A Hierarchical Conditional Random Field Model for Labeling and Classifying Images of Man-made Scenes

Michael Ying Yang, Wolfgang Förstner
Department of Photogrammetry
Institute of Geodesy and Geoinformation, University of Bonn
Nussallee 15, 53115 Bonn, Germany
michaelyangying@uni-bonn.de, wf@ipb.uni-bonn.de

Abstract

Semantic scene interpretation as a collection of meaningful regions in images is a fundamental problem in both photogrammetry and computer vision. Images of man-made scenes exhibit strong contextual dependencies in the form of spatial and hierarchical structures. In this paper, we introduce a hierarchical conditional random field to deal with the problem of image classification by modeling spatial and hierarchical structures. The probability outputs of an efficient randomized decision forest classifier are used as unary potentials. The spatial and hierarchical structures of the regions are integrated into pairwise potentials. The model is built on multi-scale image analysis in order to aggregate evidence from local to global level. Experimental results are provided to demonstrate the performance of the proposed method using images from eTRIMS dataset, where our focus is the object classes building, car, door, pavement, road, sky, vegetation, and window.

1. Introduction

The problem of scene interpretation in terms of classifying various image components (pixels, regions, and objects) in images is a challenging task due to ambiguities in the appearance of the image data (19). These ambiguities may arise either due to the physical conditions such as illumination and pose of the scene components with respect to the camera, or due to the intrinsic nature of the data itself. Images of man-made scenes, e.g. building facade images, exhibit strong contextual dependencies in the form of spatial interactions among components. Neighboring pixels tend to have similar class labels, and different regions appear in restricted spatial configurations. Modeling these spatial structures is crucial to achieve good classification accuracy, and help alleviate ambiguities. Attempts were made to exploit the spatial structure for semantic image interpretation by using random fields. Early since nineties, Markov random fields (MRFs) have been used for image interpretation (25); the limiting factor that MRFs only allow for local features has been overcome by conditional random fields (CRFs) (21; 19), where arbitrary features can be used for classification, at the expense of a purely discriminative approach. The key idea for integrating spatial structural information into the interpretation process is to combine it with low-level object-class region probabilities in a special classification process performed by CRFs. However, standard CRFs still work on a very local level and long range dependencies are not addressed explicitly in such models. Therefore, the key idea is to integrate hierarchical structural information into the interpretation process by constructing hierarchical CRFs on image regions on multiple scales.

In this paper, we propose a novel hierarchical conditional random field model. The major contribution is that a hierarchical CRF is constructed by extending the standard CRF to integrate hierarchical structure of the regions into pairwise potentials. Using multi-scale mean shift segmentation, the hierarchical CRF model aggregates evidence from local to global level. Hierarchical CRF model only exploit up to second-order cliques, which makes learning and inference much easier. We show that hierarchical CRF classification results give average performance of 69.0% compared to standard CRF 65.8% on eTRIMS 8-class dataset (18).

The complete proposed workflow for interpreting images of man-made scenes is sketched in Figure 1. The illustration shows that graphical model can provide a consistent model representation including spatial and hierarchical structures, and therefore outperforms the classical local classification approach. The following sections are organized as follows. The related works are discussed in Section 2. In Section 3, the standard CRF is introduced. In Section 4, a hierarchical conditional random field is presented as an extension of CRF by explicitly modeling spatial and hierarchical structures. In Section 5, experimental results are presented. Fi-

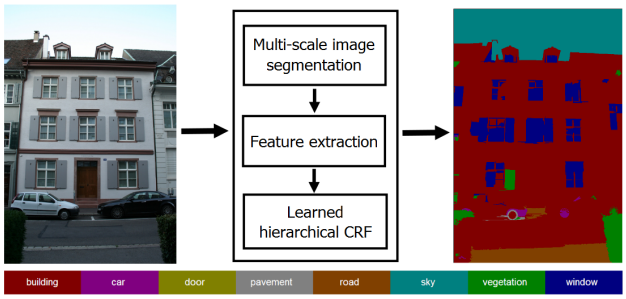


Figure 1. The workflow of our hierarchical CRF model. Given a test image, we run the segmentation and feature extraction, then we label the test image using the hierarchical CRF model learned by training images.

nally, this work is concluded and future direction is discussed in Section 6.

