

# Exclusive Constrained Discriminative Learning for Weakly-Supervised Semantic Segmentation

Peng Ying, Jing Liu, Hanqing Lu and Songde Ma  
The National Laboratory of Pattern Recognition, Institute of Automation,  
Chinese Academy of Sciences, Beijing, China  
{peng.ying, jliu, luhq}@nlpr.ia.ac.cn, masd@most.cn

## ABSTRACT

How to import image-level labels as weak supervision to direct the region-level labeling task is the core task of weakly-supervised semantic segmentation. In this paper, we focus on designing an effective but simple weakly-supervised constraint, and propose an exclusive constrained discriminative learning model for image semantic segmentation. To be specific, we employ a discriminative linear regression model to assign subsets of superpixels with different labels. During the assignment, we construct an exclusive weakly-supervised constraint term to suppress the labeling responses of each superpixel on the labels outside its parent image-level label set. Besides, a spectral smoothing term is integrated to encourage that both visually and semantically similar superpixels have similar labels. Combining these terms, we formulate the problem as a convex objective function, which can be easily optimized via alternative iterations. Extensive experiments on MSRC-21 and LabelMe datasets demonstrate the effectiveness of the proposed model.

## Categories and Subject Descriptors

H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing; I.2.10 [Artificial Intelligence]: [Vision and Scene Understanding]

## General Terms

Algorithms; Experimentation; Theory

## Keywords

Semantic Segmentation; Weak Supervision

## 1. INTRODUCTION

Image semantic segmentation aims to collaboratively perform image segmentation and region-level label assignment. Simply, it is a task of region-level annotation. Semantic segmentation is a challenging and fundamental task in computer vision. An effective image semantic segmentation can

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.  
MM'15, October 26–30, 2015, Brisbane, Australia.  
© 2015 ACM. ISBN 978-1-4503-3459-4/15/10 ...\$15.00.  
DOI: <http://dx.doi.org/10.1145/2733373.2806329>.

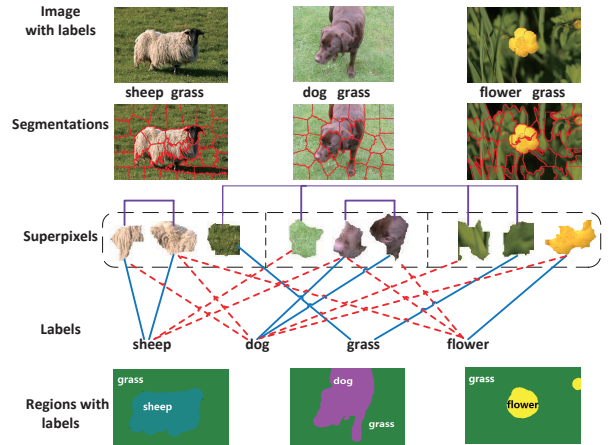


Figure 1: Overview of our model. The superpixels connected by purple lines are both visually and semantically similar. The mappings from superpixels to labels, denoted by red dash lines, are suppressed according to the exclusive constraint, while the mapping denoted by blue lines are expected.

facilitate many other vision tasks, such as scene understanding, object detection and region-based image retrieval.

Some typical image semantic segmentation methods are to train models on pixel-level annotated data [3, 5]. However, the obtainment of these pixel-level annotation costs much manpower and time. Fortunately, with the flourish of social media and the success of photo-sharing websites e.g. Flickr and Picasa, lots of images with user-contributed image-level annotation are easily to obtain. The online weakly labeled data allow us to deal semantic segmentation with weak supervision methods [16, 17, 15, 7]. For example, Liu et. al proposed to handle the image semantic segmentation problem with a weakly-supervised graph based algorithm [6]. In this algorithm, the process of label propagation from images to regions is bounded by weak supervision information. Y. Liu et. al proposed a weakly-supervised dual clustering approach to cluster the over-segmented regions and assign image-level tags to those regions simultaneously[8]. For the above weakly supervised approaches, they adopted the idea of multiple instance learning (MIL) to transfer the image-level labels to regions, i.e., the MIL-based models encourage maximum labeling responses of regions to be aligned with image-level labels. However, such a maximum optimization problem is non-convex and difficult to solve.

From the aforementioned approaches, we find the essence of weakly-supervised semantic segmentation is to learn the mapping from superpixels to labels while the co-occurrence of the labels and superpixels are also explored. Therefore, our work is motivated to find a well-optimized but effective form of weakly-supervised constraint to assist the labeling mapping for each superpixel in semantic segmentation task.

