Weakly-Supervised Dual Clustering for Image Semantic Segmentation

Yang Liu[†], Jing Liu[†], Zechao Li[†], Jinhui Tang[‡], Hanqing Lu[†]
[†]NLPR, Institute of Automation, Chinese Academy of Sciences, Beijing, China, 100190.
[‡]School of Computer Science, Nanjing University of Science and Technology, China, 210044.
{liuyang6, jliu, luhq}@nlpr.ia.ac.cn, zechao.li@gmail.com, jinhuitang@mail.njust.edu.cn

Abstract

In this paper, we propose a novel Weakly-Supervised Dual Clustering (WSDC) approach for image semantic segmentation with image-level labels, i.e., collaboratively performing image segmentation and tag alignment with those regions. The proposed approach is motivated from the observation that superpixels belonging to an object class usually exist across multiple images and hence can be gathered via the idea of clustering. In WSDC, spectral clustering is adopted to cluster the superpixels obtained from a set of over-segmented images. At the same time, a linear transformation between features and labels as a kind of discriminative clustering is learned to select the discriminative features among different classes. The both clustering outputs should be consistent as much as possible. Besides, weakly-supervised constraints from image-level labels are imposed to restrict the labeling of superpixels. Finally, the non-convex and non-smooth objective function are efficiently optimized using an iterative CCCP procedure. Extensive experiments conducted on MSRC and LabelMe datasets demonstrate the encouraging performance of our method in comparison with some state-of-the-arts.

1. Introduction

Image semantic segmentation is to automatically parse images into some semantic regions. This is a coherent task between image segmentation and region-level label assignment. That is, the two issues are inseparable and promote mutually. Intuitively, exact segmentations can provide representative features for pixel labeling. In turn, precise labeling results will boost image segmentation since the pixels with the same label can be deemed as a whole object. From this view, semantic segmentation is a kind of higher-level image understanding than any individual case. Accordingly, the solution about the problem is really challenging but valuable to support fine-grained image analysis, retrieval or other possible applications.

Recently, image semantic segmentation has become a popular research topic and some efforts contribute to the problem [3, 26]. Most works focus on fully or partially supervised setting which means each or partial pixels are manually labeled for model training [18, 11, 6, 22]. However, producing pixel-level labels is time-consuming and may be inaccurate. Fortunately, lots of image sharing websites provide us plentiful user-contributed images with social tags, in which the raw correspondences between images and labels are available. Thus, weakly-supervised methods [25, 26, 27] with only image-level labels available have emerged and attracted more attention.

In this paper, we propose a coherent framework under the weakly-supervised setting to perform holistic image understanding, i.e., obtaining meaningful image regions and simultaneously assigning image-level labels to those regions. The problem is formulated as a Weakly-Supervised Dual Clustering (WSDC) task to cluster superpixels and assign a suitable label to each cluster. The first evidence of our method is that similar superpixels have high probability to share the same label. To mine this kind of important contextual relationship, a spectral clustering term is defined over the superpixels of all images to group the visually similar ones together. The second evidence is that there is rich discriminative information among different object classes, e.g., not all the features are important and discriminative for a certain class. We define a discriminative clustering term and require its outputs to be consistent with the outputs of spectral clustering. Besides, we explicitly impose weakly-supervised constraints during the dual clustering process which can assign labels to clusters. Incorporating these three terms, the problem is formulated as a non-convex and non-smooth objective function, which is optimized via an iterative CCCP algorithm [1]. Finally, extensive experiments on the public datasets, i.e., MSRC and LabelMe, demonstrate the encouraging performance of our algorithm. Figure 1 illustrates the flowchart of the proposed method.

Our main contributions are summarized as follows.

- We propose a coherent framework to jointly solve image segmentation and region-level annotation under

1063-6919/13 $26.00 © 2013 IEEE
DOI 10.1109/CVPR.2013.270

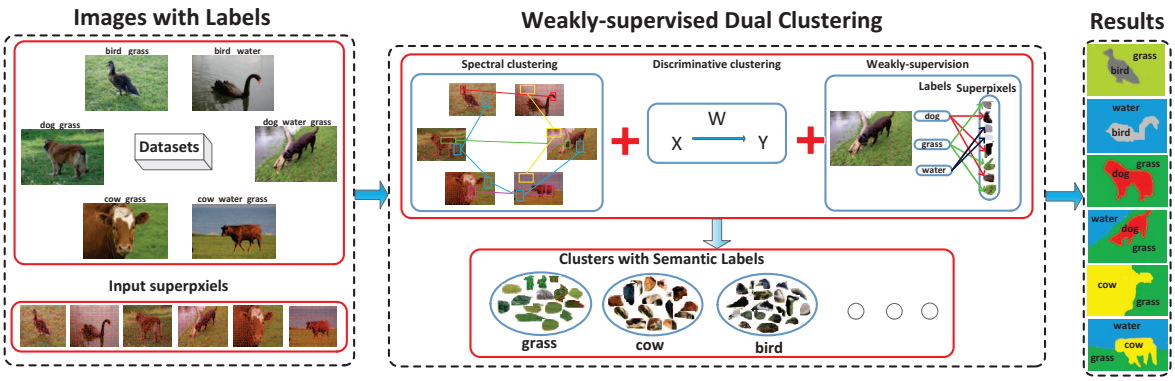

Figure 1. The flowchart of our method.

the weakly-supervised setting. Furthermore, the output of the model can also be used to semantically segment any test images with or without labels.