2. Related work

There are many previous works on contextual models that exploit spatial dependencies between objects. For this, several authors explore Markov random fields (MRFs) and conditional random fields (CRFs) for probabilistic modeling of local dependencies, e.g., (3; 25; 19; 14; 31).

(19) present a discriminative conditional random field framework for the classification of image regions by incorporating neighborhood interactions in the labels as well as the observed data. The model allows to relax the strong assumption of conditional independence of the observed data generally used in the MRF framework for tractability. (31) propose an approach for learning a discriminative model of object classes, incorporating texture, layout, and context information. Unary classification and feature selection is achieved using a boosting scheme. Image segmentation is achieved by incorporating the unary classifier in a CRF, which captures the spatial interactions between class labels of neighboring pixels. (23) propose an approach that learns a CRF to combine bottom-up and top-down cues for class specific object segmentation.

By exploiting both spatial and hierarchical structures, a number of CRF models for image interpretation address the combination of global and local features (13; 34; 27; 12; 32; 26; 28). They showed promising results and specifically improved performance compared with making use of only one type of feature - either local or global.

(13) propose a multi-layer CRF to account for global consistency, which shows improved performance. The authors introduce a global scene potential to assert consistency of local regions. Thereby, they were able to benefit from integrating the context of a given scene. (34) propose a model that combines appearance over large contiguous regions with spatial information and a global shape prior. The shape prior provides local context for certain types of objects (e.g., cars and airplanes), but not for regions representing general objects (e.g., animals, buildings, sky and grass). (32) present a proposal of a general framework that explicitly models local and global information in a CRF. Their method resolves local ambiguities from a global perspective using global image information. It enables locally and globally consistent image recognition.

Besides the above approaches, there are more popular methods to solve multi-class classification problems using higher order conditional random fields (15; 16; 20). (15) introduce a class of higher order clique potentials called P^n Potts model. Higher order clique potentials have the capability to model complex interactions of random variables, enabling them to better capture the rich statistics of natural scenes. The higher order potential functions proposed in (16) take the form of the Robust P^n model, which is more general than the P^n Potts model. (20) generalize Robust P^n model to P^n based hierarchical CRF model. Inference in these models can be performed efficiently using graph cut based move making algorithms. (9) propose the use of a soft cost over the number of labels present in an image for clustering.

Recent work by (26) comprises two aspects for coupling local and global evidences both by constructing a tree-structured CRF on image regions on multiple scales, which largely follows the approach of (27), and using global image classification information. Thereby, (26) neglect direct local neighbourhood dependencies. The work of (29) explicitly attempts to combine the power of global feature-based approaches with the flexibility of local feature-based methods in one consistent framework. Briefly, (29) extend classical one-layer CRF to multi-layer CRF by restricting pairwise potentials to regular 4-neighborhood model and introducing higher-order potentials between different layers. (28) propose a hierarchical multi-feature representation and automatically learn flexible hierarchical object models. Their work combines structure learning in conditional random fields and discriminative parameter learning of classifiers using hierarchical features. In (36), the authors present a concept of hierarchical CRF that models region adjacency graph and region hierarchy graph structure of an image. (35) present an approach for regionwise classification of building facade images using a conditional random field model. (1) propose an efficient large margin piecewise learning method for CRF model, which reduces to an equivalent convex problem with a small number of constraints.

3. Image labeling using standard CRF

Given a random field $\boldsymbol{X}$ defined over a graph $\mathcal{H} = (\mathcal{V}, \mathcal{E})$, the standard CRF model for determining the optimal labeling $\boldsymbol{x} = \{x_i\}$, based on the image data $\boldsymbol{d}$, has a

distribution of the Gibbs form

$$P(\boldsymbol{x} \mid \boldsymbol{d}) = \frac{1}{Z} \exp\left(-E(\boldsymbol{x} \mid \boldsymbol{d})\right) \tag{1}$$

with the energy function defined as (35)

$$E(\boldsymbol{x} \mid \boldsymbol{d}) = \sum_{i \in \mathcal{V}} E_1(x_i) + \alpha \sum_{\{i,j\} \in N} E_2(x_i, x_j) \tag{2}$$

where α is the weighting coefficient in the model, and Z is the normalization factor. The set $\mathcal{V}$ is the set of nodes in the complete graph. The set N is the set of pairs collecting neighboring. E_1 is the unary potentials, which represent relationships between variables and the observed data. E_2 is the pairwise potentials, which represent relationships between variables of neighboring nodes. In the remainder of this section, we will discuss the unary and pairwise potentials in details.

