FAST SEMANTIC SCENE SEGMENTATION WITH CONDITIONAL RANDOM FIELD

Wen Yang[1], *Dengxin Dai*[1], *Bill Triggs*[2], *Guisong Xia*[3], *Chu He*[1]

[1]School of Electronic Information, Wuhan University, China
[2] Laboratoire Jean Kuntzmann, CNRS-INRIA, Grenoble University, France
[3] CNRS-LTCI, TELECOM ParisTech, 46 rue Barrault, 75013 Paris, France

ABSTRACT

In this paper, we present a fast approach to obtain semantic scene segmentation with high precision. We employ a two-stage classifier to label all image pixels. First, we use the regularized logistic regression to combine different appearance-based features and the improved spatial layout of labeling information. In the second stage, we incorporate the local, regional and global cues into a conditional random field model to provide a final segmentation, and a fast max-margin training method is employed to learn the parameters of the model quickly. The comparison experiments on four multi-class image segmentation databases show that our approach can achieve comparable semantic segmentation results and work faster than that of the state-of-the-art approaches.

***Index Terms*—** Scene segmentation, image labeling, logistic regression, conditional random field.

1. INTRODUCTION

Segmentation of images into disjoint regions and interpretation of the regions for semantic meanings are two central tasks in an image analysis system. It jointly performs multi-class scene segmentation and object recognition, also called image labeling, which requires to assigning every pixel by one of the predefined semantic classes, such as buildings, trees, water, car, and etc. Though considerable efforts have been made, it remains a challenging problem due to the well-known local ambiguity.

Recently, many innovative works have been proposed to partially solve this problem by employing the informative contextual information, and this is often achieved by building a random field model over the images to encode the unary and pairwise probabilistic preferences. He et al. [1] proposed a multi-scale conditional random field (CRF) to combine multi-scale features, however, it employed an inefficient stochastic sampling for learning the model and inferencing the labels. Kumar et al. [2] presented a two-layer CRF to encode the long-range and short-range interactions. Shotton et al. [3] described a discriminative model of object classes by incorporating texture, layout, and context information efficiently. Verbeek and Triggs [4] learned a CRF from partially labeled data and incorporated top-down aggregated features to improve the segmentations. Schroff et al. [5] incorporated globally learnt class models into a random forest classifier with multiple features, and imposed spatial smoothing via a CRF model for a further increase in performance. Gould et al. [6] proposed a CRF model augmented with a novel image-dependent relative location feature which can model complicated spatial relationships, and achieved results above state of the art. Toyoda [7] presented a conditional random field that models local information and global information explicitly, which resolves local ambiguities from a global perspective using global image information, and demonstrated good performance in image labeling of two small data sets. However, many of these works still perform learning rely on hand-tuned parameters.

Most similar to us is the work of Verbeek and Triggs [4] which build a CRF segmentation model to capture the global context of image as well as the local information. However, there are several important differences with respect to our work. First, we add a new feature channel of texton and replace the absolute position information in [4] with a more informative position feature which represents the global spatial configuration of labels. Second, unlike [4], which uses a histogram of visual words representation for each patch, we represent each patch as concatenated vectors of posterior "topic" probabilities, which helps to remove the redundancy that maybe present in the basic "bag of features" model. Moreover, a lower dimensional latent topic representation speeds up computation. Third, we incorporate the regional information into our CRF model through a MRF neighborhood system, which implicitly includes the relative location information of different object classes. We finally employ the recently proposed FastPD [8] algorithm and cutting plane algorithm [9] to efficiently implement the maximum margin learning of parameters for our CRF model, and demonstrate significant improvements in computation speed and applicability.

In the remainder of this paper, we first describe how to extract and represent the local and context information in Section 2, and then we propose our two-stage segmentation model and learning method in Section 3. In Section 4, we demonstrate the experimental results. The conclusions and future work are given in Section 5.

2. LOCAL AND CONTEXT DESCRIPTORS

In this section, we first describe the extraction of visual features in more detail. Then, we introduce a low-dimensional semantic representation using supervised PLSA. Next, we introduce our improved object labels spatial layout information. We finally show how to obtain the regional and global context information.