- The proposed method incorporates the spectral clustering and discriminative clustering to cluster superpixels from all images into different clusters, and imposes image-level labels as a kind of weak supervision to assign labels to clusters.
- An efficient iterative CCCP solution is designed to solve the non-convex and non-smooth objective function.

2. Related Work

In this section, we review some works related with ours in several aspects.

Image Semantic segmentation. From the methodology view, the methods can be roughly divided into three categories: fully-supervised, semi-supervised and weakly-supervised. In the fully-supervised setting, CRF (Conditional Random Field) [21, 19] models are used typically and have lots of effective extensions [5]. Its basic formulation is defined over image pixels and various potential functions are proposed to depict the relations of multiple units. However, the CRF-style models often have complicated structures and many parameters which are hard to optimize and inference. To alleviate the dependence of fine-labeled training data, Socher et.al [22] proposed a semi-supervised model to find a mapping between visual and textual words by projecting them into a latent meaning space, in which partial fine-grained labeled images are also needed. Li et. al [6] proposed a partially-supervised hierarchical generative model to jointly classify, annotate, and segment various scene images. While the model estimation required a handful of clean images in which some object regions are marked with their corresponding tags. From this view, the above fully-supervised or semi-supervised solutions are very limited due to the high cost on the acquisition of fine-grained image labels. Weakly-supervised semantic segmentation [26, 27] arised to solve this problem. Vezhnevets et.al [26] proposed a graphical MIM model and introduced an objectness to distinguish objects from background classes. The work [27] is an extension of [26], in that work, the author built a multiple image model and adopted a parameter family of CRF models, to evaluate the quality of each model in the family, a model selection criterion is proposed.

Label To Region. Label to region means reassign the labels annotated at the image-level to those segmented image regions rather than the whole image [13, 12, 23]. Liu et.al proposed a bi-layer sparse coding formulation for reconstructing an image region using the over-segmented image patches. And they further improved the work to solve the problem by search on web [15]. Yang et.al [28] proposed the spatial group sparse coding by integrating the spatial correlations among training regions. However, all the works adopt a sequential pipeline to first over-segment images and then design suitable models to describe the intro- and inter- correlations among labels and segmented regions, while the performance of region-level tagging will be certainly degenerated by imperfect segmentation algorithms.

Image Cosegmentation. Co-segmentation [7, 8, 16] means to simultaneously segment a common salient foreground object from a set of images which can be seen as a special case of our work. Kim et.al [8] maximized the overall temperature of images associated with a heat diffusion process and the position of sources corresponding to different classes. Joulin [7] proposed a novel energy-minimization approach to cosegmentation that can handle multiple classes and images. Most existing works are only applied to a subgroup of images with same foreground and not intended to handle irregularly appearing multiple foregrounds. Besides, they did not explore any supervision like easily available image-level labels in their learning process.

3. Weakly-Supervised Dual Clustering

To uncover the correspondence between image superpixels and semantic labels, in this work we develop a weakly-supervised dual clustering model by simultaneously max-

imizing the appearance consistency of superpixels within the same class and the separability of multiple classes. The former problem leads to solving a bottom-up unsupervised clustering while the latter problem leads to methods designed for top-down discriminative clustering problem.

3.1. Notations

Assume we have a data collection with I images $\mathcal{X} = \{X_1, \cdots, X_i, \cdots, X_I\}$. Let $X = [X_1, \cdots, X_i, \cdots, X_I]$ denote the data matrix with $X_i = [x_i^1, \cdots, x_i^{n_i}]$, where $x_i^k \in \mathcal{R}^d$ is the feature descriptor of the k-th superpixel in the i-th image and n_i is the number of superpixels in the i-th image. For brevity, we denote $X = [x_1, \cdots, x_i, \cdots, x_N]$ without confusion, where $N = \sum_{i=1}^{I} n_i$. Suppose these I images are sampled from C classes and the label information is defined as $G = [g_1, ..., g_i, ...g_I] \in \{0,1\}^{C\times I}$, where $g_i \in \{0,1\}^C$ is the label vector of X_i. $g_i^c = 1$ if X_i belongs to the c-th class and 0 otherwise. The predicted superpixels-level label matrix $Y \in \mathcal{R}^{N\times C}$ is defined as

$$y_n^c = \begin{cases} 1, & \text{if the } n\text{-th superpixel belongs to the } c\text{-th class,} \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

3.2. Spectral Clustering

On the one hand, visually similar superpixels have high probability to share the same label. On the other hand, spectral techniques have been demonstrated to be effective to detect the cluster structure [20], which can integrate the consistency relationships of superpixels among different images. In light of this, we employ spectral techniques to mine the aforementioned contextual information.

