IMAGE LABELING BY MULTIPLE SEGMENTATION

Quan Zhou, Canxiang Yan, Yingying Zhu, Xiang Bai and Wenyu Liu

Huazhong University of Science and Technology
Department of Electronics and Information Engineering
Wuhan, China PR

ABSTRACT

In this paper, we provide a method for image labeling by combining the local features and contextual cues in a multiple segmentation framework. Our main insight is to weight the classification results of each image region in different levels, which are obtained by a series of learned discriminative models based on bag of features. The contextual cues are implicitly embedded as feature selection in learning process. Multiple segmentation framework provides robust representation, allowing a wide variety of cues to contribute to the confidence in each semantic label. Our algorithm has been applied on the lotus hill institute(LHI) 15-class dataset and outperforms other state-of-the-art methods.

Index Terms— Image labeling, segmentation, feature selection, classification

1. INTRODUCTION

Image labeling has been gained great concern in recent research [1, 2, 3, 13]. It provides a framework to understand the nature images by assigning a semantic label for each pixel to achieve the object recognition and accurate segmentation task. To overcome the local ambiguity for recognition, researchers have explored various contextual information in the recent literature.

The popular conditional random filed (CRF) models [4] have been widely used in recent years with two components formulated in an energy function: (a) A local data term encoding pixel-based or superpixel-based classification results (i.e. labels) [1, 2]. (b) Some pairwise relation terms expressing local or long-range context between labels such as co-occurrence [2, 3, 5], geometric context [6] and global scene template [7]. Despite their success, these CRF-based methods only represent objects implicitly through the context defined by the label co-occurrence statistics of neighbouring pixels without considering semantic information provided by multiple segmentation. They are often failure when the appearance of objects has large variations or exhibits great inter-class similarities.

This work was supported by NSFC 60873127 and 60903096.

In this paper, we propose to implicitly use contextual relation based on multi-level segmentation and adaptively weight the recognizable results in different levels to get final labeling output. Fig. 1 shows an illustration of our approach. We advocate using superpixels for image representation which adds spatial grouping to the discriminative recognition model and provides an intuitive representation for contextual based intersection. With multiple segmentation in hand, the popular bag-of-features(BoF) model [8] is employed for labeling each segment. The final labels for superpixels are voted by linear combination of multiple labeling and the weights are automatically learned from the boosting algorithm.

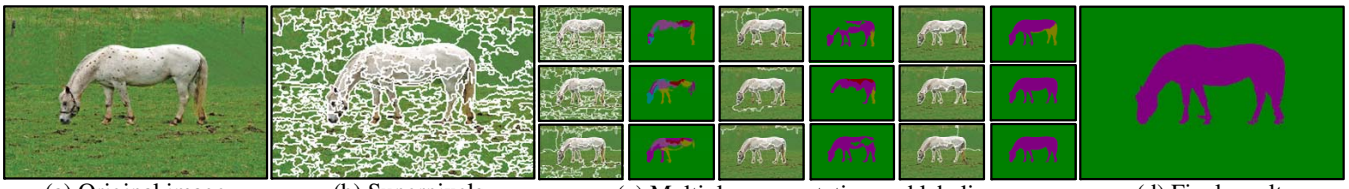
2. REPRESENTATION

2.1. Superpixels

Traditionally, an image is represented by a two-dimensional array of RGB pixels. With no knowledge of how to group these pixels, we can compute only local cues, such as pixel colors or responses convoluted with bank of filters. Our first step is to form superpixels (as shown in Fig. 1(b)) from those raw pixel intensities by using the oversegmentation technique [9]. Although those superpixels tend to be highly irregular in size and shape, the advantage of this technique is that it can often group large homogeneous regions while dividing heterogeneous regions into many smaller superpixels. This often allows reasonable oversegmentations with fewer superpixels (typically around 300 for a 320×210 image). Intuitively, using superpixels can improve the computational efficiency of our algorithm and allow more complex statistics (such as BoF) to be calculated for labeling.