In this paper, we propose a novel discriminative classification with exclusive weakly-supervised constraint to perform superpixel classification and label propagation collaboratively. As shown in Fig. 1, we first over-segment images to obtain some image patches, i.e., superpixels, as the inputs of our model. Then, we employ a discriminative linear regression model to assign some subsets of superpixels with different labels. During the assignment, we expect the labeling responses of each superpixel are suppressed on the labels outside its parent image-level label set, and an exclusive weakly-supervised constraint term is formulated to implicitly import the image-level labels as weak supervision. Besides, a spectral smoothing term is integrated to encourage that both visually and semantically similar superpixels have similar labels, while the visual features and the labeling information of superpixels are jointly considered to construct the similarity graph. Combining these terms, we formulate the problem as a convex objective function, which can be easily optimized via alternative iterations. Experimental results on two public datasets, MSRC and LabelMe, demonstrate that the proposed model outperforms some state-of-the-art methods.

## 2. THE PROPOSED MODEL

### 2.1 Problem Definition

Suppose we have a data collection  $\chi = \{X_i\}_{i=1}^I$ , where  $X_i$  indicates the  $i$ -th image. The images are from  $C$  classes and the image-level label information for all images is also given. That label information is represented symbolically by  $G = [g_1, \dots, g_i, \dots, g_I]$ , where  $g_i = [g_i^1, \dots, g_i^c, \dots, g_i^C] \in \{0, 1\}^C$  is the image-level label information for  $X_i$  and  $g_i^c = 1$  if  $X_i$  has the  $c$ -th label and otherwise  $g_i^c = 0$ . After over-segmented,  $X_i$  are divided into a set of superpixels  $\{x_{i1}, \dots, x_{in_i}\}$ , where  $x_{ik}$  indicates the  $k$ -th superpixel in the  $i$ -th image and its candidate label set is  $g_i$ . At superpixel-level, the data collection  $\chi$  is also denoted by  $\{x_1, \dots, x_i, \dots, x_N\}$ , where  $N = \sum_{i=1}^I n_i$ . The exact label information for the superpixel  $x_i$  is marked as  $y_i = [y_i^1, \dots, y_i^c, \dots, y_i^C] \in \{0, 1\}^C$ . In weakly-supervised semantic segmentation,  $Y = [y_1, \dots, y_n, \dots, y_N] \in \{0, 1\}^C$  is to be inferred with image label information  $G$  given.

### 2.2 Weakly-Supervised Discriminative learning

To deal with semantic segmentation task, we introduce the linear regression model with weak supervision constraint to mine the co-occurrence of superpixel-label pairs.

The linear regression is employed to model discriminative classification, which is formulated as

$$T_1 = \alpha \|X^T W - Y\|_F^2 + \beta \|W\|_F^2, \quad (1)$$

where  $\|W\|_F^2$  is a regularization term and  $\alpha$  and  $\beta$  are weight coefficients.

The mapping from superpixels to labels should be bounded by weak supervision information. In other words, label propagation should meet the constraint that the assigned label of the arbitrary superpixel should belong to the image-level label set. Besides, one superpixel can have and only one label. However, in consideration of the inevitable existence of noisy tags in training data for practical application, the constraint that one label has at least one superpixel mapped to it, is not required. In other words, non-image-level label should be suppressed [4]. To model the limitation on label propagation, we propose exclusive weakly-supervised constraint term:

$$\begin{aligned} T_2 &= \text{Tr}[Y(P - \underbrace{[g_1, \dots, g_1, \dots, g_i, \dots, g_i, \dots, g_I, \dots, g_I]}_{n_1 \quad n_i \quad n_I})] \\ &= \text{Tr}[Y M^T], \end{aligned} \quad (2)$$

where  $P \in \{1\}^{C \times N}$ ,  $M$  is a binaryzation penalty matrix in which punishment or impunity is denoted by 1 or 0 respectively. If a superpixel  $x_{ij}$  is given a label outside its candidate set  $g_i$ , it will be punished; If  $x_{ij}$  maps to a label which belongs to  $g_i$ , it will not be punished. Moreover, the noisy tags in  $g_i$  are not required to have superpixels mapped to it. So the weakly-supervised discriminative classification is defined as minimizing the following equation:

$$\mathcal{F} = \alpha \|X^T W - Y\|_F^2 + \beta \|W\|_F^2 + \gamma \text{Tr}[Y M^T] \quad (3)$$

It's worth mentioning that the proposed exclusive weakly-supervised constraint term is a convex function and can be easily optimized.