The interactions among superpixels are represented by an affinity matrix $S \in \mathcal{R}^{N\times N}$ defined as

$$S_{ij} = \begin{cases} \exp(-\frac{\|x_i - x_j\|^2}{\sigma^2}), & x_i \in \mathcal{N}_k(x_j) \text{ or } x_j \in \mathcal{N}_k(x_i), \\ 0, & \text{otherwise.} \end{cases}$$

Here $\mathcal{N}_k(x)$ is the set of k-nearest superpixels of x. The k-nearest superpixels are selected only from the superpixels from one image or the images sharing common labels, because the label of a superpixel is identified from labels of the image it belongs to. σ a free parameter to control the decay rate. In addition, to encourage spatially smooth labelings, the spatial neighbor superpixels within the same image are also connected. Then the spectral clustering term is defined as minimizing the following equation:

$$\mathcal{J}(Y) = \frac{1}{2}\sum_{i,j=1}^{N} S_{ij} \| \frac{y_i}{\sqrt{A_{ii}}} - \frac{y_j}{\sqrt{A_{jj}}} \|_2^2 = \text{Tr}[Y^T L Y], \tag{2}$$

where A is a diagonal matrix with $A_{ii} = \sum_{j=1}^{N} S_{ij}$ and $L = A^{-1/2}(A - S)A^{-1/2}$ is the normalized Laplacian matrix.

3.3. Discriminative Clustering

Since not all the features are important and discriminative for a certain class, a discriminative clustering strategy with $l_{2,1}$-norm regularization is introduced. Its outputs are required to be consistent with the outputs of spectral clustering. Besides, it is required to adaptively choose the discriminative features. To this end, we assume that there is a linear transformation $W \in \mathcal{R}^{d\times C}$ between features and the predicted labels. Therefore, the objective function for discriminative clustering is formulated as

$$\min \mathcal{L}(Y, W) = \alpha \sum_{i=1}^{N} loss(y_i, W^T x_i) + \beta \|W\|_{2,1}, \tag{3}$$

where $loss$ is a loss function to be defined, and α and β are two nonnegative parameters. The $l_{2,1}$-norm is defined as $\|W\|_{2,1} = \sum_{i=1}^{d} \sqrt{\sum_{j=1}^{C} W_{ij}^2}$. The $l_{2,1}$-norm regularization term is imposed to ensure W sparse in rows. In that way, the proposed method is able to handle correlated and noisy features and enable to evaluate the correlation between labels and features.

For simplify, in this work we adopt the least square loss function and then have

$$\mathcal{L}(Y, W) = \alpha \|X^T W - Y\|_F^2 + \beta \|W\|_{2,1}. \tag{4}$$

Through learning the linear transformation, i.e., a mapping function from visual features to labels, the discriminative feature representations for each class can be obtained.

3.4. Weakly-Supervised Constraint

Given an image and its associated labels, it is reasonable and natural to restrict the mapping between superpixels and labels to meet the following constraints.

- One superpixel corresponds to at most one label.
- One label has at least one superpixel mapped to it. It guarantees that if a label is assigned to an image, there is at least one superpixel supporting this label.
- Superpixels should correspond to the labels of images they belong to. This makes sure that there are no image superpixels supporting an invalid label.

To satisfy the first constraint, we impose an orthogonality constraint on Y just like [10], i.e., $Y^T Y = I_C \in \mathcal{R}^{C\times C}$, where I_C is an identity matrix. Since Y is the cluster indicator, it is reasonable to constraint $Y \geq 0$. When both nonnegative and orthogonal constraints are satisfied, only one element in each row of Y is greater than zero and all of the others are zeros. Hence the learned Y is more accurate and more capable to provide discriminative information.

To satisfy the last two conditions, we explicitly impose a weak-supervision constraint with a hyper-parameter γ:

$$\mathcal{Q}(Y) = \gamma \sum_{i=1}^{I}\sum_{c=1}^{C} | \max_{x_{ij}\in X_i} y_{ij}^c - g_{ic}|. \tag{5}$$

where y_{ij}^c is the value of Y corresponding to the j-th superpixel within the i-th image on label c. Since it is difficult to directly dealing with Eq. 5 and $y_{ij}^c \in [0,1]$, we have:

$$|\max_{j\in X_i} y_{ij}^c - g_i^c| = \begin{cases} 1-\max_{x_{ij}\in X_i} y_{ij}^c, & \text{if } g_i^c\text{=1,} \\ \max_{x_{ij}\in X_i} y_{ij}^c, & \text{else.} \end{cases} \quad (6)$$

Then the right side of Eq. 5 is rewritten as:

$$\gamma[\sum_i^I \sum_c^C (1-g_i^c) \max_{x_{ij}\in X_i} y_{ij}^c + \sum_i^I \sum_c^C g_i^c (1 - \max_{x_{ij}\in X_i} y_{ij}^c)]. \quad (7)$$

Similar with [14], the first term is further relaxed by $\sum_i \sum_c (1-g_i^c) \sum_{x_{ij}\in X_i} y_{ij}^c$. Then $\mathcal{Q}(Y)$ is rewritten as:

$$\mathcal{Q}(Y) = \gamma \sum_{i=1}^I \sum_{c=1}^C [(1-g_i^c) h_c^T Y^T q_i + g_i^c (1 - \max_{x_{ij}\in X_i} p_{ij}^T Y h_c)], \quad (8)$$

where $h_c \in \mathcal{R}^C$ is an indicator vector whose all elements except for the c-th element are zeros. $q_i \in \mathcal{R}^N$ is a vector whose all elements excepts for those elements corresponding to the i-th image are zeros. $p_{ij} \in \mathcal{R}^N$ is an indicator vector whose element corresponding to the j-th superpixel in the i-th image is one and other elements are zeros.

