Multiple Region Categorization for Scenery Images

Tamar Avraham, Ilya Gurvich, and Michael Lindenbaum

Computer Science Department, Technion - I.I.T., Haifa 32000, Israel
tammya@cs.technion.ac.il, ilya.gurvich@gmail.com, mic@cs.technion.ac.il

Abstract. We present two novel contributions to the problem of region classification in scenery/landscape images. The first is a model that incorporates local cues with global layout cues, following the statistical characteristics recently suggested in [1]. The observation that background regions in scenery images tend to horizontally span the image allows us to represent the contextual dependencies between background region labels with a simple graphical model, on which exact inference is possible. While background is traditionally classified using only local color and textural features, we show that using new layout cues significantly improves background region classification. Our second contribution addresses the problem of correct results being considered as errors in cases where the ground truth provides the structural class of a land region (e.g., mountain), while the classifier provides its coverage class (e.g., grass), or vice versa. We suggest an alternative labeling method that, while trained using ground truth that describes each region with one label, assigns both a structural and a coverage label for each land region in the validation set. By suggesting multiple labels, each describing a different aspect of the region, the method provides more information than that available in the ground truth.

Keywords: region annotation, multiple categorization, exact inference, scenery/landscape, boundary shape, contextual scene understanding.

1 Introduction

The incorporation of context into object detection and region labeling has recently come into the mainstream of computer vision (e.g., [2,3,4,5,6,7]). In these methods the identity of an image region depends both on its local properties and on the labels and appearance of the neighboring regions. To solve the region labeling problem generally, approximation methods (e.g., loopy belief propagation) are required. In our work we focus on context based region annotation for scenery images. It turns out that this more simple problem can be modeled by a rather simple graphical model on which exact inference is possible.

We follow [1], where statistical properties of scenery images were analyzed. It was observed that background regions in scenery images tend to horizontally span the image, making it possible to define a one-dimensional top-bottom order of the background regions. Moreover, it was observed that the label of a

background region correlates with the shape of the upper part of its boundary. It was shown that by using only those two *layout* properties, it is possible to capture the general appearance variability of scenery images: those cues enabled the generation of semantic sketches of scenes. However, [1] left an open question about whether those cues can assist in region annotation. In the first part of this work we answer this question; see Fig. 1(a). We suggest an exact inference model for annotating background regions that combines layout cues with texture and color cues. We show that this combination significantly improves classification over methods that rely only on local color and texture, as each type of cue contributes to a different dichotomy.

The second part of this paper considers a related but different problem. We observe that results counted as errors are not always wrong. Often, regions are associated with different labels which correspond to different aspects of them, while annotators usually provide only one label corresponding to one of these aspects. For instance, a mountain with trees may be classified by the ground truth as mountain and classified by our algorithm as trees, or vice versa. That is, a land region's annotation can either describe its structure (mountain, plain, valley) or the overlying land-cover (trees, grass, sand, rocks, etc.)

This relates to recent work on problems with large numbers of categories [8,9]. In order to get a more informative accuracy score, it was suggested that the cost of misclassification be associated with the relative location of the true classification and the estimated classification in the wordnet tree [10]. This solution is good for foreground objects for which the categorization can be for different semantic details (e.g., crow, bird or animal). However, we found this method unsuitable for background categories, as the structural descriptors and the coverage descriptors do not appear in close wordnet sub-trees. Another recent related work [11] suggests that data can be organized by several non-redundant clustering solutions, each providing a different facet.

To support multiple categorization of land regions, we suggest an alternative labeling method that, while trained using ground truth that describes each region with one label, assigns two types of annotation for test data; see examples in Fig. 3. The annotations in the training data allow us to generalize both the appearance of land coverage categories (using the regions in the training set that are labeled by their coverage) and to generalize the appearance of land structure categories (learning from regions labeled by their structure). Given a new test image, each of its land regions can now be classified by both characteristics.

Sec. 2 overviews background region classification cues. Sec. 3&4 discuss the region classification algorithm and its results. Sec. 5&6 discuss the multiple labeling algorithm and its results. Sec. 7 concludes.