### 2.3 Spectral Smoothing with Semantic Graph

The employment of spectral smoothing is based on the observation that the adjacent superpixels in feature space have good chance to share the same label. We adopt Euclidean distance to measure the similarity of superpixels. What's more, semantic information is also embed into the construction of the affinity graph, e.g. k-NN semantic graph. Unlike k-NN original graph, the k-nearest neighbors in k-NN semantic graph should be not only similar in feature space but also semantic relevant. For a specific image  $X_m$ , its image-level label set is denoted as  $g_m = \{l_1, l_2, \dots, l_n\}$ , where  $n$  is the total numbers of labels for the image  $X_m$ . For a specific label  $l_j \in g_m$  and superpixel  $x_{mi}$  from the  $m$ -th image, we select the  $k$ -NN superpixels of  $x_{mi}$  from the images with the specific label  $l_j$  and the selected superpixels corresponding to  $l_j$  is denoted as a set  $K_j^{mi}$ . Then we compute the average similarity  $A_j^{mi}$  between  $x_{mi}$  and the elements in  $K_j^{mi}$ :

$$A_j^{mi} = \sum_{a=1}^k \text{Sim}(x_{mi}, x_{ja}) \quad \text{s.t. } x_{ja} \in K_j^{mi}, \quad (4)$$

where  $\text{Sim}(x_1, x_2) = \exp(-\|x_1 - x_2\|^2/t)$ ,  $t$  is a parameter controlling decay rate. The optimal  $k$ -NN superpixels  $K^{mi}$  of  $x_{mi}$  is decided by the following equation:

$$K^{mi} = \underset{K_j^{mi}}{\text{argmax}} \{A_1^{mi}, \dots, A_j^{mi}, \dots, A_n^{mi}\} \quad (5)$$

The similarity between superpixels are formulated as a similarity matrix  $S$

$$S_{pq} = \begin{cases} \exp(-\|x_p - x_q\|^2/t) & x_q \in K^p \text{ or } x_p \in K^q \\ 0 & \text{otherwise.} \end{cases} \quad (6)$$

Accordingly, the spectral smoothing term can be formulated as follows:

$$T_3 = \frac{1}{2} \sum_{p,q=1}^N S_{pq} \left\| \frac{y_p}{\sqrt{A_{pp}}} - \frac{y_q}{\sqrt{A_{qq}}} \right\|_2^2 = \text{Tr}[Y^T L Y], \quad (7)$$

where  $A = \sum_{p,q=1}^N S_{pq}$ , and  $L = A^{-1/2}(A - S)A^{-1/2}$ .

## 2.4 Unified Objective Function

Combining the aforementioned items, the problem is formulated as a non-convex objective function as shown below:

$$\begin{aligned} \min_{Y,W} \mathcal{L} = & \alpha \left\| X^T W - Y \right\|_F^2 + \beta \|W\|_F^2 + \gamma \text{Tr}[Y M^T] \\ & + \text{Tr}[Y^T L Y] \\ \text{s.t. } & Y^T Y = I_C, Y \geq 0. \end{aligned} \quad (8)$$

where  $\alpha, \beta, \gamma$  are parameters controlling the weight of the three terms.

## 2.5 Optimization

We put the orthogonality constraint  $\|Y^T Y - I_C\|_F^2$  into the objective function. So Eq. 5 is rewritten as:

$$\begin{aligned} \min_{Y,W} \mathcal{L} = & \alpha \left\| X^T W - Y \right\|_F^2 + \beta \|W\|_F^2 + \gamma \text{Tr}[Y M^T] \\ & + \text{Tr}[Y^T L Y] + \frac{\mu}{2} \|Y^T Y - I_C\|_F^2 \\ \text{s.t. } & Y \geq 0. \end{aligned} \quad (9)$$

To solve the optimization problem, we propose an iterative optimization algorithm like in [10]. At  $r^{\text{th}}$  iteration:

1. Given  $Y$ , optimizing  $W$ : Fixing  $Y$ , let  $\partial \mathcal{L} / \partial W = 0$  and we obtain

$$W = \alpha(\alpha X X^T + \beta I)^{-1} X Y \quad (10)$$

2. Given  $W$ , optimizing  $Y$ : Fixing  $W$ , let  $\partial \mathcal{L} / \partial Y = 0$  and we have

$$Y_{ij} \leftarrow Y_{ij} \frac{2(\mu Y)_{ij}}{(2PY + \gamma M + 2\mu Y Y^T Y)_{ij}} \quad (11)$$

