DICTIONARY LEARNING BASED SUPERPIXELS CLUSTERING FOR WEAKLY-SUPERVISED SEMANTIC SEGMENTATION

Peng Ying, Jing Liu, Hanqing Lu

National Laboratory of Patten Recognition, Institute of Automation,
Chinese Academy of Sciences, Beijing, China
{peng.ying, jliu, luhq}@nlpr.ia.ac.cn

ABSTRACT

The task of weakly-supervised semantic segmentation is solved by assigning image-level labels to over-segmented superpixels. Considering that superpixels are geometrically and semantically ambiguous for label assignment, we propose a joint solution of semantic segmentation to enhance the learnability of superpixels. First, our model includes a spectral clustering item and a discriminative clustering item to obtain some clustering subsets of superpixels (ideally semantic regions), which are more separable semantically than independent superpixels. Second, sparse coding based feature for superpixel is adopted to make the representation robust to noise, and the dictionary for the sparse representation is learned together with the above clustering items. Third, a weakly supervised item for superpixels, transferred from image-level labels, is attached. We jointly formulate the above problems as a non-convex objective function, and optimize it by the constraint concave-convex programming (CCCP) algorithm. Extensive experiments on MSRC-21 and LabelMe datasets prove the effectiveness of our approach.

Index Terms— Weak supervision, semantic segmentation, dictionary learning

1. INTRODUCTION

Image semantic segmentation aims to collaboratively perform image segmentation and tag alignment with those segmented regions. Simply, it is a task of region-level annotation. This task is challenging but significant since an effective image semantic segmentation method can benefit many high-level vision tasks such as region based image retrieval and fine-grained image analysis and synthesis.

Owing to the popularity of image sharing websites, e.g. Flickr, a great deal of images with user-contributed image-level labels are available, which can provide weak supervision for the task of semantic segmentation. Hence, in contrast with the tough requirement on pixel-level annotations from the fully supervised methods [1, 2, 3, 4], several weakly-supervised approaches [5, 6, 7] have been proposed and become popular. X. Liu et al. [8] proposed a bi-layer sparse coding model for uncovering how to reconstruct an image superpixel with the over-segmented patches of an image set, and then using the learned correlation to assign labels to the corresponding superpixels. S. Liu et al. [9] proposed a weakly-supervised graph propagation approach, by which the image-level labels can be propagated to those contextually derived semantic superpixels. The above methods attempt to learn a model on superpixel level. However, the superpixels are the products of over-segmention, and hence usually have the geometric and semantic ambiguities for semantic assignment. How to enhance the learnability of superpixels and further improve the performance of semantic segmentation becomes an important and challenging task. It is also the motivation of our work.

In this paper, we make the over-segmented superpixels more learnable from the following two sides of considerations. First, semantically separable subsets of superpixels are required for label prediction. Second, we expect to obtain a robust and discriminative feature representation for each superpixel. To address both the considerations in the task of semantic segmentation, we propose a novel weakly supervised solution by jointly employing superpixel clustering and sparse coding based feature representation. Inspired from the work in [10], a dual clustering model with weak supervision is adopted, in which a spectral clustering item is employed to gather visually similar superpixels into one cluster, and a discriminative clustering item is used to learn a classifier for label assignment to each cluster. For the discriminative clustering, a sparse coding feature is used to represent each superpixel, and the dictionary for the sparse representation is jointly learned during the dual clustering process. The above problems could benefit both of them because suitable feature representation promotes good clustering results and vice versa. In addition, a weakly supervised item for superpixels, transferred from image-level labels, is attached to the dual clustering model. Combining the above items, a non-convex optimization problem is formulated, which can be solved via the constraint concave-convex programming iteratively. Finally, extensive experiments on the public datasets, i.e., MSRC and LabelMe Outdoor, demonstrate the encouraging performance of our solution.