3.1. Unary potentials

The local unary potentials E_1 independently predict the label x_i based on the image $\boldsymbol{d}$

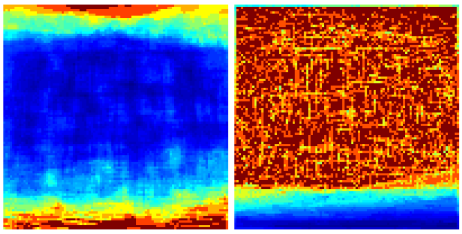
$$E_1(x_i) = -\log P(x_i \mid \boldsymbol{d}) \tag{3}$$

The label distribution $P(x_i \mid \boldsymbol{d})$ is usually calculated by using a classifier. Same as in (30), we take randomized decision forest (RDF) (6) as the classifier which operates on the regions defined by some unsupervised segmentation methods. A RDF is an ensemble classifier that consists of T decision trees (30). In order to train the classifier, each region is assigned the most frequent class label it contains. Then a RDF is trained on the labeled data for each of the object classes. The label distribution is defined directly by the probability outputs provided by RDF for each region.

Based on the fact that RDF classifier does not take class label location information explicitly, we incorporate location potentials (similar to 31; 35) in unary potentials. The location potential $-\log Q(x_i \mid \boldsymbol{d})$ is the negative logarithm of the probability function of class labels x_i given image coordinates $\boldsymbol{z}_i$ as the center of the region i, and takes the form of a look-up table with an entry for each class x_i and region center location $\boldsymbol{z}_i$, where

$$Q(x_i \mid \boldsymbol{d}) = \left(\frac{N_{x_i, \hat{\boldsymbol{z}}_i} + 1}{N_{\hat{\boldsymbol{z}}_i} + 1}\right)^2$$

The index $\hat{z}_i$ is the normalized version of the region center $\boldsymbol{z}_i$, where the normalization allows for images of different sizes: the image is mapped onto a canonical square and $\hat{\boldsymbol{z}}_i$ indicates the pixel position within this square. $N_{x_i, \hat{\boldsymbol{z}}_i}$ is the number of regions of class x_i at normalized location in $\hat{\boldsymbol{z}}_i$, and $N_{\hat{\boldsymbol{z}}_i}$ is the total number of regions at location in $\hat{\boldsymbol{z}}_i$. The location potentials capture the dependence of the class label on the rough location of the region in the image. For example, in our experiment, we use part of annotation images in 8-class eTRIMS dataset (18) to learn the location potential, but ensure no overlap between these images and testing images in the experimental part. Some learned location potentials are illustrated in Figure 2. From Figure 2, we see *sky* tends to occur at the top part of images, while *road* tends to occur at the bottom part of images, and *building* tends to occur in the middle part of images. Here, the dark blue area indicates the most likely locations of one class, while the dark red area indicates the most unlikely locations. Therefore, the unary potentials E_1 are written as



(a) building (b) road

(c) sky (d) vegetation

Figure 2. Example location potentials. Part of annotation images in 8-class eTRIMS dataset (18) is used to learn the location potentials. The annotation images are mapped onto a canonical square. The size of each image is 100×100 here. Here, the dark blue area indicates the most likely locations of one class, while the dark red area indicates the most unlikely locations.

$$E_1(x_i) = -\log P(x_i \mid \boldsymbol{d}) - \log Q(x_i \mid \boldsymbol{d}) \tag{4}$$