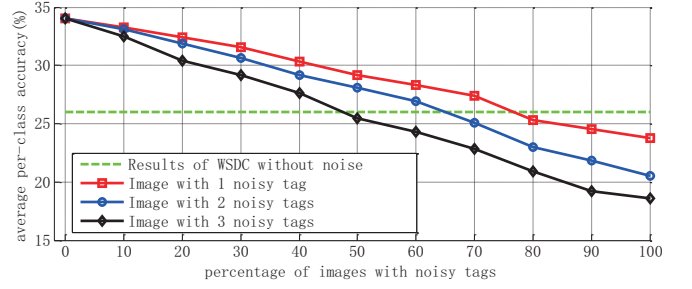
where

$$P = L + \alpha(I_N - \alpha(X X^T + \beta I)^{-1} X) \quad (12)$$

Repeat step 1 to 2 until the objective function converges to a stable state.

## 3. EXPERIMENTS

We test our model on two frequently-used data in semantic segmentation field, e.g. MSRC-21 and LabelMe Outdoor(LMO). To evaluate the effectiveness, we compare the proposed model with several related works, including weakly-supervised approaches WSDC, MIM, MRF, MTL, GMIM



**Figure 2: Average per-class accuracy with different percentage of images with 1, 2 or 3 noisy tags on LabelMe dataset.**

and fully-supervised approaches HCRF, Texboost, THSR, Supix, LT. Methods are compared by computing the average per-class accuracy (percentage of pixels with agreement between the assigned label and groundtruth for a class, averaged on the whole classes).

Besides, to evaluate the robustness of our model, we conduct experiments on LabelMe in different noise polluted variety.

### 3.1 Experiments on MSRC-21

MSRC-21 dataset is widely used in image semantic segmentation field. There are 591 images with pixel-level ground truths in the dataset. The images are from 21 class and each image contains 3 labels on average. The dataset is standardly split into 276 images for training, 59 images for verification and 256 images for testing. The images are over-segmented into 40 superpixels per image on average via S-LIC algorithm[1]. After over-segmentation, we employ SIFT as local descriptor and the standard bag-of-words model as representation of the superpixels. Besides, the parameters in formulation (9) are set as follows:  $\alpha = 10000$ ,  $\beta = 10000$ ,  $\gamma = 100000$ ,  $\mu = 10$ .

Table 2 shows the experimental results of our approach and other comparison methods. From the experimental results, we have the following observations: (1) Our approach outperforms all the listed weakly-supervised methods and is close to some fully-supervised methods like HCRF. This clearly validate the effectiveness of our approach. (2) By comparing the results of ours and WSDC, we can see the proposed approach has enhanced the semantic segmentation accuracy 6 percentage points, which is largely due to the exclusive weakly-supervised constraint.

### 3.2 Experiments on LabelMe Outdoor

LabelMe Outdoor is a more challenging dataset for semantic segmentation, which consists of 2688 images for 33 classes. Each image contains around 5 labels and has pixel-level ground truths. It is randomly split into 2488 images for training and 200 images for testing. The parameters in this dataset are set as follows:  $\alpha = 10000$ ,  $\beta = 100$ ,  $\gamma = 10000$ ,  $\mu = 100$ .

The experimental results of the proposed model and baseline algorithms are presented in Table 2. From the results, we can observe that our ECDL algorithm achieve the highest average per-class accuracy of 34 percent, which surpasses all the listed comparison methods. Specifically, compared with WSDC, our approach improves the accuracy by 8 percentage

| Method      | Texboost [11] | HCRF [2] | MTL [15] | MRF [14] | MIM [16] | WSDC [8] | ECDL      |
|-------------|---------------|----------|----------|----------|----------|----------|-----------|
| Supervision | FS            | FS       | WS       | WS       | WS       | WS       | WS        |
| Accuracy    | 58            | 75       | 37       | 50       | 67       | 68       | <b>74</b> |

**Table 1: Semantic segmentation results on MSRC dataset. WS denotes weak supervision, FS denotes full supervision.**

| Method      | Texboost [12] | LT [5] | Supix [13] | THSR [9] | MIM [16] | GMIM [17] | WSDC [8] | ECDL      |
|-------------|---------------|--------|------------|----------|----------|-----------|----------|-----------|
| Supervision | FS            | FS     | FS         | FS       | WS       | WS        | WS       | WS        |
| Accuracy    | 13            | 24     | 29         | 32       | 14       | 21        | 26       | <b>34</b> |

**Table 2: Semantic segmentation results on LabelMe dataset. WS denotes weak supervision, FS denotes full supervision.**

points. The above comparisons show our method leads to a significant improvement over previous weakly-supervised approaches and even over some fully-supervised approaches.