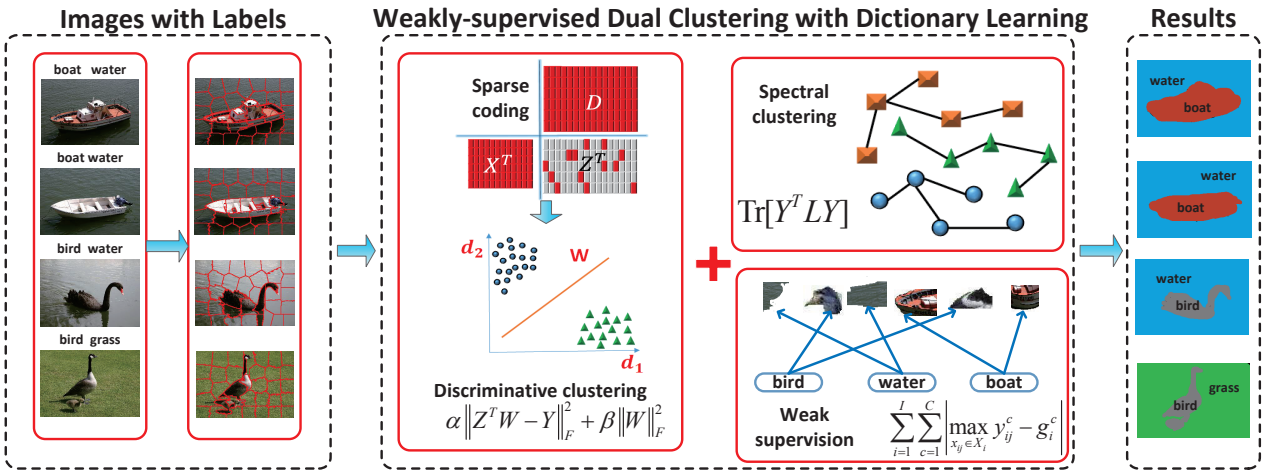

Fig. 1. The flowchart of our approach

2. APPROACH

2.1. Notations

Given an image collection $\chi = \{X_i\}_{i=1}^I$, where X_i represents the i-th image. The images are from C classes and their corresponding image-level label matrix is denoted as $G = [g_1, \ldots, g_i, \ldots, g_I]$, where $g_i = [g_i^1, \ldots, g_i^c, \ldots, g_i^C] \in \{0,1\}^C$ denotes the image-level label vector of X_i and $g_i^c = 1$ if X_i has the c-th label and otherwise $g_i^c = 0$. The images X_i are over-segmented into n_i superpixels such as $\{x_{i1}, \ldots, x_{in_i}\}$, where x_{ik} denotes the k-th superpixel in the i-th image and its candidate label set is g_i. So the data collection χ can also be marked as $\{x_1, \ldots, x_i, \ldots, x_N\}$, where $N = \sum_{i=1}^I n_i$. And its corresponding superpixel-level label matrix can be expressed as $Y = [y_1, \ldots, y_n, \ldots, y_N] \in \{0,1\}^C$, where $y_i = [y_i^1, \ldots, y_i^c, \ldots, y_i^C] \in \{0,1\}^C$ represents the label vector of the n-th superpixel.

2.2. Spectral Clustering

Since the visually similar superpixels are likely to belong to the same class, we employ a spectral clustering term to mine such relationship. Euclidean distance is adopted to measure the visually similar relationship between superpixels and thus we obtain the $k-$NN similarity graph S ($k = 50$). Let $A = \sum_{i,j=1}^N S_{ij}$, and then the Laplacian matrix can be denoted as $L = A^{-1/2}(A-S)A^{-1/2}$. The spectral clustering term can be formulated as follows:

$$T_1 = \frac{1}{2}\sum_{i,j=1}^{N} S_{ij}\|\frac{y_i}{\sqrt{A_{ii}}} - \frac{y_j}{\sqrt{A_{jj}}}\|_2^2 = \mathrm{Tr}[Y^TLY] \quad (1)$$

2.3. Discriminative Clustering with Dictionary Learning

One of the requirement for a good performance in semantic segmentation tasks is a good data representation. Sparse coding based feature representation can fit the requirement well.

