Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence

Semantic Graph Construction for Weakly-Supervised Image Parsing

Wenxuan Xie and **Yuxin Peng*** and **Jianguo Xiao**
Institute of Computer Science and Technology, Peking University, Beijing 100871, China
{xiewenxuan, pengyuxin, xiaojianguo}@pku.edu.cn

Abstract

We investigate weakly-supervised image parsing, i.e., assigning class labels to image regions by using image-level labels only. Existing studies pay main attention to the formulation of the weakly-supervised learning problem, i.e., how to propagate class labels from images to regions given an affinity graph of regions. Notably, however, the affinity graph of regions, which is generally constructed in relatively simpler settings in existing methods, is of crucial importance to the parsing performance due to the fact that the weakly-supervised parsing problem cannot be solved within a single image, and that the affinity graph enables label propagation among multiple images. In order to embed more semantics into the affinity graph, we propose novel criteria by exploiting the weak supervision information carefully, and develop two graphs: L_1 semantic graph and k-NN semantic graph. Experimental results demonstrate that the proposed semantic graphs not only capture more semantic relevance, but also perform significantly better than conventional graphs in image parsing.

Introduction

Image parsing (Liu et al. 2009a; 2012a; Yang et al. 2011; Han et al. 2012) is a fundamentally challenging problem aiming at assigning semantic labels to image pixels (Tighe and Lazebnik 2013). Being a sort of fine-grained image analysis, an effective image parsing is beneficial for many higher-level image understanding tasks, e.g., image editing (Shotton et al. 2009) and region-based image retrieval (Zhang et al. 2012). However, although the goal of image parsing is to label pixels, directly modeling pixels may result in unreliable predictions since a single pixel contains little information. In order to yield semantically consistent results, existing image parsing methods are generally based on image regions (aka, superpixels).

In the literature, most image parsing approaches suppose that a training dataset with superpixel-level labels is given and then either establish an appearance-based model which propagates labels from training superpixels to test superpixels (Yang et al. 2013) or resort to non-parametric methods to transfer labels from training images to query image (Liu, Yuen, and Torralba 2009). However, it is generally too laborious and time-consuming to annotate superpixel-level labels. Fortunately, due to the rapid spread of online photo sharing websites (e.g., Flickr), a large amount of images with user-provided image-level labels become available. These labels can be further refined by modeling sparsity and visual consistency (Zhu, Yan, and Ma 2010). In contrast to superpixel-level labels, it is more challenging to develop an image parsing algorithm based on image-level labels only. In this paper, such a problem is called weakly-supervised image parsing.

In traditional image parsing, labels are propagated from training superpixels to test superpixels; however, in weakly-supervised image parsing, the propagation is from images to superpixels. To handle such a weakly-supervised problem, several approaches have been proposed in the literature. As an example, (Liu et al. 2009a) has first proposed a bi-layer sparse coding model for uncovering how an image or superpixel could be reconstructed from superpixels of the entire image repository, and then used the learned relevance to facilitate label inference. What is more, (Liu et al. 2012a) has developed a weakly-supervised graph propagation model, where the final results can be directly inferred by simultaneously considering superpixel consistency, superpixel incongruity and the weak supervision information. It can be observed that, superpixel graphs are necessary and important to the aforementioned image parsing methods.

However, despite the effectiveness of the aforementioned approaches, the superpixel graphs are built up in relatively simpler settings. These approaches are generally based on the assumption that a given superpixel from an image can be sparsely reconstructed via the superpixels belonging to the images with common labels, and that the sparsely selected superpixels are relevant to the given superpixel. In order to state conveniently, we define *candidate superpixels* to be the set of superpixels which are possibly adjacent to a given superpixel, where the adjacency denotes a non-zero similarity in a superpixel graph. Under this definition, the candidate superpixels of the above approaches are those belonging to the images which have common labels with the image containing the given superpixel. Due to the large number of candidate superpixels in these approaches, the graph construction process tends to incur more semantically irrelevant superpixels and thus the parsing performance is degraded.

*Corresponding author.