2.1. Local patch descriptors

We perform feature extraction by dividing the images into 20×20 pixels patches with 10 pixels interval and compute three types of features from each patch: SIFT, textons, and color. Textons are computed based on an efficient implementation of computing gabor features named "simple Gabor feature space" which leads to a remarkable computational enhancements. The texton descriptor is the histogram of texton indices within the patch. To further enhance the robustness of color descriptors under photometric and geometrical changes for different scenes, we use a consolidated representation for each patch through concatenating the normalized hue descriptor and opponent angle [10]. The former is robust to scenes with saturated colors, while the latter is suitable for scenes with less saturated colors.

2.2. Low-dimensional semantic representation

In standard PLSA, each topic t is characterized by its distribution $P(w|t)$ over the words of the dictionary, and each document d is characterized by its vector of mixing weights $P(t|d)$ over topics. Then the probability model $P(w|d)$ is defined by the mixture,

$$P(w|d) = \sum_{t=1}^{T} P(w|t)P(t|d) \tag{1}$$

Generally speaking, both $P(w|t)$ and $P(t|d)$ are estimated by EM algorithm, However, here we assume $P(w|t)$ is obtained by simply count the occurrence of words and topics in the training images. In this case, the semantic topics are explicitly defined, PLSA can be thought of as a data-driven technique that use the fact that a given group of words or observations all originated from the same document or image to infer a context specific prior via statistical inversion [11]. By considering image patches as distributions of topics, we can use the topic distribution as the patch feature representation. The number of object classes defines the dimensionality of the intermediate topic space. Each topic induces a probability density on the space of low-level features, and each patch is represented as the vector of posterior topic probabilities.

2.3. Spatial layout distributions of scene categories

We have described a patch by integrating the color, structure and texture information. However, we have not included the spatial position information of patches. There are different ways to involve the spatial layout information, such as the absolutely position information in [4], simple objects relative location relations in [12], the image-dependent relative location feature in [6], the global spatial position information in [7]. Following the work of [7], we adopt a modified version of global location information by using a spatial pyramid matching (SPM) method to obtain the scene similarity, the idea behind which is that similar scenes tend to share a similar configuration of category distributions. In [7], the authors considered the contributions of all the training images to the pixelwise distribution, which is suitable for the small dataset, such as sowerby and corel dataset they used. However, for the more comprehensive and complex datasets, such as MSRC-9 class and MSRC-21 class dataset, using all the training data will leads to high computational effort and also decrease the performance slightly in practice. Therefore, we employ the k-Nearest Neighbor idea to compute the pixelwise spatial label distribution. In more detail, it selects the k most similar images to the new test image within the training database (using the similarity metric based on SPM above). Then it predicts the pixelwise spatial label distribution of the test image by weighting the category label distributions of the k similarest training images.

2.4. The regional and global context information

To describe the relationship of the central patch and its neighbors, we use the MRF neighbourhood system. The shape of a neighbor set may be regarded as the hull enclosing all the sites in the set. The regional context information for a given neighborhood is computed by taking the topic distribution of each patch and concatenating them together directly, resulting in a high-dimensional feature vector depends on the number of topics and the size of neighborhood system. The neighbour system information implicitly includes relative position information of different objects. For taking the image-level context into account, we also use the averaged topic probability on the whole image as the global aggregated features [4]. Intuitively, regional and global contexts should be complementary, as they capture different types of dependencies. The regional context partially includes the relative position information of different topics, and can yield spatially varying priors. The image-level context can capture the dependence of all the patches within the image on the same underlying scene, but it can only produce priors that are constant over the entire image.

3. SCENE SEGMENTATION USING CRF MODEL

Our labeling framework is a two-stage method involving the logistic regression model and CRF model. Fig.1 shows the flowchat of labeling process. At the first stage, LRC is trained to predict the posterior probability vectors of each patch. Essentially, it firstly uses the supervised PLSA to compute the topics probabilities of each patch with respect to the three different feature channels-SIFT, color and gabor. Then, a LRC is applied to classify each patch by concatenating the three predicted topic distribution by PLSA and the predicted spatial labels distribution as features. At the second stage, we use a CRF to learn correlations between neighboring output labels, which helps resolve ambiguities.