2.2. Multiple Segmentations

Fig.1(c) shows that, while the increased size of superpixels provides much better classification performance for foreground object (e.g. "cow", "horse" and "car"). Larger regions always provide more effectively complex cues since they explore additional contextual information than smaller ones. Our approach is to compute multiple segmentations



(a) Original image (b) Superpixels (c) Multiple segmentation and labeling (d) Final results

Fig. 1. Illumination of our framework. (best viewed in color)

based on simple basic and local cues and then use the recognizable results provided by each segment to better evaluate final labeling output. It is untractable to evaluate all possible segmentations of an image, therefore, we sample a small level of segmentations that are representative of the entire distribution. Method proposed in [9] guarantees all segments are pure (i.e. pixels inside a segment should have the same semantic label) and the careful selected granularity parameters ensure the high level segments always contain the lowest level superpixels.

3. IMAGE LABELING MODEL

3.1. Problem Formulation

Let Λ be the image lattice (e.g. $W \times H$ pixels) and I_Λ an input image defined on Λ. For I_Λ, suppose there are M superpixels denoted by $\Lambda_1, \cdots, \Lambda_M$, each of which may take a value from the set of labels: $l \in \{1, \cdots, L\}$. Any possible assignment of labels to the random variables will be called a labeling (denote by $\mathbf{x}$) which takes values from $\mathcal{L} = L^M$. Our objective is to compute $\mathbf{x}^*$ that maximizes a likelihood probability,

$$\mathbf{x}^* = \arg\max_{\mathbf{x}\in\mathcal{L}} p(\mathbf{x}|I_\Lambda) \tag{1}$$

Due to the non-overlapping of Λ_i's, the image model $p(\mathbf{x}|I_\Lambda)$ is assumed to be independent conditioning on $\mathbf{x}$, then the likelihood is further factorized onto superpixels. We have,

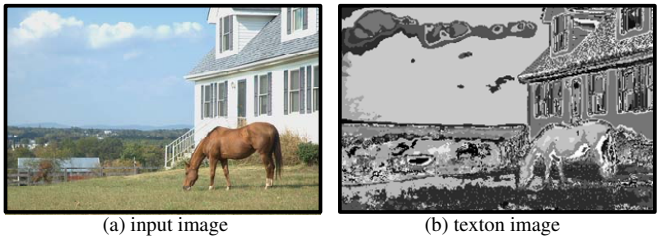
$$p(\mathbf{x}|I_\Lambda) = \prod_{i=1}^{M} p(y_i = l|I_{\Lambda_i}) \tag{2}$$

To get the label likelihood for the i^{th} superpixel, we simply marginalize over the label of sampled segments $\{s_j\}$ that contain this superpixel in all level:

$$p(y_i = l|I_{\Lambda_i}) \propto \sum_{\Lambda_i \subset s_j} \sum_{l'=1}^{L} p(y_j = l'|I_{s_j}) p(y_i = l|y_j = l') \tag{3}$$

where $p(y_j = l'|I_{s_j})$ denotes the recognizable probability when the label of s_j is l'. Eqn.(3) indicates the final label output of i^{th} superpixel is determined by the linear combination of recognizable results from all segments which contain i^{th} superpixel. Note the sum is over the segments that contain the i^{th} superpixel, rather than all the potential segmentations. We believe that small level of segmentations (denoted by S) will provide a sufficient sample. In the following, we entail the associated classifiers and features for $p(y_j = l'|I_{s_j})$ and $p(y_i = l|y_j = l')$, respectively.

3.2. Features



(a) input image (b) texton image

Fig. 2. Illumination of texton image. (see texts for details)

For learning $p(y_j = l'|I_{s_j})$, we advocate using BoF features. Given an input training image as shown in Fig.2(a), we first use a set of band-pass filters adopted in [1] to map each pixel to a high dimensional feature space. These features are densely abstracted from original training images and clustered to T centers by K-means algorithm [10]. We call these cluster centers as **textons** and the pixels with similar appearance cues are grouped with the same texton index. This generates the texton images (as shown in Fig.2(b)) which achieves a pre-segmentation results for nature images. Given multiple segmentation in hand, we count the frequency of textons to encode segments and thus each of they is represented by a T-dimensional vector.

3.3. Classifiers

In this paper, we use boosted decision trees for learning $p(y_j = l'|I_{s_j})$ and the weights of $p(y_i = l|y_j = l')$ with the logistic regression version of Adaboost [11]. Decision trees make good weak learners, since they provide automatic feature selection and limited modeling of the joint statistics of data. Each decision tree provides a partitioning of the data and outputs a confidence-weighted decision which is the class-conditional log-likelihood ratio for the current weighted distribution. The logistic regression version of Adaboost differs from the original confidence weighted version by only a

slight change in the weight update rule, but it results in confidence outputs that tend to be well-calibrated probabilities (after applying the simple sigmoid conversion to the log-ratio output). We train separate classifiers in a one vs. all fashion, for instance, to distinguish with "tree" class, we train the classifiers that estimate the probability of a segment being the remaining semantic labels. These are then normalized to ensure that the estimated probabilities sum to one. The whole training algorithm is given in Algorithm 1.