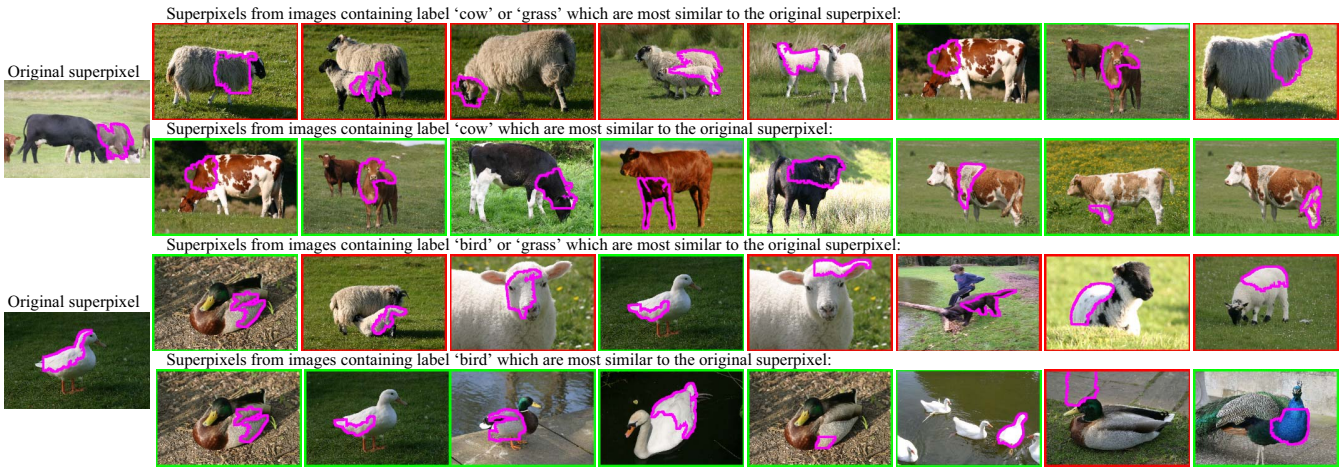

Figure 1: Illustrations of our motivation in constructing semantic graphs by reducing the number of *candidate superpixels*. An image is in a green box if its corresponding superpixel (which is bounded by a magenta closed curve) is semantically relevant to (i.e., has the same ground-truth label with) the original superpixel, otherwise it is in a red box.

Therefore, it is important to construct a superpixel graph with more semantic relevance. In order to handle this problem, we start from the following two empirical observations.

- *An ideal graph yields nearly perfect results.* Suppose there is an ideal graph, in which all pairs of semantically relevant superpixels are adjacent, and all pairs of semantically irrelevant superpixels are non-adjacent. The parsing accuracy with such an ideal graph is nearly 100%. Although the ideal graph is unavailable due to the fact that the ground-truth labels of superpixels are unknown in advance, it is worthwhile to construct a superpixel graph with more semantic relevance.
- *It is beneficial to reduce the number of candidate superpixels.* As shown by the illustrative examples in Fig. 1, through reducing the number of candidate superpixels, the graph can be made more descriptive.

Based on the above two observations, superpixel graph is a key factor to the parsing performance, and we can construct a descriptive graph by reducing the number of candidate superpixels. Concretely, we impose novel criteria on conventional graphs by exploiting the weak supervision information carefully, and develop two graphs: L_1 semantic graph and k-NN semantic graph. These two graphs are shown to be effective in weakly-supervised image parsing.

The rest of this paper is organized as follows. Section 2 presents a brief overview of related studies. The graph propagation approach to image parsing is introduced in Section 3 as a preliminary. Then, we introduce the proposed semantic graph construction approach in Section 4. In Section 5, the proposed method is evaluated on two standard datasets in image parsing. Finally, Section 6 concludes our paper.

Related Work

Image Parsing

The image parsing problem has received wide interests in the vision community, and numerous approaches have been proposed. Earlier studies mainly focus on modeling shapes (Winn and Jojic 2005; Chen et al. 2009). These methods, however, can only handle images either with a singe object or without occlusions between objects. Some other approaches are mostly based on discriminative learning techniques, e.g., conditional random field (Yuan, Li, and Zhang 2008), dense scene alignment (Liu, Yuen, and Torralba 2009) and deep learning (Farabet et al. 2013). All of these algorithms require pixel-level labels for training, however, which are very expensive to obtain in practice.