To further explore the robustness of the proposed model to noisy tags, we perform experiments on LMO in different noise polluted variety. To be specific, the  $p \in \{10, 20, \dots, 100\}$  percentage images are randomly chosen and accordingly each chosen image is added  $N \in \{1, 2, 3\}$  randomly chosen labels as noise. The experimental results are shown in Figure 2. From Fig. 2, we can find the accuracy reposefully descends with the increase of noisy tag number, which demonstrates to some extent our method are robust to noisy tags. Furthermore, with 70 percent images added one random noisy tag, our method can still achieve 27 percent accuracy, which is better than other weakly-supervised methods.

## 4. CONCLUSIONS

In this paper, we propose an exclusive weakly-supervised constraint with discriminative learning to collaboratively perform image segmentation and region-level annotation. We impose the exclusive constraint term on linear regression to construct a weakly-supervised discriminative classification model. To ensure that both visually and semantically similar superpixels have similar labels, a semantic graph based spectral smoothing term is integrated into the framework. Extensive experiments on two public datasets demonstrate the effectiveness of our approach.

## 5. ACKNOWLEDGEMENT

This work was supported by 863 Program (2014AA015104) and National Natural Science Foundation of China (61332016, 61272329, and 61472422).

## 6. REFERENCES

- [1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2274–2282, 2012.
- [2] L. Ladicky, C. Russell, P. Kohli, and P. H. Torr. Associative hierarchical crfs for object class image segmentation. In *ICCV, 2009*, pages 739–746, 2009.
- [3] L.-J. Li, R. Socher, and L. Fei-Fei. Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. In *CVPR 2009.*, pages 2036–2043. IEEE, 2009.
- [4] Y. Li, J. Liu, Y. Wang, H. Lu, and S. Ma. Weakly supervised rbm for semantic segmentation. In *IJCAI*, pages 1888–1894, 2015.
- [5] C. Liu, J. Yuen, and A. Torralba. Nonparametric scene parsing: Label transfer via dense scene alignment. In *CVPR 2009.*, pages 1972–1979. IEEE, 2009.
- [6] S. Liu, S. Yan, T. Zhang, C. Xu, J. Liu, and H. Lu. Weakly supervised graph propagation towards collective image parsing. *IEEE Transactions on Multimedia.*, 14(2):361–373, 2012.
- [7] X. Liu, S. Yan, J. Luo, J. Tang, Z. Huango, and H. Jin. Nonparametric label-to-region by search. In *CVPR, 2010*, pages 3320–3327. IEEE, 2010.
- [8] Y. Liu, J. Liu, Z. Li, J. Tang, and H. Lu. Weakly-supervised dual clustering for image semantic segmentation. In *CVPR, 2013*, pages 2075–2082, 2013.
- [9] H. Myeong and K. M. Lee. Tensor-based high-order semantic relation transfer for semantic scene segmentation. In *CVPR*, pages 3073–3080, 2013.
- [10] S. Qiu, J. Cheng, X. Zhang, B. Niu, and H. Lu. Community discovering guided cold-start recommendation: A discriminative approach. In *Multimedia and Expo (ICME), 2014 IEEE International Conference on*, pages 1–6. IEEE, 2014.
- [11] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *ECCV 2006*, pages 1–15. 2006.
- [12] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *IJCV*, 2009.
- [13] J. Tighe and S. Lazebnik. Superparsing: scalable nonparametric image parsing with superpixels. In *ECCV 2010*, pages 352–365. Springer, 2010.
- [14] J. Verbeek and B. Triggs. Region classification with markov field aspect models. In *CVPR, 2007*, pages 1–8.
- [15] A. Vezhnevets and J. M. Buhmann. Towards weakly supervised semantic segmentation by means of multiple instance and multitask learning. In *CVPR, 2010*, pages 3249–3256. IEEE, 2010.
- [16] A. Vezhnevets, V. Ferrari, and J. M. Buhmann. Weakly supervised semantic segmentation with a multi-image model. In *ICCV*, pages 643–650, 2011.
- [17] A. Vezhnevets, V. Ferrari, and J. M. Buhmann. Weakly supervised structured output learning for semantic segmentation. In *CVPR*, pages 845–852, 2012.