We now describe how the features for RDF classifier are constructed from low-level descriptors. For each region, we compute an 178-dimensional feature vector, first incorporating region area and perimeter, its compactness and its aspect ratio. For representing spectral information of the region, we use nine color features as (2): the mean and the standard deviation of the RGB and the HSV color spaces. Texture features derived from the Walsh transform (22) are also used. Additionally we use mean SIFT descriptors (24) of the image region. SIFT descriptors are extracted for each pixel of the region at a fixed scale and orientation, which is practically the same as the HOG descriptor (8), using the fast SIFT framework in (33). The extracted descriptors are

then averaged into one l_1-normalized descriptor vector for each region. Other features are derived from generalization of the region's border and represent parallelity or orthogonality of the border segments. We select the four points of the boundary which are farthest away from each other. From this polygon region with four corners, we derive three central moments, and eigenvalues in direction of major and minor axis, aspect ratio of eigenvalues, orientation of polygon region, coverage of polygon region, and four angles of polygon region boundary points.

3.2. Pairwise potentials

The pairwise potentials E_2 describe category compatibility between neighboring labels x_i and x_j given the image $\boldsymbol{d}$, which take the form

$$E_2(x_i, x_j) = \frac{1 + 4\exp(-2c_{ij})}{0.5(N_i + N_j)} \delta(x_i \neq x_j) \qquad (5)$$

where c_{ij} is the l_2 norm of the color difference between regions in the HSV color space. N_i is the number of regions neighbored to region i, and N_j is the number of regions neighbored to j. The potentials E_2 are scaled by N_i and N_j to compensate for the irregularity of the graph $\mathcal{H}$. We refer the reader to (31; 12) for more details about designing pairwise potentials.

4. Image labeling using hierarchical CRF

As seen from last section, standard CRF acts on a local level and represents a single view on the image data typically represented with unary and pairwise potentials. To overcome those local restrictions, we analyze the image at multiple scales to enhance the model by evidence aggregation on a local to global level. Furthermore, we integrate pairwise potentials to regard the hierarchical structure of the regions.

4.1. The Hierarchical CRF model

Given a random field $\boldsymbol{X}$ defined over a graph $\mathcal{H} = (\mathcal{V}, \mathcal{E})$, the goal is to determine a label x_i for each region i based on the image data $\boldsymbol{d}$. The complete model for determining the optimal labeling $\boldsymbol{x} = \{x_i\}$ has a distribution of the Gibbs form as Equation 1 with the energy function defined as

$$E(\boldsymbol{x} \mid \boldsymbol{d}) = \sum_{i \in \mathcal{V}} E_1(x_i) + \alpha \sum_{\{i,j\} \in N} E_2(x_i, x_j) + \beta \sum_{\{i,k\} \in H} E_3(x_i, x_k) \qquad (6)$$

where α and β are the weighting coefficients in the model, and Z is the normalization factor. The set $\mathcal{V}$ is the set of nodes in the complete graph. The set N is the set of pairs collecting neighboring nodes within each scale, and H is the set of pairs collecting parent-child relations between regions with neighboring scales. E_1 is the unary potential, and E_2 is the local pairwise potential representing relationships between variables of neighboring regions within each scale. E_3 is the hierarchical pairwise potential, which represents relationships between regions with neighboring scales.

The full graphical model is illustrated in Figure 3. Three layers are connected via region hierarchy. The development of the regions over several scales is used to model the region hierarchy. (10) defined a region hierarchy with directed edges between regions of successive scales. Furthermore, the relation is defined over the maximal overlap of the regions. If the edges would be undirected, the region hierarchy would only consist of trees. Nodes connection and numbers correspond to multi-scale segmentation of a building facade image. The blue edges between the nodes represent the neighborhoods at one scale, and the red edges represent the hierarchical relation between regions. Note that this model only exploits up to second-order cliques, and combines different views on the data by the hierarchical structure accounting for longer range dependencies.