Besides the aforementioned approaches, there have been a few studies on weakly-supervised image parsing, where superpixel labels are propagated along a predefined graph. As a first attempt, (Liu et al. 2009a) has proposed a bi-layer sparse coding model for mining the relation between images and superpixels. The model has also been extended to a continuity-biased bi-layer sparsity formulation (Liu et al. 2012b). In (Liu et al. 2012a), a weakly-supervised graph propagation model is developed to directly infer the superpixel labels. Moreover, in (Liu et al. 2010), a multi-edge graph is established to simultaneously consider both images and superpixels, and is then used to obtain superpixel labels through a majority voting strategy. Different from the above approaches which pay main attention to the formulation of the weakly-supervised learning problem, our focus is constructing a superpixel graph with more semantic relevance by using the weak supervision information carefully.

Weakly-Supervised Image Segmentation

The weakly-supervised image segmentation task is similar to weakly-supervised image parsing. The only difference is that, images are split into a training set and a test set, and the aim is to infer the labels of test image pixels by exploiting only the image-level labels in the training set. In the literature, (Verbeek and Triggs 2007) has proposed to handle this task by using the Markov field aspect model. In (Vezhnevets and Buhmann 2010), multiple instance learning

and multi-task learning strategies are adopted. Multi-image model (Vezhnevets, Ferrari, and Buhmann 2011) and criteria on multiple feature fusion (Vezhnevets, Ferrari, and Buhmann 2012) have also been studied.

What is more, recent approaches include criteria on classification evaluation (Zhang et al. 2013a), weakly-supervised dual clustering (Liu et al. 2013) and probabilistic graphlet cut (Zhang et al. 2013b). However, in practice, due to the easy access of image-level labels on photo sharing websites such as Flickr, we assume all image-level labels are available in this paper, which is different from the aforementioned weakly-supervised image segmentation task.

Graph Construction

A number of methods have been proposed for graph construction, among which the most popular ones include sparse linear reconstruction (L_1) graph (Yan and Wang 2009), ϵ-ball graph and k-nearest neighbor (k-NN) graph. Recent studies are mostly based on the combinations and extensions of these graphs. For example, (Lu and Peng 2013) deals with latent semantic learning in action recognition through L_1 graph and hypergraph. In (He et al. 2013), a two-stage non-negative sparse representation has been proposed for face recognition. Furthermore, a k-NN sparse graph is applied to handle image annotation in (Tang et al. 2011).

However, different from conventional graph construction in either supervised or unsupervised setting, constructing a descriptive graph under weak supervision in this paper is a novel and interesting problem to handle.

Image Parsing by Graph Propagation

The proposed semantic graph construction approach is based on the weakly-supervised graph propagation model in (Liu et al. 2012a). As a preliminary, we introduce the graph propagation model by first defining some notations, and then present the formulation and solution. Due to the space limit, we only show the key steps here. Detailed derivations can be found in (Liu et al. 2012a).

Notation

Given an image collection $\{X_1, \ldots, X_m, \ldots, X_M\}$, where X_m denotes the m-th image, and its label information is denoted by an indicator vector $y_m = [y_m^1, \ldots, y_m^c, \ldots, y_m^C]^\top$, where $y_m^c = 1$ if X_m has the c-th label, and $y_m^c = 0$ otherwise. C denotes the number of classes, and $Y = [y_1, \ldots, y_m, \ldots, y_M]^\top$ denotes the image-level label collection. After image over-segmentation with a certain approach, e.g., SLIC (Achanta et al. 2012), X_m is represented by a set of superpixels $X_m = \{x_{m1}, \ldots, x_{mi}, \ldots, x_{mn_m}\}$, where n_m is the number of superpixels in X_m. x_{mi} stands for the i-th superpixel of X_m, and its corresponding label information is also denoted by an indicator vector $f_{mi} = [f_{mi}^1, \ldots, f_{mi}^c, \ldots, f_{mi}^C]^\top$, where $f_{mi}^c = 1$ if superpixel x_{mi} has the c-th label, and $f_{mi}^c = 0$ otherwise. Moreover, $N = \sum_{m=1}^{M} n_m$ denotes the number of superpixels in the image collection, and $F \in \mathbb{R}^{N \times C}$ denotes all the superpixel labels. In the weakly-supervised setting, all the image labels Y are given, and the superpixel labels F are to be inferred.