3.5. The Proposed Formulation

Jointly considering the above three aspects, we obtain a unified objective function $\mathcal{J}(Y) + \mathcal{L}(Y,W) + \mathcal{Q}(Y)$:

$$\begin{aligned} &\min_{Y,W} \mathrm{Tr}[Y^T L Y] + \alpha \|X^T W - Y\|_F^2 + \beta \|W\|_{2,1} \\ &+ \gamma \sum_{i=1}^I \sum_{c=1}^C [(1-g_i^c) h_c^T Y^T q_i + g_i^c (1 - \max_{x_{ij}\in X_i} p_{ij}^T Y h_c)] \\ &s.t. \quad Y^T Y = I_C, \quad Y \geq 0. \end{aligned} \quad (9)$$

Note that the $l_{2,1}$-norm regularization is non-smooth and the max term is non-convex. So the objective function is not convex over Y and W simultaneously. In Section 4, we focus on how to solve this optimization problem.

4. Optimization Algorithm

4.1. CCCP Algorithm

The CCCP algorithm solves the optimization problem using an iterative process. At each round t, given an initial value, CCCP substitutes the concave part of the objective function using the 1-st order Taylor expansion. The sub-optimum solution is achieved by iteratively optimizing the subproblem until convergence.

Since the last term in Eq. 9 is a sum term, we consider only the term related with g_{ic}. Let $l = [y_{i1}^c, ..., y_{ij}^c, ..., y_{in_i}^c]^T$, we pick the subgradient of l with $\eta \in \mathcal{R}^{n_i}$ and its j-th element is given by:

$$\eta_j = \begin{cases} \frac{1}{n_\alpha}, & \text{if } l_j^{(t)} = \max(l^{(t)}), \\ 0, & \text{otherwise.} \end{cases} \quad (10)$$

where n_α is the number of superpixels with the largest label value $\max l^{(t)}$. At the $(t+1)$-th iteration, we estimate the current l based on $l^{(t)}$ and the corresponding $\eta^{(t)}$. As $\eta^T l^{(t)} = \sum_j \eta_j l_j^{(t)} = \max l^{(t)} \sum_{\eta_j \neq 0} \eta_j = \max l^{(t)}$, for the function $\max(l)$, its 1-st order Taylor expansion is approximated as $(\max l)_{l^{(t)}} \approx \max l^{(t)} + \eta^T (l - l^{(t)}) = \max l^{(t)} + \eta^T l - \max l^{(t)} = \eta^T l$, which could be also written as :

$$\sum_{i=1}^I \sum_{c=1}^C g_i^c (1 - h_c B U_i Y h_c^T) \quad (11)$$

where $B = [B_1, ..., B_i, ..., B_I]$, each $B_i = [b_{i1}^T, ..., b_{ic}^T, ...b_{iC}^T] \in \mathcal{R}^{C\times n_i}$ is a matrix corresponding the image i and $b_{ic} = \eta^T$. $U_i \in \mathcal{R}^{N\times N}$ is a diagonal block matrix, $U_i = diag(u_1, .., u_i)$, $u_k = 0_{n_k \times n_k}$ for $k = 1, ...i-1, i+1, ...I$ and $u_i = I_{n_i \times n_i}$.

4.2. Iterative Optimization

Now, we adopt an iterative optimization process. First, we relax the orthogonal constraint and the optimization problem (9) becomes

$$\begin{aligned} &\min_{Y,W} \mathscr{L}(Y,W) = Tr(Y^T L Y) + \alpha \|X^T W - Y\|_F^2 + \beta \|W\|_{2,1} \\ &+ \gamma \sum_{i=1}^I \sum_{c=1}^C [(1-g_i^c) h_c Y^T q_i + g_i^c (1 - h_c B U_i Y h_c^T)] \\ &+ \frac{\mu}{2} \|Y^T Y - I_C\|_F^2 \\ &s.t. \quad Y \geq 0. \end{aligned} \quad (12)$$

where $\mu \geq 0$ is a parameter to control the orthogonality constraint. In our experiments it is set large enough to ensure the orthogonality constraint satisfied. We have

$$\begin{aligned} &\frac{\partial \mathscr{L}(Y,W)}{\partial W} = 2(\alpha X (X^T W - Y) + \beta D W) = 0 \\ &\Rightarrow W = \alpha (\alpha X X^T + \beta D)^{-1} X Y. \end{aligned} \quad (13)$$

Here D is a diagonal matrix with $D_{ii} = \frac{1}{2\|w_i\|_2}$. Substituting W by Eq. 13, Eq. 12 can be rewritten as:

$$\begin{aligned} \min_Y \quad &\mathscr{L} = Tr[Y^T M Y] + \gamma [\sum_i^I \sum_c^C (1-g_i^c) h_c Y^T q_i \\ &+ \sum_i^I \sum_c^C g_i^c (1 - h_c B U_i Y h_c^T)] + \frac{\mu}{2} \|Y^T Y - I_C\|_F^2 \\ &s.t. \quad Y \geq 0. \end{aligned} \quad (14)$$

where $M = L + \alpha (I_N - \alpha X^T (\alpha X X^T + \beta D)^{-1} X)$ and $I_N \in \mathcal{R}^{N\times N}$ is an identity matrix. To optimize the above problem, we introduce multiplicative updating rules. Letting ϕ_{ij} be the Lagrange multiplier for constraint $Y_{i,j} \geq 0$

Algorithm 1 Weakly-supervised Dual clustering.