2 Cues for Background Region Classification

2.1 Color and Texture

The most natural choice for classifying a background region is by its color and textural attributes. In [12], such attributes were used for classifying patches

Fig. 1. (a) Demonstrating the contribution of ‘layout’: given this general image layout (without the texture), we can easily decide between the two suggested annotations. (b) The image is divided into n horizontal background regions, $R = (R_1, \dots, R_n)$. Each region $R_i = (H_i, T_i, S_i, O_i)$ is described by its average height H_i , its texture&color descriptor T_i , by the curve separating it from the region above it, S_i , and by its relative location (order), $O_i = i$. (c) the HMM (Hidden Markov Model) representing the ORC model. The regions’ labels, $l_1, \dots, l_n$, are the hidden variables/states. The observed variables are $H_1, \dots, H_n$, $T_1, \dots, T_n$, and $S_2, \dots, S_n$.

of images, as a first step in a method for scene categorization. We adapt the description suggested there and describe each region by: 1. Color histograms: a histogram for each of the components of the HSV color representation (36, 32, 16 bins for H, S, and V, respectively). 2. Edge direction histograms: we apply a Canny edge detector and collect the gradient directions in the edge locations (72 bins). 3. GLCM (Gray Level Co-occurrence Matrix [13]): the region is quantized to 32 gray levels. A GLCM is computed for 4 offsets $(-1,1)$, $(0,1)$, $(1,1)$ and $(1,0)$. For each, we compute the contrast, energy, entropy, homogeneity, inverse difference moment, and the correlation, for a total of 24 additional components.

2.2 Relative Location Statistics

In [1] it was shown that types of background regions tend to have typical relative locations. For example, a sand region will usually appear below a sea region, mountains are usually higher in the image than fields, and of course the sky is usually above all. These top-bottom relations were modeled by a Markov network. Let $\{L_1, \dots, L_m\}$ be the possible background labels. The network has $m + 2$ nodes. The first m are associated with the m labels. In addition, there is a starting status denoted ‘top’ and a sink status denoted ‘bottom’. $M(L_i, L_j)$ is the probability that a region labeled L_i appears above a region labeled L_j in an image. $M(\text{‘top’}, L_i)$ and $M(L_i, \text{‘bottom’})$ are the probabilities that a region with label L_i is at the top/bottom of an image, respectively. The transition probabilities are estimated from the training image set ¹.

¹ We specify the top-bottom order by the height of the highest pixel in each region. Therefore, the background regions do not always have to horizontally span the image.

2.3 Boundary Shape Characteristics

In [1], the characteristics of a contour separating two background regions were shown to correlate with the lower region's identity. The boundary on top of a sea, grass or field region is usually smooth and horizontal, resembling a DC signal. The boundary on top of a region of trees or plants can be considered as a high frequency 1D signal. The boundary on top of a mountain region usually resembles 1D signals of rather low frequency and high amplitude.

Following this observation, it was suggested that a signal representation be adopted. For each background labeled region, the upper part of its contour is extracted and cut to chunks of 64-pixel length. Each such chunk is actually a descriptor vector of length 64. In the model described in Sec. 3 we use these descriptor vectors as cues for region annotation, and use an SVM that provides probability estimates [14]. Let S_i describe the boundary above a region indexed i with identity $l \in \{L_1, \dots, L_m\}$. S_i is cut to K_i chunks, $S_{i1}, \dots, S_{iK_i}$. $K_i = \lfloor \frac{|S_i|}{64} \rfloor$. When the region belongs to a training image, the K_i chunks, each labeled l , are members of the training set. When the region is being classified, the SVM classifier returns a probability estimate for each of the K_i chunks, $p(l = L_j | S_{ik})$, $k = 1, \dots, K_i$, $j = 1, \dots, m$. The class probability for the whole signal (boundary) is then $p(l = L_j | S_i) = \frac{1}{z_1} \prod_{k=1}^{K_i} p(l = L_j | S_{ik})$, where z_1 is a normalizing factor.

3 Background Region Classification: The ORC Algorithm

In [1], it was shown that it is possible to capture the general appearance variability of scenery images using only the cues described in sections 2.2&2.3. Those cues enabled the generation of semantic sketches of scenes. However, [1] left an open question about whether those cues can assist in region annotation. In this section we answer this question by proposing a mechanism for combining all the cues described in Sec. 2: the *ORC* (Ordered Region Classification) algorithm.