Formulation

First of all, given an $N \times N$ matrix W denoting the affinity graph of superpixels, the smoothness regularizer is shown as follows, which also resembles the idea of spectral clustering (Ng, Jordan, and Weiss 2001)

$$\operatorname{tr}(F^\top L F) \tag{1}$$

where L is a Laplacian matrix $L = D - W$, and D is the degree matrix of W. The smoothness regularizer enforces similar superpixels in feature space to share similar labels. Furthermore, the image-level supervision information can be formulated in the following form

$$\sum_m \sum_c |\max_{x_{mi} \in X_m} f_{mi}^c - y_m^c| \tag{2}$$

According to Eq. 2, if $y_m^c = 1$, at least one superpixel should interpret the label. Moreover, if $y_m^c = 0$, no superpixels will be assigned to that label, which is equivalent to require $\max f_{mi}^c = 0$. According to such equivalence, and due to the fact that the image-level label y_m^c can only be either 1 or 0, Eq. 2 can be rewritten in the following form

$$\begin{aligned} &\sum_m \sum_c (1 - y_m^c) h_c F^\top q_m \\ &+ \sum_m \sum_c y_m^c (1 - \max_{x_{mi} \in X_m} g_{mi} F h_c^\top) \end{aligned} \tag{3}$$

where h_c is a $1 \times C$ indicator vector whose all elements, except for the c-th element, are zeros, and q_m is an $N \times 1$ indicator vector whose all elements, except for those elements corresponding to the m-th image, are zeros. Moreover, g_{mi} is a $1 \times N$ vector whose elements corresponding to the i-th superpixel in X_m are ones and others are zeros. Through simultaneously considering Eq. 1 and Eq. 3, the final formulation is shown as follows

$$\begin{aligned} \min_F \ & \lambda \operatorname{tr}(F^\top L F) + \sum_m \sum_c (1 - y_m^c) h_c F^\top q_m \\ & + \sum_m \sum_c y_m^c (1 - \max_{x_{mi} \in X_m} g_{mi} F h_c^\top) \\ s.t. \ & F \geq 0, \ \ F e_1 = e_2 \end{aligned} \tag{4}$$

where λ is a positive parameter. It should be noted that, the equality $\sum_{c=1}^{C} f_{mi}^c = 1$ always holds due to $F e_1 = e_2$, where $e_1 = \mathbf{1}_{C \times 1}$, and $e_2 = \mathbf{1}_{N \times 1}$.

Solution

Eq. 4 can be efficiently solved via concave-convex programming (Yuille and Rangarajan 2003) iteratively. Let η be the subgradient of $l = [f_{m1}^c, \ldots, f_{mi}^c, \ldots, f_{mn_m}^c]^\top$, which is an $n_m \times 1$ vector and its i-th element is shown as follows

$$\eta_i = \begin{cases} \frac{1}{n_\alpha}, & f_{mi}^c = \max_j f_{mj}^c \ \text{ where } \ x_{mj} \in X_m \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

where n_α is the number of superpixels with the largest label value. According to (Yuille and Rangarajan 2003), Eq. 4

can be derived and further relaxed as the following quadratic programming problem

$$\begin{aligned}\min_{F}\ & \lambda \operatorname{tr}(F^\top LF) + \sum_m \sum_c (1 - y_m^c) h_c F^\top q_m \\ & + \sum_m \sum_c y_m^c (1 - h_c \beta U_m F h_c^\top) + \delta \|Fe_1 - e_2\|^2 \\ s.t.\ & F \geq 0\end{aligned} \tag{6}$$

where U_m is an $N \times N$ diagonal block matrix , whose diagonal elements are equal to q_m. β is a $C \times n_m$ matrix corresponding to X_m and $\beta_{mc} = \eta^\top$. Moreover, δ is a weighting parameter. To efficiently solve Eq. 6, the non-negative multiplicative updating procedure in (Liu et al. 2009b) is adopted, which facilitates the following element-wise updating rule

$$F_{ij} = F_{ij} \times \frac{[2\lambda WF + 2\delta e_2 e_1^\top + \sum_m \sum_c y_m^c U_m^\top \beta^\top h_c^\top h_c]_{ij}}{[2\lambda DF + 2\delta e_1 e_1^\top + \sum_m \sum_c (1 - y_m^c) q_m h_c]_{ij}} \tag{7}$$

Therefore, Eq. 4 can be solved via the above iterative steps, and the superpixel labels F are obtained as final results.