In general, a sampled superpixel x_n may contain only a small part of information about the destination object. Hence, if only based on x_n, the judgement of the presence or absence of the destination object may be improper since the superpixel-level representations have geometric and semantic ambiguities . However, if we have a learned dictionary D that contains the most representative object parts (visual words), by sparse coding formulation, the discriminative visual information stored in D can divert into the reconstruction sparse vectors. The sparse coding term is defined as follows:

$$T_2 = \epsilon\|x - Dz\|_F^2 + \lambda\|z\|_1, \quad (2)$$

where z denotes sparse reconstruction vector corresponding to x and $Z = [z_1, \ldots, z_i, \ldots, z_N]$ denotes the sparse representation of the whole N superpixels which is more discriminative than the origin representation $[x_1, \ldots, x_i, \ldots, x_N]$. Besides, the number of atoms in D is set to 250.

Then, a linear mapping W from Z to the predicted labels Y is introduced as a classifier. The discriminative clustering term can be defined as minimizing the following equation:

$$T_3 = \alpha\|Z^TW - Y\|_F^2 + \beta\|W\|_F^2, \quad (3)$$

where $\|W\|_F^2$ is a regularization term.

Joint optimization of T_2 and T_3 enables to learn the classifier W and the dictionary D together. Given the classifier W, the formulation can be regarded as discriminative clustering supervised dictionary learning, whereas given the dictionary D, it can be regarded as discriminative clustering learning with sparse coding.

2.4. Weakly-Supervised Constraint

The weakly-supervised constraint is used to transfer image-level labels to superpixels, which is based on the three premises: (1) The label of the superpixel x_{ij} should belong to the label set of the image X_i. (2) One superpixel can have one and only one label. (3) Every label in the image-level label

set should have at least one superpixel corresponding to them in the image. The weakly-supervised constraint can be formulated as follows:

$$T_4 = \gamma \sum_{i=1}^{I} \sum_{c=1}^{C} | \max_{x_{ij} \in X_i} y_{ij}^c - g_{ic}| \\ s.t.\ Y^T Y = I_C,\ Y \geq 0, \tag{4}$$

where I_C is an unit matrix.

2.5. Unified Objective Function

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

$$\min_{Y,W,Z,D} \mathcal{L} = \text{Tr}\left[Y^T L Y\right] + \epsilon \|X - DZ\|_F^2 + \lambda \|z\|_1 \\ + \alpha \left\| Z^T W - Y \right\|_F^2 + \beta \|W\|_F^2 + \gamma \sum_{i=1}^{I} \sum_{c=1}^{C} | \max_{x_{ij} \in X_i} y_{ij}^c - g_i^c| \\ s.t.\ Y^T Y = I_C,\ Y \geq 0. \tag{5}$$

where $\epsilon, \lambda, \alpha, \beta, \gamma$ are weight coefficients. By optimizing the objective function, we can learn a good dictionary D and a discriminative classifier W simultaneously. Finally, the learned W is used to predict the label of superpixels.

2.6. Optimization

The objective function is non-convex due to the max function in T_4. In order to translate the original optimization problem into a convex optimization problem, at each iteration, we replace the non-convex part with the first-order Taylor expansion according to CCCP algorithm [11]. Furthermore, we put the orthogonality constraint $\|Y^T Y - I_C\|_F^2$ into the objective function. So Eq. 5 is rewritten as:

$$\min_{Y,W,Z,D} \text{Tr}\left[Y^T L Y\right] + \epsilon \|X - DZ\|_F^2 + \lambda \|z\|_1 \\ + \alpha \left\| Z^T W - Y \right\|_F^2 + \beta \|W\|_F^2 + \frac{\mu}{2} \|Y^T Y - I_C\|_F^2 \\ + \gamma \sum_{i=1}^{I} \sum_{c=1}^{C} [(1 - g_i^c) h_c Y^T q_i + g_i^c (1 - h_c B U_i Y h_c^T)] \\ s.t.\ Y \geq 0. \tag{6}$$