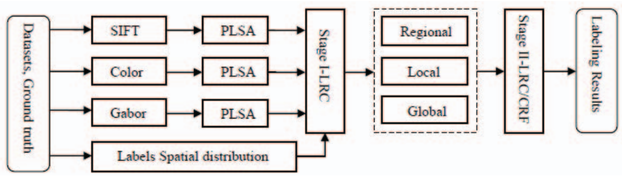

Fig. 1. Pipeline of our two-stage segmentation method

3.1. CRF model

Standard CRF has the following distribution form:

$$P(X|Y) = \frac{1}{Z} exp \left\{ - \left[\sum_{i \in \mathcal{V}} \phi_i(f_i) + \sum_{(i,j) \in \mathcal{E}} \psi_{ij}(x_i, x_j) \right] \right\} \tag{2}$$

where Y is an input data, X is the corresponding labels, Z is the partition function, $\mathcal{V}$ is a set of nodes of the image, and $\mathcal{E}$ is the pairs of adjacent nodes. $\phi(\cdot)$ is the unary potential term, and $\psi(\cdot)$ is the pairwise term. In this paper we label images at the level of small patches, using CRF to incorporate the local feature (current patch is considered) functions, the regional neighbors of the current patch and more global "context capturing" feature functions that depend on aggregates of observations over the whole image. Our energy formulation can be written as follows,

$$E(X,Y) = \sum_{i \in \mathcal{V}} \sum_{w=1}^{W} \left(\alpha_{wl} y_{iw}^{loc} + \sum_{n=1}^{N} \beta_{nwl} y_{iwn}^{reg} + \gamma_{wl} y_{w}^{glo} \right) + \sum_{(i,j) \in \mathcal{E}} \psi_{ij}(x_i, x_j) \tag{3}$$

where y_i denotes a W-dimensional feature vector, $x_i \in \{1, \ldots, l, \ldots, L\}$ denotes the label of node i. (i, j) denotes the set of all adjacent (4-neighbor) pairs of patches i, j. The parameters α_{wl} , β_{nwl} and γ_{wl} are $W \times L$,$N \times W \times L$ and $W \times L$ matrices of coefficients to be learned, respectively. For the pairwise potential, we choose a simple Potts model as done in [4].

3.2. Max-margin learning of CRF model

Max-margin learning method employs the energy function of CRF model as a discriminative function. The advantage of the margin-based approach is that the learning can be formulated as a quadratic programming problem. Inspired by [13], we use a 1-slack cutting-plane training method instead of the n-slack method [13] used. The 1-slack algorithm is substantially faster than n-slack algorithm on multi-class classification by several orders of magnitude [9]. In addition, we employ the FastPD algorithm [8] instead of the alpha-expansion graph cut in [14] as the final energy optimization method. It can be proved that FastPD is as powerful as alpha-expansion, in the sense that it computes exactly the same solution, but with a substantial speedup of a magnitude ten over existing techniques. Moreover, contrary to alpha-expansion, the derived algorithms generate solutions with guaranteed optimality properties for a much wider class of problems, even for non-metric potentials [8].

4. EXPERIMENTAL RESULTS

In this section we present our experimental results, and compare the performance of our method to recently published state-of-the-art results on four datasets: the 9-class and 21-class MSRC datasets; and the 7-class Sowerby and Corel datasets used in [1]. For all datasets, we randomly partition the images into balanced training and test data sets as done in [6].

Results for the 9-class MSRC database are shown in Table 1, our LRC/CRF classifier surpass the state of the art method slightly by 0.1% on this dataset. Table 2 gives the comparison of pixel-level accuracy with other algorithms on MSRC-21 class datasets. Using our two stage classifier LRC/CRF achieves 76.8% on five folds average (Note here our result obtained with a five folds average as done in [6], from 75.5% to 78.0%). Other works are only reported on a single fold.

Table 1. Comparison of pixel-level accuracy on MSRC-9(%)

Method	[15]	[4]	[16]	[5]	[6]	Proposed
Accuracy	82.3	84.9	86.7	87.2	88.5	88.6

Table 2. Comparison of pixel-level accuracy on MSRC-21(%)

Method	[17]	[3]	[15]	[6]	[18]	Proposed
Accuracy	72.2	72.2	73.5	76.5	77.7	76.8

One of the most fascinating parts of our algorithm is the speed of training and testing. Our algorithm runs on a 3.4 Ghz machine with 3.8 GB memory. The total training time and test time per image are listed in Table 3. The training time with max-margin learning is about $30 - 35$ minutes, the testing time per image is less than 0.02 second by applying FastPD as inference algorithm. Note that the training time in Table 3 does not contain the time consume on feature extraction and codebook formation.