Let $R = (R_1, \dots, R_n)$ be n background regions in an image I , ordered by their top-bottom location. Each $R_i = (H_i, S_i, T_i, O_i)$ is characterized by its size, H_i^2 , its color&texture, T_i , the '1D signal' S_i describing the boundary separating it from R_{i-1} ($S_1 = \emptyset$), and by its order in the image, $O_i = i$; see Fig. 1(b).

Taking a contextual approach, the identity l_i of region R_i depends on its appearance, its location, and on the appearance and relative location of the other image regions. Therefore, the probability for l_i to be L_j is a marginalization over all joint assignments $(l_1, \dots, l_n)$ in which $l_i = L_j$:

$$p_{\text{ORC}}(i, L_j) = p(l_i = L_j | R) = \sum_{\substack{(k_1, \dots, k_{i-1}, k_{i+1}, \dots, k_n) \\ \in \{1, \dots, m\}^{n-1}}} p(l_1 = L_{k_1}, \dots, l_i = L_j, \dots, l_n = L_{k_n} | R).$$

We use the Markovian property described in Sec. 2.2, i.e., the identity of region R_i directly depends only on the identity of region R_{i-1} . Also, we ignore the

² Since we are discussing horizontal patches that usually span the image from side to side, their size is described only by their average height.

direct dependency between the color, texture, and height of the different regions inside an image, and between the appearance of separating boundaries inside an image. The probability for a joint assignment is then

$$p(l_1 = L_{k_1}, \dots, l_n = L_{k_n} | R) = p(l_1 = L_{k_1} | R_1) \prod_{i=2}^n p(l_i = L_{k_i} | l_{i-1} = L_{k_{i-1}}, R_i). \quad (1)$$

Assuming T_i , S_i , H_i , and O_i are independent, and that T_i , S_i , H_i are independent in R_{i-1} , every term in the product can be expressed as

$$p(l_i = L_{k_i} | l_{i-1} = L_{k_{i-1}}, R_i) = \frac{p_{T_i} p_{S_i} p_{H_i} p(l_i = L_{k_i} | l_{i-1} = L_{k_{i-1}}, O_i)}{p(T_i, S_i, H_i)} = \quad (2)$$

$$\frac{p(l_i = L_{k_i} | T_i) p(l_i = L_{k_i} | S_i) p_{H_i} p(l_i = L_{k_i} | l_{i-1} = L_{k_{i-1}}, O_i)}{z_2 [p(l_i = L_{k_i})]^2},$$

where $p_{T_i} = p(T_i | l_i = L_{k_i})$, $p_{S_i} = p(S_i | l_i = L_{k_i})$, $p_{H_i} = p(H_i | l_i = L_{k_i})$ and $z_2 = \frac{p(T_i, S_i, H_i)}{p(T_i)p(S_i)}$. Since z_2 is not a function of the labels, we can infer it by normalization.

Given a label, the distribution $p(H_i | l_i = L_{k_i})$ is modeled by a simple Gaussian distribution (as suggested in [1]). For modeling the dependency of the label in the color and texture, $p(l_i = L_{k_i} | T_i)$, and in the boundary shape, $p(l_i = L_{k_i} | S_i)$, we use an extension of SVM that provides probability estimates [14] (the SVMs used for color&texture and for boundary shapes are separate). The prior probability $p(l_i = L_{k_i})$ is computed from the occurrences of labels in the training set. Finally,

$$p(l_i = L_{k_i} | l_{i-1} = L_{k_{i-1}}, O_i) = \begin{cases} M(\text{'top'}, l_i) & i = 1 \\ M(l_{i-1}, l_i) & 1 < i < n \\ M(l_{i-1}, l_i) M(l_i, \text{'bottom'}) & i = n \end{cases}, \quad (3)$$

where M is the transition matrix described in Sec. 2.2.

Eqs. (1)-(3) are equivalent to describing the problem of estimating the class probabilities for the region labels as the problem of calculating the (multi-class) marginals in the HMM (Hidden Markov Model) in Fig. 1(c). We estimate the class probabilities by the sum product algorithm [15]. ORC classifies each land region by $\text{ORC}(i) = \arg\max_{L_j} p_{\text{ORC}}(i, L_j)$.