Input:

Data matrix $X \in \mathcal{R}^{d\times N}$;
Label matrix $G \in \mathcal{R}^{C\times I}$;
Parameters α, β, γ, μ.

Construct the k-nearest neighbor graph and calculate L;
The iteration step $t = 1$;
Initialize $Y \in \mathcal{R}^{N\times C}$;
Set $D^t \in \mathcal{R}^{d\times d}$ as an identity matrix.
repeat
  $W^t = \alpha(\alpha XX^T + \beta D^t)^{-1}XY^t$;
  $M^t = L + \alpha(I_N - \alpha X^T(\alpha XX^T + \beta D^t)^{-1}X)$;
  calculate B^t;
  calculate P^t according Eq. 16;
  $Y_{ij}^{t+1} \leftarrow Y_{ij}^t \frac{2(\mu Y^t)_{ij}}{(2M^tY^t+P^t+2\mu Y^t(Y^t)^TY^t)_{ij}}$
  update the diagonal matrix D^{t+1} as $D_{ii}^t = \frac{1}{2\|W_i^t\|_2}$;
  t=t+1;
until Convergence criterion satisfied

Output:

label matrix Y;
multi-class classifier W.

and $\Phi = [\phi_{ij}]$, the lagrange function is $\mathscr{L} + Tr(\Phi Y^T)$. Setting its derivative with respect to Y to 0, we obtain

$$2MY + P + 2\mu YY^TY - 2\mu Y + \Phi = 0, \quad (15)$$

where

$$P = \gamma \sum_{i=1}^{I}\sum_{c=1}^{C}[(1 - g_i^c)q_i h_c - g_i^c U_i^T B^T h_c^T h_c]. \quad (16)$$

Using the Karush-Kuhn-Tuckre (KKT) condition [9] $\phi_{ij}Y_{ij} = 0$, we obtain the updating rules:

$$Y_{ij} \leftarrow Y_{ij}\frac{2(\mu Y)_{ij}}{(2MY + P + 2\mu YY^TY)_{ij}} \quad (17)$$

Then we normalize Y such that $(Y^TY)_{ii} = 1, i = 1, ..., C$. The optimization algorithm is summarized in Algorithm 1.

5. Experiments

In this section, we conduct extensive experiments to validate the performance of the proposed method and discuss the experimental analysis.

5.1. Datasets

To verify the effectiveness of our method, we conduct experiments on two public and challenging datasets, i.e., MSRC [21] and LabelMe [11].

MSRC: It is a widely used dataset in semantic segmentation task. It contains 591 images from 21 different classes and there are 3 labels per image on average. The dataset is split into 276 training images and 256 test images.

LabelMe [11]: It is a more challenging dataset than MSRC. It contains 2688 images from 33 classes. There are 2488 training images and 200 test images.

The both datasets are provided with pixel-level groundtruth. We adopt SLIC algorithm [2] to obtain the superpixels for each image, and describe each superpixel by the typical bag-of-words representation while using SIFT [17] as the local descriptor. To present fair comparisons with other methods, we use training images to learn our model, and use test images to evaluate the performance.

We evaluate the performance of semantic segmentation from two views: the labeling performance and segmentation performance. The labeling performance is usually evaluated via two kinds of quantitative measures: total accuracy (T_Acc) which measures the percentage of classified pixels, and average per-class accuracy (Aver_Acc) which measures the percentage of correctly classified pixels for a class then averaged over all classes. Because the various baselines on the both datasets adopt different evaluation standards so we report different measures to accord with the corresponding baselines. For segmentation evaluation metric we adopt the intersection-over-union score (IOU score) [7] which is a standard measure in PASCAL challenges. It is defined as $\max_k \frac{1}{|I|}\sum_{i\in I}\frac{GT_i \bigcap R_i^k}{GT_i \bigcup R_i^k}$, where GT_i is the groundtruth and R_i^k the region associated with the k-th class in the image i.