Table 4 shows a comparison of semantic segmentation results on Sowerby-7 and Corel-7 datasets. The accuracies within the ten randomly partition tests on Sowerby-7 are from

Table 4. Comparison of pixel level labeling accuracy(%) to other algorithms on the Sowerby and Corel datasets

Algorithm	Sowerby			Corel		
	Accuracy	Training time	Test time	Accuracy	Training time	Test time
Shotton et al. [3]	88.6	5h	10s	74.6	12h	30s
He et al. [1]	89.5	Gibbs	Gibbs	80.0	Gibbs	Gibbs
Verbeek et al. [4]	87.4	20min	5s	74.6	15min	3s
Toyoda et al. [7]	90.0	–	–	83.0	–	–
Gould et al. [6]	87.5	–	–	77.3	–	–
Our LRC/CRF	89.1	7∼8min	<0.02s	77.0	6∼7min	<0.02s

Table 3. Comparison of computation speed on MSRC-21

Method	Training time	Test time
TextonBoost [3]	2 days	30 sec/image
PLSA-MRF [15]	1 hour	2 sec/image
STF-ILP [17]	2 hours	<0.125 sec/image
AC(ACP) [18]	a few days	30∼70 sec/image
Proposed	30∼35 min	<0.02 sec/image

86.8% to 91.2%, and from 71.5% to 81.3% for Corel-7 , respectively.

5. CONCLUSION

We have presented a fast high performance approach to semantic image segmentation that incorporate the local, regional and global features. One limitation of our approach is that the images are represented as rectangular patches at a single scale which could not capture many classes whose appearance and co-occurrence varies significantly with scale. One way to capture this would be to learn separate topic models for the patch appearances or label mixtures at each scale, and use these as features. We intend to explore this in future work.

6. REFERENCES

[1] X. He, R.S. Zemel, and M. Á. Carreira-Perpinan, “Multiscale conditional random fields for image labeling,” in *CVPR*, 2004, vol. II, pp. 695–702.

[2] S. Kumar and M. Hebert, “A hierarchical field framework for unified context-based classification,” in *ICCV*, 2005, vol. II, pp. 1284–1291.

[3] J. Shotton, C. Rother J. Winnand, and A. Criminisi, “Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation,” *ECCV*, vol. I, pp. 1–15, 2006.

[4] J. Verbeek and B. Triggs, “Scene segmentation with crf learned from partially labeled images,” *NIPS*, pp. 1553–1560, 2008.

[5] F. Schroff, A. Criminisi, and A. Zisserman, “Object class segmentation using random forests,” in *BMVC*, 2008.

[6] S. Gould, J. Rodgers, D. Cohen, G. Elidan, and D. Koller, “Multi-class segmentation with relative location prior,” *IJCV*, vol. 80, pp. 300–316, March 2008.

[7] T. Toyoda and O. Hasegawa, “Random field model for integration of local information and global information,” *TPAMI*, vol. 30, pp. 1483–1480, August 2008.

[8] N. Komodakis and G. Tziritas, “Approximate labeling via graph-cuts based on linear programming,” *TPAMI*, vol. 29, pp. 1436–1453, August 2007.

[9] T. Joachims, T. Finley, and C. N. Yu, “Cutting-plane training of structural svms,” *Machine Learning*, vol. 76, January 2009.

[10] J. van de Weijer and C. Schmid, “Coloring local feature extraction,” in *ECCV*, 2006, vol. II, pp. 334–348.

[11] S. Lazebnik and M. Raginsky, “An empirical bayes approach to contextual region classification,” in *CVPR*, 2009.

[12] C. Galleguillos, A. Rabinovich, and S. Belongie, “Object categorization using co-occurrence, location and appearance,” in *CVPR*, 2008.

[13] M. Szummer, P. Kohli, and D. Hoiem, “Learning crf using graph cuts,” in *ECCV*, 2008.

[14] Y. Boykov, O. Veksler, and R. Zabih, “Fast approximate energy minimization via graph cuts,” *TPAMI*, vol. 23, pp. 1222–1239, November 2001.

[15] J. Verbeek and B. Triggs, “Region classification with markov field aspect models,” in *CVPR*, 2007.

[16] X. He and R. S. Zemel, “Learning hybrid models for image annotation with partially labeled data,” *NIPS*, 2008.

[17] J. Shotton, M. Johnson, and R. Cipolla, “Semantic texton forests for image categorization and segmentation,” in *CVPR*, 2008.

[18] Z. W. Tu, “Auto-context and its application to high-level vision tasks,” in *CVPR*, 2008.