The formulation of unary potentials E_1 and local pairwise potential E_2 is same in Section 3.2, except that regions in multi-scale are taken into account instead of regions in only one scale. The hierarchical pairwise potentials E_3 describe category compatibility between hierarchically neighboring labels x_i and x_k given the image $\boldsymbol{d}$, which take the form of a data-dependent model

$$E_3(x_i, x_k) = (1 + 4\exp(-2c_{ik}))\, \delta(x_i \neq x_k) \qquad (7)$$

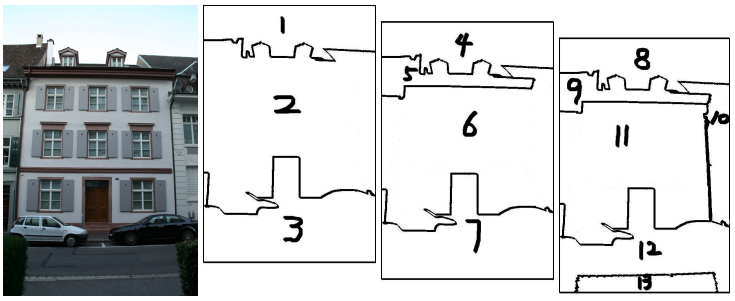
where c_{ik} is the l_2 norm of the color difference between regions in the HSV color space. Hierarchical pairwise potentials act as a link across scale, facilitating propagation of information in the model.

4.2. Learning and inference

In our formulation, we have two weights α and β which represent the trade-off among hierarchical, local pairwise regularization and the confidence in unary potentials. We estimate α and β by cross validation on the training data. Once the model has been learned, inference is carried out with the multi-label graph optimization library of (5; 17; 4) using α-expansion. Since the hierarchical CRF is defined on the graph of regions, inference is very efficient, taking less than half a second per image.

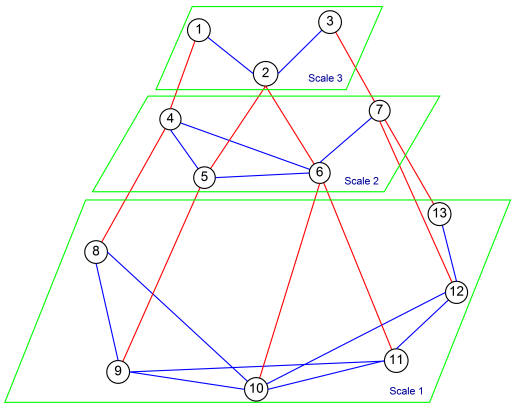
5. Experimental results

We conduct experiments to evaluate the performance of the hierarchical CRF model on the 8-class eTRIMS dataset (18). The dataset consists of 60 building facade images, labeled with 8 classes: *building*, *car*, *door*, *pavement*, *road*,



(a) Original building facade image

(b) Multi-scale segmentation (from left to right: top, middle and bottom scale)



(c) The graphical model

Figure 3. Illustration of the hierarchical CRF model architecture. Three layers are connected via region hierarchy. Nodes connection and numbers correspond to multi-scale segmentation of a building facade image. The blue edges between the nodes represent the neighborhoods at one scale, the red edges represent the hierarchical relation between regions.

sky, *vegetation*, *window*. These classes are typical objects which can appear in images of building facades. In the experiments, we take the ground-truth label of a region to be the majority vote of the ground-truth pixel labels. We randomly divide the images into a training set with 40 images and a testing set with 20 images.

We segment the images using mean shift algorithm (7). Our approach uses the Gaussian scale-space for obtaining regions at several scales. For each scale, we convolve each image channel with a Gaussian filter and apply mean shift algorithm to segment the smoothed image. As a result of the mean shift algorithm, we obtain a complete partitioning of the image for each scale, where every image pixel belongs to exactly one region. In all 60 images, we extract around 61 000 regions. We use three layers in the scale space, resulting the ground layer often contains around 500 regions, and the number decrease down to 200 in the highest layer. For the example image, Figure 4 shows one example result from multi-scale mean shift segmentation, where the color of each region is assigned randomly that neighboring regions are likely to have different colors.