4 Experiments: ORC

We experiment on the *coast*, *mountain*, and *open country* datasets from Labelme [16] presented in [17]. This provides a set of 1144 256X256 images of natural scenery. With the Labelme toolbox, a Web user marks polygons in the image and freely provides a textual annotation for each. This freedom encourages the use of synonyms and spelling mistakes. Following [16], synonyms were grouped together and spelling mistakes were corrected.

In this work we do not deal with automatic segmentations and rely on manual ones provided with the dataset. We select all regions whose annotation describes background. This gives us 4979 regions annotated by 19 background labels: sky (1120), mountain (1489), sea (401), trees (622), field (366), river (150), sand

Table 1. Left: Accuracy of ORC using different combinations of cues. A significant improvement from 61.5% to 68.2% is achieved by utilizing the new cues. Right: Accuracy of M-ORC demonstrating its ability to recognize separately and simultaneously structural categories and coverage categories of land regions.

Cue	Accuracy	cue	Acc. Sky-Water-Land	Acc. Land Structure	Acc. Land Cover
Color&Texture	0.615	Color&Texture	0.865	0.849	0.570
Relative Location	0.503	Relative Location	0.862	0.750	0.468
Boundary Shape	0.452	Boundary Shape	0.866	0.836	0.517
Relative Loc. + Boundary Shape	0.573	Relative Loc. + Boundary Shape	0.881	0.835	0.530
Color&Texture + Relative Loc.	0.676	Color&Texture + Relative Loc.	0.905	0.860	0.612
Color&Texture + Boundary Shape	0.641	Color&Texture + Boundary Shape	0.887	0.862	0.567
All (ORC)	0.682	All (M-ORC)	0.909	0.876	0.605

(182), ground (94), grass (36), land (41), rocks (201), plants (143), snow (50), plateau (28), valley (20), bank (20), lake (9), beach (3), and cliff (4).

For extracting the color&textural features describing a background region, we first compute a mask that includes all its parts that do not intersect with other annotated regions (e.g., foreground objects that occlude part of it). We compute the descriptors described in Sec. 2.1 over the pixels inside this mask.

To compute the probability estimates from color&textural features and boundary appearance, we use SVM with an RBF kernel (using LIBSVM [18]). To test *ORC* we perform a 5-fold cross-validation at the image level. Each iteration starts with a parameter selection stage (c and γ ; see [18]) in which the training set is split into a training and validation test (also at image level).

We found that the height cue is noninformative and report results without using it. Fig. 2 demonstrates ORC results using each cue alone and all cues together. We can clearly see that the incorporation of new cues gives better results than using only color&texture. For instance, colored sky that is misclassified using only color&textural features is classified correctly using the relative location cue. Sea and mountain regions that are sometimes misclassified using color&texture, are correctly classified using the boundary shape cue. In the bottommost example, a narrow sea region that is missed in the ground truth is recognized. A region is sometimes misclassified when each cue is used separately, but classified correctly when all cues act as a committee. The total accuracies are reported in Table 1(a). A significant improvement from 61.5% to 68.2% is achieved by utilizing the new cues. The color&texture cues are better for classifying trees, field, rocks, plants, and snow. The new cues alone give more accurate results for sky, mountain, sea, and sand. The performance for the other classes is low for all cues due to their low occurrences in the dataset.

5 Multiple Categorization: The M-ORC Algorithm

In the second part of this work we examine the causes for errors. It turns out that results counted as errors are not always wrong. Some ambiguities are a matter of synonyms. A mountain is sometimes annotated as a cliff and vice versa. Other

Fig. 2. Demonstrating region classification results using each cue alone and all cues together, using the ORC algorithm

ambiguities, however, are created by annotations that refer to different aspects of the labeled object. In particular, it seems that land regions are sometimes annotated with a description of the land structure (mountain, plain, valley) and sometimes annotated with a description of the overlying coverage (trees, grass, sand, rocks, etc.). For instance, a mountain with trees may be classified by the ground truth as mountain and classified by ORC as trees, or vice versa. A snowy plain can be correctly labeled both as ‘snow’ and as ‘field’ or ‘plateau’.

