same ways of computing them. One is by checking the bounding boxes of the regions and the other is by using a computationally efficient version of the "weighted walk-through"" approach [5]. The bounding box method is easy to implement, but may encounter difficulties when the bounding boxes of the regions overlap. The lookup table method is robust to the size and location of regions, but is computationally more complex than the bounding box method. We use a hybrid scheme to determine the spatial relationship of two regions in an image. First, we check the bounding boxes of the regions. If the

bounding boxes are not overlapping, we can use their coordinates to quickly derive the spatial relationship.

The spatial context models are built by learning probability density functions corresponding to the spatial relationships described above, as opposed to handcrafted rules. A simple frequency counting approach suffices to generate all the discrete pdfs. Example pdfs for the relationship *above* are shown in Tables 1 and 2 for *beach* and *open-water*. For example, the sixth row in Table 1 shows that any region above a water region is most likely to be sky or cloud; snow and pavement do not occur at all. Note that the relationship between sand and water is much different for open-water scenes and beach scenes; for beach scenes, water almost always occurs above sand, while for open-water scenes, sand, when it occurs, occurs above water (as discussed in the Introduction).

Table 1: *Beach-specific* pdf for B *above* A for ten materials for which we have detectors: Blue Sky, Clouds, Grass, Foliage, Snow, Water, Sand, Pavement, Rocks, and Manmade structures.

A\B	Sky	Clo	Gra	Fol	Sno	Wat	San	Pav	Roc	Man
Sky	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00
Clo	0.04	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00
Gra	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Fol	0.02	0.03	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00
Sno	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Wat	0.12	0.14	0.00	0.03	0.00	0.00	0.01	0.00	0.01	0.02
San	0.03	0.04	0.00	0.04	0.00	0.17	0.01	0.00	0.03	0.03
Pav	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Roc	0.01	0.02	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.00
Man	0.03	0.02	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00

Table 2: *Open-water-specific* pdf for B *above* A.

A\B	Sky	Clo	Gra	Fol	Sno	Wat	San	Pav	Roc	Man
Sky	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Clo	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Gra	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Fol	0.03	0.08	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Sno	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Wat	0.01	0.24	0.00	0.07	0.00	0.00	0.01	0.00	0.01	0.20
San	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Pav	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Roc	0.00	0.02	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00
Man	0.03	0.19	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.01

2.3. Comparison with other spatial models

In previous work [9], we used a full spatial configuration model, equivalent to replacing the pairwise spatial factor with a single factor connecting the scene with all n regions. The advantage was that we obtained an exact representation of the probability density over scenes and regions and inference was guaranteed to converge. Graph-based smoothing was used to populate the high-dimensional pdf. However, a different pdf was needed for each spatial arrangement and each number of regions, which can preclude this model's use in real applications. Pairwise approximations of spatial relationships can alleviate the problem.

Singhal, *et al*. [5] approximated spatial constraints using a series of non-loopy Bayesian networks, solved iteratively. Our single factor graph eliminates the need for a *series* of models. While loopy belief propagation is not guaranteed to converge, it has been shown to have nice properties [10]. While our topology is different, it is still symmetric, as was in [5].

Markov Random Fields also have been used to incorporate spatial context in low-level vision applications (e.g., [10]). For a pairwise spatial-context only model, we could take a similar approach; in fact, removing the scene node from the factor graph yields a graph equivalent to a typical two-level MRF used in [10] but with a fully-connected scene structure. However, without a scene node, we would be required to have an MRF for each scene type. The factor graph provides a unified model for both *scene* context and *spatial* context.

3. EXPERIMENTAL RESULTS

We have a database composed of 865 consumer and stock photo images in six classes: *Beach, Field, Mountain, Open-water, Urban-street,* and *Suburban*. Note that this database is not the same as the one used in [5] because it only contains these scene types.

Each image in the database is automatically segmented [11], and the semantically-critical regions are manually labeled with their true materials (i.e., ground truth). The ground truth labels correspond to those ten materials for which we have detectors, listed in the caption to Table 1. Other regions are left unlabeled.