5.2. Parameter Analysis

Five parameters need to be set in WSDC, k in the k-nearest graph construction, α and β in Eq. 4, γ in Eq. 5, μ in Eq. 12. We set $k = 50$ to construct the k-nearest graph. In the experiment we find that α is insensitive so we fixed $\alpha = 1000$ empirically. μ is set to be 10^8 which is large enough to guarantee the orthogonality constraint satisfied. Specifically, we focus on the effects of β and γ, because the two parameters are crucial to our results. The range of β and γ are $\{10, 10^2, 10^3, 10^4, 10^5\}$ and $\{10^2, 10^3, 10^4, 10^5, 10^6, 10^7\}$ respectively. The semantic segmentation performance is used to tune parameters. The results on both datasets are shown in Fig. 2. We can observe the following conclusions. Firstly, when β and γ increase from small to large, the performance varies apparently, which shows that the $l_{2,1}$-norm term and weakly-supervision constraint have great impacts on the performance. Secondly, accuracies reach the peak points when $\beta = 10^3, \gamma = 10^4$ and $\beta = 10^4, \gamma = 10^6$ on both datasets respectively which all lie in the middle range and the accuracies do not increase monotonically when β and γ increase. Because extremely large β makes the rows sparsity overwhelming and extremely small β will fail to select the discriminative features. Extremely large γ will lead to neglect the effects of other terms which is also inadvisable. In the following experiments, we adopt the best

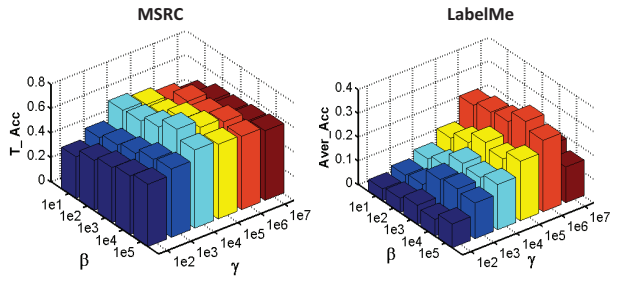

Figure 2. Parameter tuning results of paramters β and γ for MSRC and LabelMe.

parameter settings on both datasets.

5.3. Experiments on MSRC dataset

We compare the proposed algorithm with LAS [15], MTL-RF [25], MIM [26] and RLSIM [4] to evaluate the semantic segmentation performance. We summarize these methods from the three sides as in Table 1: Supervision, ILP (Image Label Prior) and MOF (Multiple of Features). 'Full' supervision means each pixel is labeled with a tag and 'Weak' supervision means only image-level labels are available. 'With' ILP represents during the predicting period, the images' labels are available and we only predict the labels of superpixels from the image's labels. 'Without' ILP indicates the labels of images are absolutely unknown. MOF = 'yes' stands for the method using multiple features.

Table 2 shows the overall semantic segmentation performance. Two facts can be observed. First, our method achieves the best result which can validate the effectiveness of our method. Even we use single feature and without ILP, our method is comparable even better than other methods. Secondly, unlike RLSIM_1 gets much higher accuracy with ILP than RLSIM_2 without ILP, the results of WSDC_1 and WSDC_2 are very near and both achieve high accuracies, which approves during the prediction period the image-level labels have negligible effect on our algorithm's performance. In addition, requiring image-level priors to boost performance is also a weakness of many semantic segmentation methods. Figure 3 illustrates the per-class accuracies on MSRC. Our method gets the best results on 10 out of 21 classes and especially works well on some very hard classes such as bird, cat, dog, etc.

For segmentation performance, we compare our method with [7, 8, 16]. All the three methods divide the images into some subgroups which images with a same label are deemed as a subgroup, and they process the images from a subgroup at one time. In our experiments, we report the segmentation performance under two settings: one is WSDC_3 which segments images from a subgroup at one time, the other one we directly report the segmentation performance of WSDC_2. Segmenting images of the whole dataset with multiple foregrounds and backgrounds at one time is a big challenge for most segmentation methods.

Table 3. Segmentation performances of our method comparing with other baselines on MSRC dataset.

class	WSDC_2	WSDC_3	[7]	[8]	[16]
bike	27.6	39.9	**43.3**	29.9	42.8
bird	48.2	**48.3**	47.7	29.9	-
car	48.0	52.3	**59.7**	37.1	52.5
cat	**56.0**	52.3	31.9	24.4	5.6
chair	**72.1**	54.3	39.6	28.7	39.4
cow	30.5	43.2	**52.7**	33.5	26.1
dog	42.8	**50.8**	41.8	33.0	-
face	25.3	45.8	**70.0**	33.2	40.8
flower	71.0	**84.9**	51.9	40.2	-
house	28.2	48.6	51.0	32.3	**66.4**
plane	15.6	**35.9**	21.6	25.1	33.4
sheep	56.3	**66.3**	**66.3**	60.8	45.7
sign	51.2	**59.5**	58.9	43.2	-
tree	**71.3**	58.1	67.0	61.2	55.9
average	46.1	**52.9**	50.2	36.6	40.9

Table 3 shows the segmentation performance on MSRC dataset. First, under the same setting, WSDC_3 gets the highest average IOU score comparing with [7, 8, 16]. It can proves that the weakly-supervision information can promote the segmentation performance. Secondly, WSDC_2 obtains comparable results with other methods. It is worth to noting that, segmenting a subgroup of images which share the same foreground itself is a strong supervision. Even WSDC_2 segments the whole image set with multiple foregrounds and backgrounds our method still outperform [8, 16]. The results of WSDC_2 will certainly be effected by the imbalanced labels and irregular appearing foregrounds and backgrounds. Thirdly, our method obtains the best results on 6 out of 14 classes especially on cat and dog categories which are easily confusing objects classes. This reflects that the guidance of weakly-supervision can boost the segmentation performance and especially helpful to disambiguate the easily confusing categories which is also a second target of our method.