These observations lead us to suggest an alternative labeling method, denoted *M-ORC* (Multiple Categorization Ordered Region Classifier). With this method, the correct labeling for land regions is multi-valued. Each region is first classified into one of the three categories SKY, WATER or LAND. Each region classified as LAND is further classified by a structural category and a coverage category.

More formally, $M-ORC(i) = \{c_{i1}, c_{i2}, c_{i3}\}$, where $c_{i1} \in \{\text{‘SKY’}, \text{‘WATER’}, \text{‘LAND’}\}$, $c_{i2} \in C_{\text{struct}} = \{\text{‘MOUNTAIN’}, \text{‘PLAIN’}, \text{‘VALLEY’}, \text{‘BANK’}, \emptyset\}$, and $c_{i3} \in C_{\text{cover}}$, where $C_{\text{cover}} = \{\text{‘SAND’}, \text{‘GROUND’}, \text{‘ROCKS’}, \text{‘PLANTS’}, \text{‘TREES’}, \text{‘GRASS’}, \text{‘SNOW’}, \emptyset\}$.

When $c_{i1} \in \{\text{‘SKY’}, \text{‘WATER’}\}$, $c_{i2}, c_{i3} = \emptyset$. To assign c_{i1} a score r_{main} is defined: $r_{\text{main}}(i, \text{‘SKY’}) = p_{\text{ORC}}(i, \text{‘sky’})$, $r_{\text{main}}(i, \text{‘WATER’}) = p_{\text{ORC}}(i, \text{‘sea’}) + p_{\text{ORC}}(i, \text{‘lake’}) + p_{\text{ORC}}(i, \text{‘river’})$, and $r_{\text{main}}(i, \text{‘LAND’}) = 1 - r_{\text{main}}(i, \text{‘SKY’}) - r_{\text{main}}(i, \text{‘WATER’})$. The selection of c_{i1} is according to the maximal r_{main} . For regions for which $c_{i1} = \text{‘LAND’}$, c_{i2} and c_{i3} are set. A score r_{struct} is defined for $c \in C_{\text{struct}}$: $r_{\text{struct}}(i, c) = \sum_{l \in G_c} p_{\text{ORC}}(i, l)$, where $G_c = \{\text{‘mountain’}, \text{‘cliff’}\}$ for $c = \text{‘MOUNTAIN’}$, $G_c = \{\text{‘land’}, \text{‘field’}, \text{‘plateau’}\}$ for $c = \text{‘PLAIN’}$, $G_c = \{\text{‘valley’}\}$ for $c = \text{‘VALLEY’}$, and $G_c = \{\text{‘bank’}, \text{‘beach’}\}$ for $c = \text{‘BANK’}$. The selection of c_{i2} is according to the maximal r_{struct} . For $c \in C_{\text{cover}}$, $r_{\text{cover}}(i, c) = p_{\text{ORC}}(i, l_c)$, where $l_c = \text{lowercase}(c)$. The selection of c_{i3} is according to the maximal r_{cover} .

6 Experiments: M-ORC

Each land region training example is provided with ground-truth (human labeling) of one type of label, structural or coverage. Nevertheless, by the categorization scheme suggested here, each test region, if recognized as a land region, is assigned with both a coverage and a structural label, as can be seen in Fig. 3.

When evaluating the categorization accuracy, we can only check if the available label is correct. Out of the total of 4979 regions, 1120 are annotated sky, 560 are annotated by a word describing water, and 3299 by words describing land. Out of the 3299 land regions, 1328 are described by their cover and 1971 by their structure. For land regions that are hand labeled with a structural label, we check whether it matches the structural label assigned by M-ORC, and for land regions hand labeled with a coverage label, we check whether it matches the coverage label assigned by M-ORC; see Table 1(b). Again we see the advantage of incorporating the new cues in comparison to using color&texture alone. The accuracy for the SKY-WATER-LAND categorization grows from 86% to 91%. The accuracy for classifying land structure grows from 85% to 88%, and the accuracy for classifying land coverage grows from 57% to 60%. The lower accuracy of the latter is probably due to the larger number of cover categories.