Our baseline detectors are based on color and texture features, similar to the common approach used in [1-4]. First, color (LUV) and texture (6 high-frequency coefficients from a 2-level wavelet transform) features are computed on the input image. The features are fed to trained neural-networks, which produce a probability or belief value for each pixel in the image according to the color and texture characteristics. The collection of pixel belief values forms a pixel belief map. After pixel classification, spatially contiguous regions are obtained from the raw pixel belief map after thresholding the belief values. The belief value of each region is the average belief values of all pixels in the region.

While we have actual detectors, we are also interested in determining the usefulness of the context models on a wider range of detector performance. To simulate different faulty detectors, we randomly perturbed the ground truth to create simulated detector responses. We set the detection rates of individual material detectors on each true material (both *true positive rates*, e.g., how often the grass detector fires on grass regions, and *false positive rates*, e.g., how often the grass detector fires on water regions) by counting performance of corresponding actual detectors on a validation set (or estimating them in the case of detectors to be developed in the future). When they fire, they are assigned a belief that is distributed normally with mean μ. The parameter μ can be set differently for true and false positive detections; varying the ratio between the two is a convenient way to simulate detectors with different operating characteristics.

Table 3 shows the comparison between material detection accuracy in two cases (average 67% and 75% baselines); we compare the proposed scene context model (MAP), with the baseline - a context-free model (MLE), and the previous general spatial context model (MAPGen). We use cross validation, learning the pdfs (e.g. Tables 1 and 2) from part of the database while testing on the remainder. Note that we *simulate* the general spatial context

model by fixing the scene variable so that each scene is equally likely. (This has the same effect as replacing the scene-specific pdfs with a general pdf obtained by averaging the scene-specific pdfs.) Therefore, it is not identical to the one in [5], which was derived from more than the six scene types in this study and encoded using a Bayes network. We intend to perform a more rigorous comparison in the near future.

Table 3: Improvement due to scene-context model (MAP) vs. spatial-only context (MAPGen) and no context (MLE).

Class	Scene-context (MAP)	Spatial-only context (MAPGen)	Context-free (MLE)
Beach	78.7	53.2	75.6
Field	86.0	84.7	76.2
Mountain	75.1	45.3	74.8
Open-water	87.9	86.7	62.6
Street	81.2	78.8	76.5
Suburban	79.6	73.6	77.3
All	80.5	66.8	75.3

Figure 3 shows an example where scene context corrects a mislabeling due to misdetection and due to a spatial-only model. The ground truth consists of clouds, sky, and grass. In the MLE case, the snow detector fires, and the cloud is mislabeled as snow (a common error due to the similar characteristics of clouds and snow), whereas the field-specific spatial model corrects it, as snow is unlikely in field scenes. For the grass region, both the grass and foliage detectors fired, giving approximately equal belief in each label. The general spatial model changed the labeling to foliage, as foliage occurs below sky more often in the data set as a whole (particularly in suburban and mountain images). However, grass more commonly occurs under sky in Field scenes, so the field-specific spatial model retains the original belief in grass.

4. CONCLUSIONS AND FUTURE WORK

We have demonstrated a context-based approach to improved object detection. In particular, the scene type-specific context, in the form of factor graphs, is obtained by learning and subsequently used via MAP estimation to reduce misclassification by constraining the object detection beliefs to conform to the spatial context models. Experimental results show that the richer spatial context models improve the accuracy of object detection over the individual object detectors and the general outdoor scene model.

One obvious area of improvement in the future is in the individual natural object detectors. Our current sky detector performs well individually with mid-90% accuracy [12], even without orientation. However, the remaining detectors have accuracies of in the range of 50 to 85% and can be substantially improved. However, we are keenly aware of the fact that, with improved individual detectors, the benefit of spatial context diminishes; with *near perfect* detectors, at which point there is no need for improvement, the use of spatial context is expected to *hurt* the performance.

5. REFERENCES

[1] M. Naphade and T. S. Huang, "A Factor Graph Framework for Semantic Video Indexing," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 12, no. 1, January, 2002.