Semantic Graph Construction

Although the graph propagation method shown in the previous section is capable of inferring superpixel labels, the superpixel graph W is constructed by adopting relatively simpler settings. For example, the L_1 graph used in (Liu et al. 2012a) is built up by reconstructing each given superpixel via the superpixels belonging to the images with common labels. However, as a key factor to the final performance of image parsing, the superpixel graph W can be made more descriptive by exploiting the weak supervision information carefully. In this section, we present the construction process of two novel superpixel graphs, i.e., L_1 semantic graph and k-NN semantic graph.

L_1 Semantic Graph

Based on the two observations in Section 1, we propose to construct graphs with more semantic relevance by reducing the number of candidate superpixels. To begin with, we denote all the feature vectors of the superpixels as $Z \in \mathbb{R}^{d \times N}$, where d is the dimensionality of a feature vector. Furthermore, based on the image-level labels, all the superpixels belonging to images with the c-th label is denoted as $Z_c \in \mathbb{R}^{d \times N_c}$. According to the illustrative examples in Fig. 1, given a superpixel x_{mi} (belonging to image X_m) whose ground-truth label is c and whose corresponding feature vector is denoted as p_{mi}, using Z_c as candidate superpixels can provide better results than using Z or other Z_j, where $j \neq c$. This fact can be easily verified due to the following reasons: 1) Since all the superpixels which belong to Z but not Z_c are semantically irrelevant to p_{mi}, it is beneficial to reconstruct p_{mi} by excluding these superpixels, and thus using Z_c may yield better results than Z; 2) Z_c contains more semantically relevant superpixels and fewer irrelevant superpixels to p_{mi} than other Z_j, where $j \neq c$. Therefore, our aim is to find the most appropriate candidate superpixels for each superpixel.

Notably, this is a paradox, since we can precisely obtain Z_c according to the ground-truth label of x_{mi} (i.e., c) and thus provide better reconstruction results. However, the superpixel label c is to be inferred and unknown in advance. In order to handle this problem, we propose criteria in selecting Z_c according to the sparse reconstruction formulation, whose objective function is shown as follows.

$$\begin{aligned}&\min_{\varphi_j} \|\varphi_j\|_1 \\ &s.t.\ Z_j \varphi_j = p_{mi},\ y_m^j = 1,\ z_{mi} = 0\end{aligned} \tag{8}$$

where φ_j denotes the coefficients of superpixels belonging to Z_j in reconstructing p_{mi}, and z_{mi} denotes the coefficient of p_{mi} (i.e., p_{mi} cannot be used to reconstruct itself). Given that the image X_m may contain multiple labels (i.e., its label $y_m^j = 1$ may hold true for different j), and that the label c of p_{mi} is unknown, we aim to select the candidate superpixels according to the criterion shown as follows.

$$\begin{aligned}&\min_{j,\varphi} \|\varphi\|_1 \\ &s.t.\ \varphi = \arg\min_{\varphi_j} \|\varphi_j\|_1,\ Z_j \varphi_j = p_{mi}, \\ &y_m^j = 1, z_{mi} = 0,\ j \in \{1, \ldots, C\}\end{aligned} \tag{9}$$

Eq. 9 is optimized in two steps: 1) Compute φ_j for all possible values of j which satisfy $y_m^j = 1$ (i.e., all labels of the image containing the given superpixel p_{mi}); 2) Select the specific φ_j which minimizes Eq. 9 as the final result.

According to Eq. 9, we select the candidate superpixels whose corresponding reconstruction coefficient vector φ has the smallest L_1 norm. Eq. 9 makes sense due to the fact that, the L_1 norm of φ indicates the correlation between Z_j and p_{mi}. If the correlation coefficient between Z_j and p_{mi} is adequately large, all the elements in φ are non-negative and thus $\|\varphi\|_1$ remains to be small; however, if the correlation coefficient is small, there may be both positive and negative elements in φ and thus $\|\varphi\|_1$ is large. Hence, we select Z_j with the smallest $\|\varphi\|_1$, which is most correlated with p_{mi} and is assumed to be the desired Z_c.