Figure 5 shows the classification results from hierarchical CRF with multi-scale mean shift. We run the experiment five times, and get the overall classification accuracy 69.0%. For comparison, we also run the the experiments on RDF and standard CRF. The number T of decision tree for RDF is 250. RDF classifier alone gives an overall accuracy of 58.8%, and standard CRF gives an overall accuracy of 65.8%. Therefore, the hierarchical potentials increase the accuracy by 3.2%. Table 1 and Table 2 show two confu-

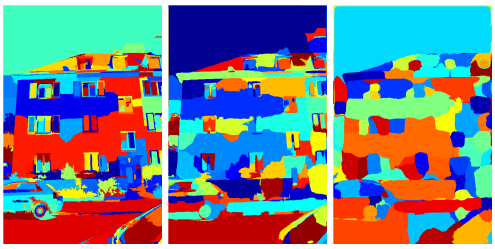


Figure 4. One example region images of the mean shift segmentation (7) result at scale 1, 2, 3, respectively. The color of each region is assigned randomly that neighboring regions are likely to have different colors.

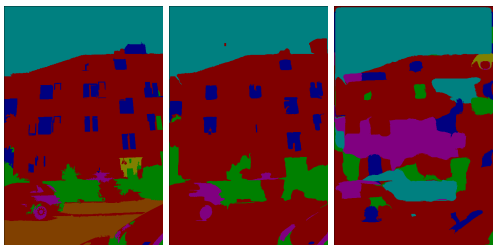


Figure 5. Classification results using hierarchical CRF from 3-scale mean shift segmentation in Figure 4. From *left* to *right*: classification result at scale 1, 2, 3, respectively.

Pr \ Tr	b	c	d	p	r	s	v	w
b	**71**	2	1	1	1	2	10	12
c	12	**35**	0	12	11	0	30	0
d	42	0	**16**	1	6	0	8	27
p	11	15	0	**22**	36	0	14	2
r	4	8	0	44	**35**	0	9	0
s	13	0	0	0	0	**78**	8	1
v	18	5	2	1	1	0	**66**	7
w	19	1	1	0	0	1	3	**75**

Table 1. Pixelwise accuracy of image classification using standard CRF on the eTRIMS 8-class dataset. The confusion matrix shows the classification accuracy for each class (rows) and is row-normalized to sum to 100%. Row labels indicate the true class (Tr), and column labels the predicted class (Pr). (b = *building*, c = *car*, d = *door*, p = *pavement*, r = *road*, s = *sky*, v = *vegetation*, w = *window*.)

sion matrices obtained by applying standard CRF and hierarchical CRF to the whole test set, respectively. Accuracy values in the table are computed as the percentage of image pixels assigned to the correct class label, ignoring pixels labeled as void in the ground truth. Compared to the confusion matrix showing standard CRF in Table 1, hierarchical CRF performs significantly better on *pavement*, *vegetation*, *road*, and *window* classes, slightly better on *car* and *sky* classes, and slightly worse on *building* and *door* classes. The weighting parameter settings, learned by cross validation on the training data, are $\alpha = 0.1$, $\beta = 0.65$.

Qualitative results of hierarchical CRF with multi-scale mean shift on the eTRIMS dataset are presented in Figure 6. The quality inspection of the results in these images shows that hierarchical CRF yields significant improvement. By visual inspection of classification results for a challenging test image in Figure 6, we have demonstrated that hierarchical CRF outperform the method either with only spatial information or without context information.

The best accuracies are for classes which have low visual variability and many training examples (such as window, vegetation, building, and sky) whilst the lowest accuracies are for classes with high visual variability or few

Pr \ Tr	b	c	d	p	r	s	v	w
b	**67**	3	1	4	5	1	8	11
c	17	**36**	0	11	9	0	26	1
d	50	5	**14**	8	0	0	7	16
p	6	4	0	**85**	1	0	4	0
r	0	11	0	21	**53**	0	15	0
s	11	0	0	0	0	**80**	8	1
v	9	5	1	0	1	0	**78**	6
w	15	0	1	0	0	2	2	**80**

Table 2. Pixelwise accuracy of image classification using hierarchical CRF with multi-scale mean shift on the eTRIMS 8-class dataset. The confusion matrix shows the classification accuracy for each class (rows) and is row-normalized to sum to 100%. Row labels indicate the true class (Tr), and column labels the predicted class (Pr). (b = *building*, c = *car*, d = *door*, p = *pavement*, r = *road*, s = *sky*, v = *vegetation*, w = *window*.)

training examples (for example door and car). We expect more training data will improve the classification accuracy. Objects such as car, door, and window are sometimes incorrectly classified as *building*, due to the dominant presence of building in the image.