[2] E. Saber, A.M. Tekalp, R. Eschbach, and K. Knox, "Automatic Image Annotation using Adaptive Color Classification," *CVGIP: Graphical Models and Img. Proc.*, vol. 58, pp. 115–126, 1996.

[3] J. R. Smith and C.-S. Li, "Decoding Image Semantics using Composite Region Templates," in *Proc. IEEE Int. Workshop on Content-based Access of Image and Video Database*, 1998.

[4] A. Vailaya and A. Jain, "Detecting Sky and Vegetation in Outdoor Images," *Proc. SPIE*, vol. 3972, 2000.

[5] A. Singhal, J. Luo, and W. Zhu, "Probabilistic Spatial Context Models for Scene Content Understanding," in *Proc. IEEE Int. Conf. On Computer Vision and Pattern Recognition*, 2003.

[6] K. Murphy, A. Torralba, and W. T. Freeman "Using the Forest to See the Trees: A Graphical Model Relating Features, Objects, and Scenes. In *Proc. Neural Info. Proc. Sys.,* 2003.

[7] P. Mulhem, W. Leow, and Y. Lee, "Fuzzy Conceptual Graphs for Matching Images of Natural Scenes," in *Proc. IJCAI*, 2001.

[8] A. Vailaya, M. Figueiredo, A. Jain, and H.-J. Zhang, "Content-Based Hierarchical Classification of Vacation Images," in *Proc. IEEE Multimedia Systems*, 1999.

[9] M. Boutell, J. Luo, and C. Brown, "Learning Spatial Configuration Models using Modified Dirichlet Priors," in *Workshop on Statistical Relational Learning (in conjunction with ICML)*, 2004.

[10] W.T. Freeman and E. Pasztor, "Learning Low-Level Vision," in *Proc. IEEE Int. Conf. on Computer Vision,* 1999.

[11] D. Comaniciu and P. Meer, "Mean Shift: A Robust Approach Toward Feature Space Analysis," *IEEE Trans. Pattern Anal. Mach. Intell.* Vol 24, no. 5: 603-619, 2002.

[12] J. Luo and S. P. Etz, "A Physical Model-based Approach to Detecting Sky in Photographic Images," *IEEE Transactions on Image Processing*, vol. 11, pp. 201–212, 2002.

[13] J. Batlle, A. Casals, J. Freixenet, and J. Marti, "A Review on Strategies for Recognizing Natural Objects in Colour Images of Outdoor Scenes," *Img. Vis. & Comput.*, vol. 18, pp 515–530, 2000.

[14] P. Lipson, E. Grimson, and P. Sinha, "Configuration Based Scene Classification and Image Indexing," *Proc. IEEE Int. Conf. Comput. Vision and Pattern Recog.*, 1997.

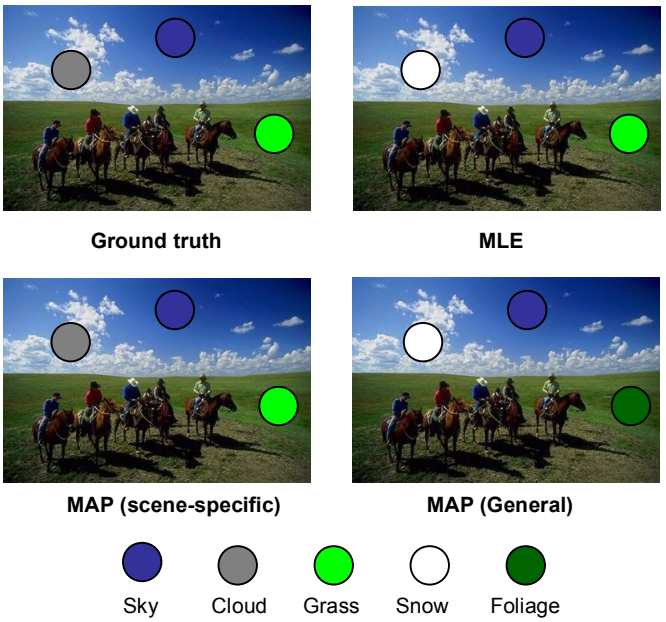

Fig. 3. A field example showing improvement due to scene-specific spatial model over both baselines.