Algorithm 1: Training boosted decision trees

Input: $D_1, \cdots, D_m$: training data;$\omega_1^1, \cdots, \omega_m^1$: initial weights;$y_1, \cdots, y_m \in \{1, -1\}$: labels; n: number of nodes per decision trees; N: number of iteration

Output: $T_1, \cdots, T_N$: decision trees; $f_1^1, \cdots, f_N^n$: weighted log-ratio for each node of each tree

1 **for** *t = 1 to N* **do**

2 Learn n-node decision tree T_t based on weighted distribution $\mathbf{w}_t$.

3 Assign to each node T_t^k: $f_t^k = \frac{1}{2}\log\frac{\sum_{i:y_i=1, D_i\in T_t^k}\omega_i^t}{\sum_{i:y_i=-1, D_i\in T_t^k}\omega_i^t}$.

4 Update weights: $\omega_i^{t+1} = \frac{1}{1+\exp(y_i\sum_{t'=1}^{t} f_{t'}^{k_{t'}})}$ with $k_{t'} : D_i \in T_{t'}^{k_{t'}}$.

5 Normalize weights so that $\sum_i \omega_i^{t+1} = 1$.

6 **end**

4. EXPERIMENTS

In this section, we evaluate the performance of our algorithm on LHI 15-class dataset [12] which consists of 370 images including 15 classes that include: building, tree, grass, water, sky, mountain, road, car, horse, cow, sheep, elephant, motorbike, rhinoceros and airplane. They are randomly split into roughly 40% for training, 10% for evaluation and 50% for testing. Empirically, we select $T = 500$ and $S = 9$ in our experiments. We first present the overall results compared with other approaches and then analysis some aspects of our algorithm in details.

Overall results. Fig.3 illustrates the confusion matrices obtained by applying our algorithm and other state-of-the-art methods to the test set. The percentage of image pixels assigned to the correct class label are scared from blue to red. Tab.1 shows the overall results on the testing set of LHI 15-class and it demonstrates our algorithm outperforms the state-of-the-art methods both in pixel-wise accuracy and implemental efficiency. In Fig.4, we also illustrate some labeling results for nature photograph. Our algorithm can handle large variations of view-point and scale of foreground objects and large variations in the appearance of background regions. On the last column, we show two examples in which the labeling are not good enough ("tree" labeled as "building" and "mountain" misclassified as "tree").

Table 1. Overall pixel-wise accuracy and implement time on the LHI 15-class.

Methods	LHI 15-class	time(min)
Our results	**80.47%**	**19.28**
Auto-Context [2]	76.52%	24.67
Multi-class segmentation [3]	74.19%	23.91
DecomposingScene [13]	71.08%	142.07
BoF [8]	68.73%	58.26
TextonBoost [1]	62.70%	37.80

Effects of texton number and segmentation level. In our experiment, two factors affecting the performance are the number of textons and level of segmentations. Fig.5(a) plots of accuracy curve v.s. texton number. We observe that after the number of texton increases greater than 500, the accuracy is not improved much. For the level of segmentations, empirically, we use 500 textons in our experiments. Fig.5(b) shows the plot of accuracy v.s. segmentation level. We also discover that there is no significant effect on performance when the level of segmentation is greater than 9.

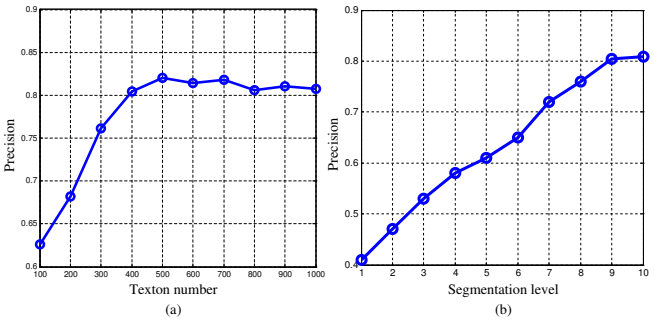


Fig. 5. Effects of the number of Textons and multiple segmentations. (see texts for details)

5. CONCLUSION

In this paper, we presented a multiple segmentation framework for image labeling. The linear combination of multiple labeling results can efficiently explore local and contextual cues in a unified framework. The experimental results demonstrate our algorithm outperforms the state-of-the-art methods in terms of the pixel-level accuracy of the labeling and segmentation on the LHI 15-class dataset. Our future work will address the extensions including high-level semantic contextual constraints and incorporation of object detection as supplementary cues.