With the current settings for the local and hierarchical pairwise potential functions, our methods tend to produce rather low classification rate for object classes with minor instances (e.g., *car* and *door*). An investigation into more sophisticated potential functions might resolve this problem. In computer vision, the pairwise potentials are usually represented by a weighted summation of many features functions (e.g., (31)), and the parameters with the size as same as feature number are learned from the training data. By maximizing the conditional log-likelihood, better accuracy usually obtained. But this kind of parameter learning remains a difficult problem and also is most time-consuming part (1). While in our proposed graphical model formulations, we simply have at most two weighting parameters (similar to (12; 11; 20)). So this is the trade-off between efficiency and accuracy.

6. Conclusion

In this paper, we have addressed the problem of incorporating two different types of context information, i.e., spatial structure and hierarchical structure for image interpretation of man-made scenes. We present an approach to classify images into regions of *building*, *car*, *door*, *pavement*, *road*, *sky*, *vegetation*, and *window*. To exploit different levels of contextual information in images, a hierarchical conditional random field is described as an extension of standard CRF. The hierarchical structure of the regions is integrated into pairwise potentials. The model is built on multi-scale mean shift segmentation in order to aggregate evidence from local to global level. We have evaluated our approach on bench-

Figure 6. Qualitative classification results on testing images from the eTRIMS dataset. The hierarchical CRF model yields more accurate and cleaner results than the standard CRF and RDF. (1st-row) Testing images. (2nd-row to 5th-row) Classification results using the RDF classifier, the CRF model, the hierarchical CRF model (HCRF), and the ground truth (GT), respectively. (6th-row) Legend.

mark dataset. The results show that hierarchical CRF outperforms both the standard CRF and local RDF classifier.

The proposed hierarchical CRF operates on region level resulting from mean shift segmentation algorithm, which allows for fast inference. However, one disadvantage of such an approach is that mistakes in the initial unsupervised segmentation, in which regions span multiple object classes, cannot be recovered from. For each region from segmentation, a class label is commonly assigned to the region according to the majority vote of the ground-truth pixel labels. At the starting point, ambiguity is introduced in the region ground-truth labeling. As a future work, we will try to resolve this problem by assigning a class probability vector to the region, not assigning most probable label to the region. We could result in a probability estimation model of image segmentation regions. One could also eliminate inconsistent regions by employing another hierarchical CRFs (20), which allow for the integration of region-based CRFs with a low-level pixel based CRF.

References

[1] K. Alahari, C. Russell, and T. Philip. Efficient piecewise learning for conditional random fields. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 895–901, 2010. 2, 6

[2] K. Barnard, P. Duygulu, N. D. Freitas, D. Forsyth, D. Blei, and M. Jordan. Matching Words and Pictures. *Journal of Machine Learning Research*, 3:1107–1135, 2003. 3

[3] K. Barnard and D. Forsyth. Learning the semantics of words and pictures. In *International Conference on Computer Vision*, volume 2, pages 408–415, 2001. 2

[4] Y. Boykov and V. Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26:1124–1137, 2004. 4

[5] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23:1222–1239, 2001. 4

[6] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. 3

[7] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):603–619, 2002. 5

[8] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 886–893, 2005. 3

[9] A. Delong, A. Osokin, H. N. Isack, and Y. Boykov. Fast approximate energy minimization with label costs. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2173–2180, 2010. 2

[10] M. Drauschke. An irregular pyramid for multi-scale analysis of objects and their parts. In *IAPR-TC-15 Workshop on Graph-based Representations in Pattern Recognition*, pages 293–303, 2009. 4

[11] B. Fulkerson, A. Vedaldi, and S. Soatto. Class segmentation and object localization with superpixel neighborhoods. In *International Conference on Computer Vision*, pages 670–677, 2009. 6