5.4. Experiments on LabelMe dataset

Fully-supervised methods [21, 24, 11] and weakly-supervised methods [27, 26] are used as compared methods and the condition settings of them are displayed in Table 4.

Semantic segmentation comparisons on LabelMe are presented in Table 5. Our method outperforms the weakly-supervised methods substantially and is comparable with fully-supervised approaches. The segmentation average IOU score is 20.1%. To our best knowledge, no works have reported the segmentation performance on LabelMe dataset.

Table 1. The experimental settings of our method and baselines on MSRC dataset.

method	MTL-RF [25]	LAS [15]	MIM [26]	RLSIM_1 [4]	RLSIM_2 [4]	WSDC_1	WSDC_2
Supervision	Weak	Weak	Weak	Weak	Weak	Weak	Weak
ILP	Without	Without	With	With	Without	Without	With
MOF	No	No	Yes	No	No	No	No

Table 2. Total accuracy (T_Acc) of our method comparing with baselines on MSRC dataset.

method	MTL-RF [25]	LAS [15]	MIM [26]	RLSIM_1 [4]	RLSIM_2 [4]	WSDC_1	WSDC_2
T_Acc	51	63	67	69	47	69	**71**

Table 6. Total accuracy (T_Acc) of our method under different data settings on MSRC dataset.

order	Learning	Predicting	ILP	T_Acc
1	$\tau_1 + \tau_2$	$\tau_1 + \tau_2$	yes	68.5
2	$\tau_1 + \tau_2$	τ_2	yes	70.7
3	τ_1	τ_2	yes	71.0
4	τ_1	τ_2	no	69.0
5	τ_2	τ_2	yes	**71.4**

Table 7. Average per-class accuracy (Aver_Acc) of our method under different data settings on LabelMe dataset.

order	Learning	Predicting	ILP	Aver_Acc
1	$\tau_1 + \tau_2$	$\tau_1 + \tau_2$	yes	25.0
2	$\tau_1 + \tau_2$	τ_2	yes	**26.3**
3	τ_1	τ_2	yes	26.0
4	τ_1	τ_2	no	25.0
5	τ_2	τ_2	yes	23.2

5.5. Out-of-Sample and Label Prior Discussion

To further investigate the ability of solving the out-of-sample problem of our method, we use different data settings during the learning and predicting periods. We name the standard training set and test set as τ_1 and τ_2 respectively. The results of our method under different data settings on both datasets are reported in Table 6 and Table 7. Several facts can be obtained. First, the highest and lowest accuracies on both datasets under different settings make little difference which proves our method is relatively stable and robust. Second, compared setting 2 and 3, whether the test set τ_2 is explored in the model learning process or not, the obtained accuracies are comparable. Maybe due to the simplicity of MSRC, the out-of-sample setting (setting 3) on the dataset achieves better performance than the in-sample setting (setting 2). Third, the results with ILP are only a little higher than without ILP. This demonstrates that our method is effective to semantically parsing images even no labels are provided. Finally, the proposed algorithm achieves the best performance with setting 5 and setting 2 on the MSRC and LabelMe datasets, respectively. The reason may be that τ_2 of MSRC has more images and fewer class labels than τ_2 of LabelMe.

6. Conclusion

In this paper, we propose a Weakly-Supervised Dual Clustering (WSDC) method to automatically segment the images into localized semantic regions. We combine spectral clustering and discriminative clustering into a unified framework to integrate the contextual relationships between superpixels and discriminative features of multiple classes. To fully exploit discriminative features, we impose the non-negative constraint on the label matrix Y and $l_{2,1}$-norm regularization on the linear transformation. The image-level labels are imposed as weakly-supervised constraints to assign each cluster a semantic label. Extensive experiments on public challenging datasets have shown the effectiveness of our method.

Acknowledgements. This work was supported by 973 Program (2010CB327905) and National Natural Science Foundation of China (61272329,61070104,61202325).

References

[1] R. A Yuille. The concave-convex procedure. *Neural Computation*, 15(4):915–936, 2003. 1

[2] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Ssstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE TPAMI*, 22(8):888 –905, 2012. 5

[3] P. Arbelaez, B. Hariharan, C. Gu, S. Gupta, L. Bourdev, and J. Malik. Semantic segmentation using regions and parts. In *CVPR*, 2012. 1

[4] F. Briggs, X. Z. Fern, and R. Raich. Rank-loss support instance machines for miml instance annotation. In *KDD*, 2012. 6, 7

[5] B. Fulkerson, A. Vedaldi, and S. Soatto. Class segmentation and object localization with superpixel neighborhoods. In *ICCV*, 2009. 2

[6] L. jia Li, R. Socher, and L. Fei-fei. Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. In *CVPR*, 2009. 1, 2

[7] A. Joulin, F. Bach, and J. Ponce. Multi-class cosegmentation. In *CVPR*, 2012. 2, 5, 6

[8] G. Kim, E. P. Xing, L. Fei-Fei, and T. Kanade. Distributed cosegmentation via submodular optimization on anisotropic diffusion. In *ICCV*, 2011. 2, 6