where $h_c \in \mathcal{R}^{\mathcal{C}}$ is an indicator vector with all elements but $c-$th element are zeros and $B = [B_1, ..., B_i, ..., B_I]$, each $B_i = [b_{i1}^T, ..., b_{ic}^T, ..., b_{iC}^T] \in \mathcal{R}^{C \times n_i}$ is a matrix corresponding to the image X_i and $b_{ic} = \eta^T$. And $U_i = diag(u_1, ..., u_i)$, $u_k = 0_{n_k \times n_k}$ for $k = 1, ..., i-1, i+1, ..., I$ and $u_i = I_{n_i} \times n_i$.

To solve the optimization problem, we propose an iterative optimization algorithm. At r^{th} iteration:

1. Compute the sparse coding Z with the previous round of dictionary by gradient descent method.

$$\frac{\partial \mathcal{L}}{\partial Z} = 2\alpha W (W^T Z - Y^T) + 2\epsilon D^T (DZ - X) + \lambda \text{sign}(Z) \tag{7}$$

$$Z = Z - \rho_t \frac{\partial \mathcal{L}}{\partial Z} \tag{8}$$

where $\rho_t = \rho / t$ is the learning rate at t^{th} iteration in the gradient descent algorithm and ρ is the initial learning rate.

2. After Z is updated, let $\partial \mathcal{L} / \partial W = 0$ and we obtain

$$W = \alpha (\alpha Z Z^T + \beta I)^{-1} Z Y \tag{9}$$

3. After W is updated, let $\partial \mathcal{L} / \partial Y = 0$ and we have

$$Y_{ij} \leftarrow Y_{ij} \frac{2(\mu Y)_{ij}}{(2MY + P + 2\mu Y Y^T Y)_{ij}} \tag{10}$$

where $P = \gamma \sum_{i=1}^{I} \sum_{c=1}^{C} [(1 - g_i^c) q_i h_c - g_i^c U_i^T B^T h_c^T h_c]$.

4. After Y is updated, let $\partial \mathcal{L} / \partial D = 0$ and we obtain

$$D = X Z^T (Z Z^T)^{-1} \tag{11}$$

Repeat step 1 to 4 until the objective function converges to a stable state.

3. EXPERIMENTS

We evaluate the proposed model on two public datasets:

MSRC-21: This dataset contains 591 images from 21 categories with pixel-level ground truths and there are about 3 labels for each image on average. Here we use the standard spilt into training and test sets as defined in [12].

LabelMe Outdoor (LMO): The dataset consists of 2688 images from 33 classes also with pixel-level gound truths. There are around 5 labels per image on average. It is randomly spilt into 2488 training images and 200 test images.

We employ SLIC algorithm [13] to divide all the images into over-segmented patches (superpixels). After image oversegmentation, we use SIFT [14] as the local descriptor and the typical bag-of-words model as the original representation of each superpixel. To evaluate the performance of semantic segmentation, we adopt 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, the parameter settings are given in table 1.

3.1. Experiments on MSRC Dataset

We compare the proposed dictionary learning based superpixels clustering (DLSC) approach with two fully-supervised methods [12, 15] and four weakly-supervised methods [7, 16, 5, 10]. Table 2 shows the general comparison and the detailed

Parameters	ρ	ϵ	λ	α	β	γ	μ
MSRC	0.1	2×10^5	8×10^4	10^4	10^4	115	10
LabelMe	0.1	2×10^5	8×10^4	10^4	10^4	400	10

Table 1. The parameter settings on MSRC and LabelMe.