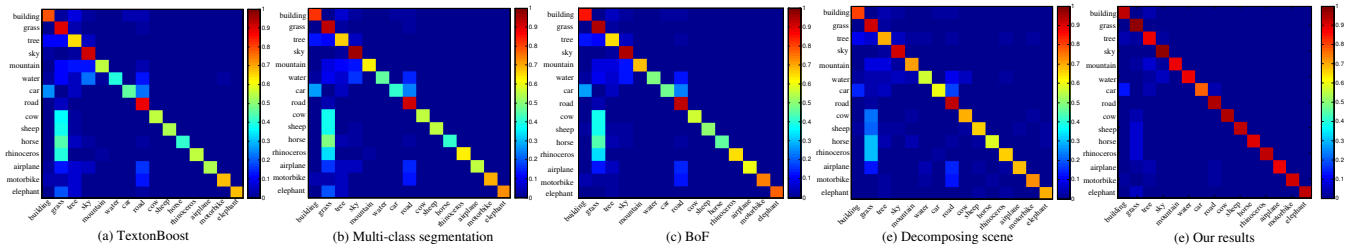


Fig. 3. Comparisons of the confusion matrices. From (a) to (e), the confusion matrix is evaluated by using Textonboost [1], Multi-class segmentation [3], BoF [8], Decomposing scenes [13] and ours. (best viewed in color)

Fig. 4. Illumination of some labeling results. For clarity, textual labels have also been superimposed on the resulting segmentations and different color denotes different category. (best viewed in color)

6. REFERENCES

[1] J. Shotton, J.W. Winn, C. Rother, and A. Criminisi, "Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context," *IJCV*, vol. 81, no. 1, pp. 2–23, 2009.

[2] Z. W. Tu and X. Bai, "Auto-context and its application to high-level vision tasks and 3d brain image segmentation," *IEEE Trans. PAMI*, vol. 32, no. 10, pp. 1744–1757, 2010.

[3] S. Gould, J. Rodgers, D. Cohen, G. Elidan, and D. Koller, "Multi-class segmentation with relative location prior," *IJCV*, vol. 80, no. 3, pp. 1239–1253, 2008.

[4] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: probabilistic models for segmenting and labeling sequence data," in *ICML*, 2001, pp. 282–289.

[5] C. Galleguillos, A. Rabinovich, and S. Belongie, "Object categorization using co-occurrence, location and appearance," in *CVPR*. IEEE, 2008, pp. 1–8.

[6] D. Hoiem, A.A. Efros, and M. Hebert, "Geometric context from a single image," in *ICCV*. IEEE, 2005, pp. 654–661.

[7] X.M. He, R.S. Zemel, and D. Ray, "Learning and incorporating top-down cues in image segmentation," in *ECCV*. Springer, 2006, pp. 338–351.

[8] L. Yang, P. Meer, and D.J. Foran, "Multiple class segmentation using a unified framework over mean-shift patches," in *CVPR*. IEEE, 2007, pp. 1–8.

[9] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Trans. PAMI*, vol. 24, no. 5, pp. 603–619, 2002.

[10] C. Elkan, "Using the triangle inequality to accelerate k-means," in *ICML*, 2003, pp. 147–153.

[11] M. Collins, R. Schapire, and Y. Singer, "Logistic regression, adaboost and bregman distances.," *Machine Learning*, vol. 48, no. 1-3, pp. 253–285, 2002.

[12] B. Yao, X. Yang, and S.C. Zhu, "Introduction to a large scale general purpose groundtruth dataset: Methodology, annotation tool, and benchmark," in *EMMCVPR*, 2007.

[13] S. Gould, R. Fulton, and D. Koller, "Decomposing a scene into geometric and semantically consistent regions," in *ICCV*. IEEE, 2009, pp. 1–8.