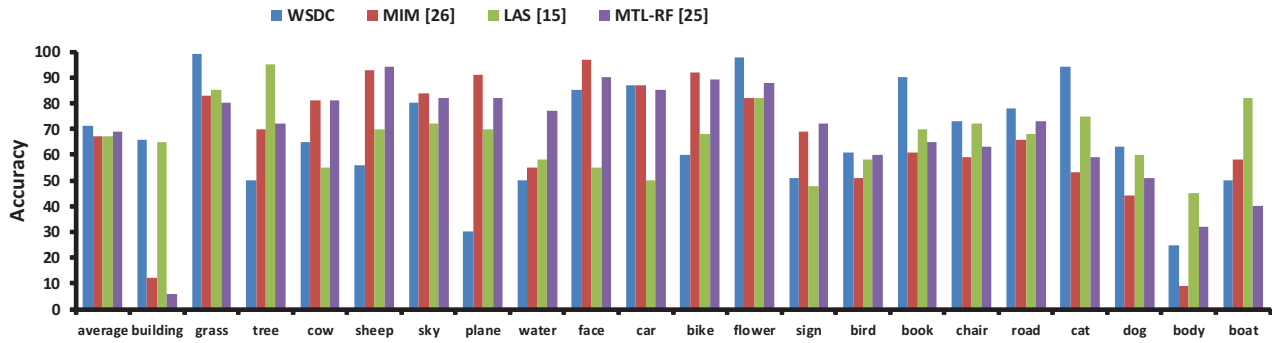

Figure 3. Detailed performance of our method on MSRC dataset.

Table 4. The experimental settings of our method and baselines on LabelMe dataset.

method	Texboost [21]	LT [11]	Supix [24]	MIM [26]	GMIM [27]	WSDC_1	WSDC_2
Supervision	Full	Full	Full	Weak	Weak	Weak	Weak
ILP	Without	Without	Without	With	With	Without	With
MOF	Yes	No	Yes	Yes	Yes	No	No

Table 5. Average per-class accuracy (Aver_Acc) of our method comparing with other baselines on LabelMe dataset.

method	Texboost [21]	LT [11]	Supix [24]	MIM [26]	GMIM [27]	WSDC_1	WSDC_2
Aver_Acc	13	24	**29**	14	21	25	**26**

[9] H. Kuhn and A. Tucker. Nonlinear programming. In *Berkeley Symposium on Mathematical Statistics and Probabilistics*, 1951. 5

[10] Z. Li, Y. Yang, J. Liu, X. Zhou, and H. Lu. Unsupervised feature selection using nonnegative spectral analysis. In *AAAI*, 2012. 3

[11] C. Liu, J. Yuen, and A. Torralba. Nonparametric scene parsing: label transfer via dense scene alignment. In *CVPR*, 2009. 1, 5, 6, 8

[12] J. Liu, M. Li, Q. Liu, H. Lu, and S. Ma. Image annotation via graph learning. *Pattern Recognition*, 42(2):218–228, 2009. 2

[13] J. Liu, B. Wang, M. Li, Z. Li, W. Ma, H. Lu, and S. Ma. Dual cross-media relevance model for image annotation. In *ACM MM*, 2007. 2

[14] S. Liu, S. Yan, T. Zhang, C. Xu, J. Liu, and H. Lu. Weakly supervised graph propagation towards collective image parsing. *Multimedia, IEEE Transactions on*, 14(2):361–373. 4

[15] X. Liu, S. Yan, J. Luo, J. Tang, Z. Huango, and H. Jin. Nonparametric label-to-region by search. In *CVPR*, 2010. 2, 6, 7

[16] J. P. Lopamudra Mukherjee, Vikas Singh. Scale invariant cosegmentation for image groups. In *CVPR*, 2011. 2, 6

[17] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60:91–110, 2004. 5

[18] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie. Objects in context. In *ICCV*, 2007. 1

[19] C. Russell, P. H. S. Torr, and P. Kohli. Associative hierarchical crfs for object class image segmentation. In *ICCV*, 2009. 2

[20] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE TPAMI*, 22(8):888 –905, 2000. 3

[21] 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*, 81:2–23, 2009. 2, 5, 6, 8

[22] R. Socher and L. Fei-fei. Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora. In *CVPR*, 2010. 1, 2

[23] J. Tang, S. Yan, R. Hong, G.-J. Qi, and T.-S. Chua. Inferring semantic concepts from community-contributed images and noisy tags. In *ACM MM*, 2009. 2

[24] J. Tighe and S. Lazebnik. Superparsing: Scalable nonparametric image parsing with superpixels. In *ECCV*, 2010. 6, 8

[25] A. Vezhnevets and J. M. Buhmann. Towards weakly supervised semantic segmentation by means of multiple instance and multitask learning. In *CVPR*, 2010. 1, 6, 7

[26] A. Vezhnevets, V. Ferrari, and J. Buhmann. Weakly supervised semantic segmentation with a multi-image model. In *ICCV*, 2011. 1, 2, 6, 7, 8

[27] A. Vezhnevets, V. Ferrari, and J. M. Buhmann. Weakly supervised structured output learning for semantic segmentation. In *CVPR*, 2012. 1, 2, 6, 8

[28] Y. Yang, Y. Yang, Z. Huang, H. T. Shen, and F. Nie. Tag localization with spatial correlations and joint group sparsity. In *CVPR*, 2011. 2