Method	average	building	grass	tree	cow	sheep	sky	airplane	water	face	car	bicycle	flower	sign	bird	book	chair	road	cat	dog	body	boat
Texboost [12], FS	58	62	**98**	**86**	58	50	83	60	53	74	63	75	63	35	19	92	15	86	54	19	62	7
HCRF [15], FS	**75**	**80**	96	**86**	74	87	**99**	74	**87**	86	**87**	82	**97**	**95**	30	86	31	**95**	51	69	**66**	9
MTL [7], WS	37	7	96	18	32	6	**99**	0	46	**97**	54	74	54	14	9	82	1	28	47	5	0	0
MRF [16], WS	50	45	64	71	75	74	86	81	47	1	73	55	88	6	6	63	18	80	27	26	55	8
MIM [5], WS	67	12	83	70	**81**	**93**	84	**91**	55	**97**	**87**	**92**	82	69	51	61	59	66	53	44	9	**58**
WSDC [10], WS	68	65	**98**	50	65	55	79	30	50	85	86	59	96	51	60	89	72	77	**93**	63	25	49
DLSC, WS	70	71	97	64	77	68	79	52	66	71	84	50	**97**	69	**64**	**93**	**76**	62	**93**	**70**	19	56

Table 2. Semantic segmentation results on MSRC-21 dataset. WS denotes weak supervision and FS denotes full supervision.

Method	Texboost [17]	LT [3]	Supix [18]	THSR [19]	MIM [5]	GMIM [6]	WSDC [10]	DLSC
Supervision	FS	FS	FS	FS	WS	WS	WS	WS
Accuracy	13	24	29	32	14	21	26	**31**

Table 3. Semantic segmentation results on LabelMe dataset. WS denotes weak supervision, FS denotes full supervision.

comparison results for individual classes. From the results, we can draw three conclusions: (1) In general comparison with the weakly-supervised methods, our method achieve the highest average accuracy of 70 percent, which validates the effectiveness of our method. And the average accuracy of ours is close to the best fully-supervised approach HCRF. (2) In detailed comparison with the baselines, our DLSC algorithm get the best results on 6 out of 21 categories. (3) Compared with WSDC, the proposed DLSC method has better performance on the average accuracy and 13 out of 21 categories. It demonstrates that the employment of sparse coding as feature representation with dictionary learning is more effective than the original feature representation in WSDC.

3.2. Experiments on LabelMe Dataset

Seven popular methods [17, 3, 18, 19, 5, 6, 10] are implemented as benchmark baselines for comparison with our model, as shown in Table 3. From the results, we can observe that the proposed DLSC algorithm surpasses the other weakly-supervised methods and even some full-supervised approaches (e.g. LT) significantly. Besides, THSR achieves a little higher performance than our method on account that THSR is under full-supervised constraint. In other words, our method is comparable with full-supervised approaches, which shows the effectiveness of DLSC. It's also worth pointing out that the performance has been improved more obviously on LabelMe than on MSRC dataset. It is reasonable since MSRC is a simple and over-development dataset, while labelMe is more challenging and diverse one.

3.3. Out-of-Sample Discussion

To evaluate the generalization performance of our model, we conduct experiments on out-of-sample problem. We adopt different data settings during the learning and predicting periods. The standard training set and test set are denoted by τ_1 and τ_2 respectively. Table 4 display the results of our proposed algorithm under different settings on MSRC and LabelMe respectively. Several observations can be obtained. In general, the fluctuation of the accuracy rate across different settings is quite minimal, which proves the stability of our approach. Comparing setting 2 and setting 3, we can find the performance is scarcely affected whether the test images are added into the training set or not. The aforementioned phenomena indicate our approach has certain stability and high generalization performance.

Order	Learning	Predicting	Accuracy	
			MSRC	LabelMe
1	$\tau 1 + \tau 2$	$\tau 1 + \tau 2$	**71.0**	30.8
2	$\tau 1 + \tau 2$	$\tau 2$	70.6	**31.5**
3	$\tau 1$	$\tau 2$	70.4	31.3

Table 4. Results of DLSC under different settings.