As a consequence, $|\varphi|$ (i.e., the absolute value of φ) is used as the similarity between superpixels (Yan and Wang 2009), and thus the affinity graph W is constructed. To ensure the symmetry, we assign $W = \frac{1}{2}(W + W^\top)$, and further use it as the L_1 semantic graph in this paper.

k-NN Semantic Graph

Besides L_1 semantic graph, we can similarly construct k-NN semantic graph through reducing the number of candidate superpixels. Since using Z as candidate superpixels is always a suboptimal choice, we focus on selecting candidate superpixels from Z_j where $j \in \{1, \ldots, C\}$. Given a superpixel x_{mi} (belonging to image X_m) whose feature vector is p_{mi}, we begin by denoting S_j as the set of k-NN superpixels of p_{mi} in Z_j, and S_j^{cp} as the set of k-NN superpixels in Z_j^{cp}, where Z_j^{cp} is the complementary set of Z_j, i.e., $Z_j^{cp} = Z \backslash Z_j$. Based on these notations, we select the k-NN superpixels of p_{mi} according to the following criterion.

$$\begin{aligned}&\min_{j,S_j} \sum_{a=1}^{k} \sum_{b=1}^{k} \operatorname{sim}(S_{ja}, S_{jb}^{cp}) \\ &s.t.\ S_{ja} \in S_j,\ S_{jb}^{cp} \in S_j^{cp},\ y_m^j = 1,\ j \in \{1, \ldots, C\}\end{aligned} \tag{10}$$

where S_{ja} and S_{jb}^{cp} are superpixels belonging to sets S_j and S_j^{cp}, respectively. Moreover, $\mathrm{sim}(\cdot,\cdot)$ denotes a similarity measure of two feature vectors of superpixels. Similarly with Eq. 9, Eq. 10 is optimized by first enumerating all possible j and then select the specific S_j which minimizes Eq. 10. According to Eq. 10, we select S_j as the k-NN superpixels of p_{mi}, where the sum of pairwise similarity between superpixels in S_j and S_j^{cp} is minimized. Eq. 10 makes sense due to the following reasons. Generally, superpixels with the same labels tend to be visually similar, whereas the similarity between superpixels belonging to different classes tends to be small. Through minimizing the pairwise similarity between superpixels in S_j and S_j^{cp}, the superpixels in the selected S_j are likely to have the same label with p_{mi}.

For example, given an image X_m with labels 'grass' and 'bird', we denote a 'grass' superpixel and a 'bird' superpixel in X_m as p_{grs} and p_{brd} respectively. Moreover, candidate superpixels Z_{grs}, Z_{grs}^{cp}, Z_{brd} and Z_{brd}^{cp} are defined accordingly. Given p_{grs}, since 'grass' superpixels may appear as neighbors in both Z_{brd} (superpixels in 'bird' images) and Z_{brd}^{cp} (superpixels in 'non-bird' images), the pairwise similarity between superpixels in S_{brd} and S_{brd}^{cp} is relatively large. In contrast, the pairwise similarity between superpixels in S_{grs} and S_{grs}^{cp} is small since 'grass' superpixels are absent in S_{grs}^{cp}. Therefore, the selected set of neighbors for p_{grs} is S_{grs} but not S_{brd}. Moreover, the same applies to p_{brd}, where S_{brd} is chosen as its k-NN superpixels.

As a result, after selecting neighbors for each superpixel by reducing the number of candidate superpixels according to Eq. 10, the affinity graph W is constructed. To ensure its symmetry, we assign $W = \frac{1}{2}(W + W^\top)$, and further use it as the k-NN semantic graph in this paper.

Experiments

In this section, we evaluate the performance of the proposed semantic graphs in weakly-supervised image parsing.

Experimental Setup

We conduct experiments on two standard datasets: PASCAL VOC'07 (PASCAL for short) (Everingham et al. 2010) and MSRC-21 (Shotton et al. 2009). Both datasets contain 21 different classes and are provided with pixel-level labels, which are used to evaluate the performance measured by classification accuracy. In the weakly-supervised image parsing task, we assume all the image-level labels are known for both training and test set, i.e., 632 images in PASCAL dataset and 532 images in MSRC-21 dataset (Shotton, Johnson, and Cipolla 2008). Moreover, we adopt SLIC algorithm (Achanta et al. 2012) to obtain superpixels for each image, and represent each superpixel by the bag-of-words model while using SIFT (Lowe 2004) as the local descriptor. To present fair comparisons, we adopt the same parameters for the graph propagation model shown in Eq. 4. In the experiments, we discover that the parameter k in all k-NN-based graphs are relatively insensitive to the performance, and we set $k = 20$ empirically.