Fig. 3. Demonstrating M-ORC results: land regions are assigned with both a coverage and a structural label, while the ground-truth provides only one label for each region

7 Discussion

This paper presents: 1. A new model for contextual background region classification that uses statistical characteristics of scenery images and suggests an exact inference solution. 2. A novel method for categorizing land regions separately by their structural and their coverage categories, while learning from a training set in which each region is annotated only by one label type.

Note that an alternative solution for this scheme would be to use mixed class labels covering all combinations (e.g., "tree covered mountain") . However, this will lead to a quadratic number of classes, which implies lower performance and much greater efforts to label the training set.

One may argue that the images dealt with here are rather simple, and that modeling more complex images (street scenes, indoor scenes) is a more appropriate challenge. We agree. However, while computer vision has taken major steps forward in object detection and recognition also in complex scenes (e.g., vehicle and pedestrian detection), computer vision still has a long way to go before it can provide full interpretation and understanding of scenes. Our focus on scenery/landscape images can be considered a step backwards to deal with the tasks that

occupied the human vision system in the early stages of its evolution. Only after the visual system was able to cope with such scenery did it gradually evolve to cope with more complex scenes. Nevertheless, it is of course desirable to extend the models suggested here to more complex image classes and objects.

In this work we used pre-segmented images. In future work we intend to check our model's stability given automatic segmentation results, and to investigate the usability of the layout cues for the task of semantic segmentation.

References

1. Avraham, T., Lindenbaum, M.: Non-local characterization of scenery images: Statistics, 3D reasoning, and a generative model. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6315, pp. 99–112. Springer, Heidelberg (2010)
2. Torralba, A.B.: Contextual priming for object detection. IJCV 53(2), 169–191 (2003)
3. Kumar, S., Hebert, M.: A hierarchical field framework for unified context-based classification. In: ICCV (2005)
4. He, X., Zemel, R.S., Ray, D.: Learning and incorporating top-down cues in image segmentation. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3951, pp. 338–351. Springer, Heidelberg (2006)
5. Rabinovich, A., Vedaldi, A., Galleguillos, C., Wiewiora, E., Belongie, S.: Objects in context. In: ICCV (2007)
6. Desai, C., Ramanan, D., Fowlkes, C.: Discriminative models for multi-class object layout. In: ICCV (2009)
7. Galleguillos, C., Belongie, S.: Context based object categorization: A critical survey. Comput. Vis. Image Understand (2010)
8. Deng, J., Berg, A.C., Li, K., Fei-Fei, L.: What does classifying more than 10,000 image categories tell us? In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6315, pp. 71–84. Springer, Heidelberg (2010)
9. Fergus, R., Bernal, H., Weiss, Y., Torralba, A.: Semantic label sharing for learning with many categories. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6311, pp. 762–775. Springer, Heidelberg (2010)
10. Feenbaum, C.: Wordnet: An Electronic Lexical Database. Bradford Books (1998)
11. Niu, D., Dy, J.G., Jordan, M.I.: Multiple non-redundant spectral clustering views. In: ICML (2010)
12. Vogel, J., Schiele, B.: Semantic modeling of natural scenes for content-based image retrieval. IJCV 72(2), 133–157 (2007)
13. Haralick, R.M., Shanmugam, K., Dinstein, I.: Textural features for image classification. IEEE Transactions on Systems, Man, and Cybernetics, 610–621 (1973)
14. Wu, T.F., Lin, C.J., Weng, R.C.: Probability estimates for multi-class classification by pairwise coupling. Journal of Machine Learning Research 5, 975–1005 (2004)
15. Kschischang, F.R., Frey, B.J., Loeliger, H.A.: Factor graphs and the sum-product algorithm. IEEE Tran. on Information Theory 47(2), 498–519 (2001)
16. Russell, B.C., Torralba, A.: Labelme: a database and web-based tool for image annotation. IJCV 77, 157–173 (2008)
17. Oliva, A., Torralba, A.: Modeling the shape of the scene: a holistic representation of the spatial envelope. IJCV 42(3), 145–175 (2001)
18. Chang, C., Lin, C.: LIBSVM: a library for support vector machines. (2001) Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>