4. CONCLUSION

In this paper, we propose a novel dictionary learning based superpixels clustering (DLSC) approach to collaboratively perform image segmentation and tag alignment with those segmented regions. Our model combines dictionary learning, dual clustering and weakly-supervised constraint. To obtain discriminative features, sparse coding is employed to represent each superpixel, and the dictionary for sparse representation is jointly learned during the clustering process. On base of the proposed framework, image semantic segmentation can be effectively implemented.

5. ACKNOWLEDGEMENT

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

6. REFERENCES

[1] Li-Jia Li, Richard Socher, and Li Fei-Fei, "Towards total scene understanding: Classification, annotation and segmentation in an automatic framework," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 2036–2043.

[2] Andrew Rabinovich, Andrea Vedaldi, Carolina Galleguillos, Eric Wiewiora, and Serge Belongie, "Objects in context," in *Computer vision, 2007. ICCV 2007. IEEE 11th international conference on*. IEEE, 2007, pp. 1–8.

[3] Ce Liu, Jenny Yuen, and Antonio Torralba, "Nonparametric scene parsing: Label transfer via dense scene alignment," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 1972–1979.

[4] Richard Socher and Li Fei-Fei, "Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 966–973.

[5] Alexander Vezhnevets, Vittorio Ferrari, and Joachim M Buhmann, "Weakly supervised semantic segmentation with a multi-image model," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 643–650.

[6] Alexander Vezhnevets, Vittorio Ferrari, and Joachim M Buhmann, "Weakly supervised structured output learning for semantic segmentation," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 845–852.

[7] Alexander Vezhnevets and Joachim M Buhmann, "Towards weakly supervised semantic segmentation by means of multiple instance and multitask learning," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 3249–3256.

[8] Xiaobai Liu, Shuicheng Yan, Jiebo Luo, Jinhui Tang, Zhongyang Huango, and Hai Jin, "Nonparametric label-to-region by search," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 3320–3327.

[9] Si Liu, Shuicheng Yan, Tianzhu Zhang, Changsheng Xu, Jing Liu, and Hanqing Lu, "Weakly supervised graph propagation towards collective image parsing," *Multimedia, IEEE Transactions on*, vol. 14, no. 2, pp. 361–373, 2012.

[10] Yang Liu, Jing Liu, Zechao Li, Jinhui Tang, and Hanqing Lu, "Weakly-supervised dual clustering for image semantic segmentation," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. IEEE, 2013, pp. 2075–2082.

[11] Alan L Yuille and Anand Rangarajan, "The concave-convex procedure," *Neural Computation*, vol. 15, no. 4, pp. 915–936, 2003.

[12] Jamie Shotton, John Winn, Carsten Rother, and Antonio Criminisi, "Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation," in *Computer Vision–ECCV 2006*, pp. 1–15. Springer, 2006.

[13] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Susstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 11, pp. 2274–2282, 2012.

[14] David G Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[15] Lubor Ladicky, Christopher Russell, Pushmeet Kohli, and Philip HS Torr, "Associative hierarchical crfs for object class image segmentation," in *Computer Vision, 2009 IEEE 12th International Conference on*. IEEE, 2009, pp. 739–746.

[16] Jakob Verbeek and Bill Triggs, "Region classification with markov field aspect models," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*. IEEE, 2007, pp. 1–8.

[17] Jamie Shotton, John Winn, Carsten Rother, and Antonio Criminisi, "Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context," *International Journal of Computer Vision*, vol. 81, no. 1, pp. 2–23, 2009.

[18] Joseph Tighe and Svetlana Lazebnik, "Superparsing: scalable nonparametric image parsing with superpixels," in *Computer Vision–ECCV 2010*, pp. 352–365. Springer, 2010.

[19] Heesoo Myeong and Kyoung Mu Lee, "Tensor-based high-order semantic relation transfer for semantic scene segmentation," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. IEEE, 2013, pp. 3073–3080.