[12] S. Gould, J. Rodgers, D. Cohen, G. Elidan, and D. Koller. Multi-class segmentation with relative location prior. *International Journal of Computer Vision*, 80(3):300–316, 2008. 2, 4, 6

[13] X. He, R. Zemel, and M. Carreira-perpin. Multiscale conditional random fields for image labeling. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 695–702, 2004. 2

[14] X. He, R. Zemel, and D. Ray. Learning and incorporating top-down cues in image segmentation. In *European Conference on Computer Vision*, pages 338–351, 2006. 2

[15] P. Kohli, M. Kumar, and P. Torr. P3 & Beyond: Solving Energies with Higher Order Cliques. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007. 2

[16] P. Kohli, L. Ladicky, and P. Torr. Robust Higher Order Potentials for Enforcing Label Consistency. *International Journal of Computer Vision*, 82(3):302–324, 2009. 2

[17] V. Kolmogorov and R. Zabih. What energy functions can be minimized via graph cuts? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(2):147–159, 2004. 4

[18] F. Korč and W. Förstner. eTRIMS Image Database for Interpreting Images of Man-Made Scenes. In *TR-IGG-P-2009-01, Department of Photogrammetry, University of Bonn*, 2009. 1, 3, 4

[19] S. Kumar and M. Hebert. Discriminative random fields: A discriminative framework for contextual interaction in classification. In *IEEE International Conference on Computer Vision*, volume 2, pages 1150–1157, 2003. 1, 2

[20] L. Ladicky, C. Russell, P. Kohli, and P. Torr. Associative hierarchical crfs for object class image segmentation. In *International Conference on Computer Vision*, pages 739–746, 2009. 2, 6, 7

[21] J. Lafferty, A. McCallum, and F. Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *International Conference on Machine Learning*, pages 282–289, 2001. 1

[22] G. Lazaridis and M. Petrou. Image registration using the Walsh transform. *IEEE Transactions on Image Processing*, 15(8):2343–2357, 2006. 3

[23] A. Levin and Y. Weiss. Learning to combine bottom-up and top-down segmentation. In *European Conference on Computer Vision*, pages 581–594, 2006. 2

[24] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004. 3

[25] J. W. Modestino and J. Zhang. A Markov Random Field Model-Based Approach to Image Interpretation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(6):606–615, 1992. 1, 2

[26] N. Plath, M. Toussaint, and S. Nakajima. Multi-class image segmentation using conditional random fields and global classification. In *International Conference on Machine Learning*, pages 817–824, 2009. 2

[27] J. Reynolds and K. Murphy. Figure-ground segmentation using a hierarchical conditional random field. In *Canadian Conference on Computer and Robot Vision*, pages 175–182, 2007. 2

[28] P. Schnitzspan, M. Fritz, S. Roth, and B. Schiele. Discriminative structure learning of hierarchical representations for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2238–2245, 2009. 2

[29] P. Schnitzspan, M. Fritz, and B. Schiele. Hierarchical support vector random fields: Joint training to combine local and global features. In *European Conference on Computer Vision*, pages 527–540, 2008. 2

[30] J. Shotton, M. Johnson, and R. Cipolla. Semantic texton forests for image categorization and segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008. 3

[31] J. Shotton, JohnWinn, C. Rother, and A. Criminisi. Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *European Conference on Computer Vision*, pages 1–15, 2006. 2, 3, 4, 6

[32] T. Toyoda and O. Hasegawa. Random field model for integration of local information and global information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(8):1483–1489, 2008. 2

[33] A. Vedaldi and B. Fulkerson. Vlfeat: An open and portable library of computer vision algorithms. *http://www.vlfeat.org/*, 2008. 3

[34] L. Yang, P. Meer, and D. Foran. Multiple class segmentation using a unified framework over mean-shift patches. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007. 2

[35] M. Y. Yang and W. Förstner. Regionwise Classification of Building Facade Images. In *Photogrammetric Image Analysis*, LNCS 6952, pages 209–220. Springer, 2011. 2, 3

[36] M. Y. Yang, W. Förstner, and M. Drauschke. Hierarchical Conditional Random Field for Multi-class Image Classification. In *International Conference on Computer Vision Theory and Applications*, pages 464–469, 2010. 2