Besides comparing with the state-of-the-arts (Liu et al. 2009a; 2012a), we mainly focus on the comparisons among the following three types of graphs: 1) Original graph (OG), where all superpixels are candidates for a given superpixel; 2) Label intersection graph (LIG), where all the candidate superpixels belong to images which have at least one common label with the image containing the given superpixel; 3) Semantic graph (SG), where the candidate superpixels are derived by criteria shown in Eq. 9 or Eq. 10. Based on L_1 graph and k-NN graph, there are six graphs in total, i.e., L_1 OG, L_1 LIG, L_1 SG, k-NN OG, k-NN LIG and k-NN SG. Notably, besides the parsing accuracy, we also measure the semantic relevance captured by a graph with a percentage value calculated as follows

$$\text{percentage} = \frac{\#(\text{adjacent superpixels with the same label})}{\#(\text{adjacent superpixels})} \tag{11}$$

where the term *adjacent superpixels* denotes a pair of superpixels whose similarity in a graph is non-zero.

Empirical Results

The per-class accuracies on PASCAL dataset and MSRC-21 dataset are listed in Table 1 and Table 2, respectively. It can be observed that trends on both datasets are similar, where L_1 SG and k-NN SG achieve the best performances among all L_1-based graphs and all k-NN-based graphs, respectively. Since all the settings are the same except for the superpixel graphs, the results indicate the effectiveness of the proposed semantic graphs in weakly-supervised image parsing. Please note that, the reason why the performances of (Liu et al. 2012a) and L_1 LIG slightly differ is that, (Liu et al. 2012a) utilizes an additional superpixel distance graph.

Furthermore, we report the semantic relevance captured by different graphs along with the corresponding mean parsing accuracy on PASCAL dataset and MSRC-21 dataset in Table 3 and Table 4, respectively. We observe from these two tables that, generally, the more semantic relevance captured by the graph, the better the parsing accuracy is. However, although the semantic relevance captured by L_1 LIG (k-NN LIG) is much more than L_1 OG (k-NN OG), there is nearly no improvement in parsing accuracy on MSRC-21 dataset. It may be due to the fact that the label intersection graphs only discard the adjacencies of superpixels whose corresponding images have no common labels. These adjacencies do not affect the parsing accuracy much, since the inferred label of a superpixel is constrained to be one of the labels of its corresponding image. In contrast, L_1 SG (k-NN SG) further improves the percentage of semantically adjacent superpixels, which is beneficial for the final performance.

Notably, the criteria shown in Eq. 9 and Eq. 10 are to select candidate superpixels belonging to images containing a specific label, which can be viewed as initial predictions for all superpixel labels, although these predictions are not used to evaluate the parsing accuracy directly. However, we can still calculate an accuracy for these predictions. We empirically discover that these predictions achieve relatively lower results. For example, on MSRC-21 dataset, the accuracy achieved by initial predictions in constructing L_1 SG (k-NN SG) is 60% (64%), whereas the accuracy of label propagation with L_1 SG (k-NN SG) is 62% (73%). These results

Table 1: Accuracies (%) of the proposed semantic graphs for individual classes on PASCAL dataset, in comparison with other methods. The last column shows the mean accuracy over all classes.

Methods	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	bkgd	mean
(Liu et al. 2009a)	24	25	40	25	32	35	27	45	16	49	24	32	13	25	**56**	28	17	16	33	18	82	32
(Liu et al. 2012a)	28	20	52	28	46	**41**	**39**	60	25	68	25	35	17	35	**56**	36	46	17	31	20	65	38
L_1 OG	10	8	8	10	17	13	12	12	6	8	7	15	5	15	38	15	10	3	20	7	**85**	15
L_1 LIG	6	12	63	30	47	22	16	58	8	53	7	39	10	18	30	27	58	4	46	26	66	31
L_1 SG	16	14	75	43	42	34	29	**64**	7	57	15	**46**	**38**	**43**	29	**39**	**83**	6	**58**	21	59	39
k-NN OG	20	16	16	16	12	16	14	15	15	22	11	13	14	13	25	17	24	16	11	20	76	19
k-NN LIG	41	20	58	41	**48**	30	38	44	**31**	42	**31**	36	28	26	37	30	50	**25**	42	**40**	47	37
k-NN SG	**85**	**55**	**87**	**45**	42	31	34	57	21	**81**	23	16	6	11	42	31	72	24	49	**40**	41	**42**

Table 2: Accuracies (%) of the proposed semantic graphs for individual classes on MSRC-21 dataset, in comparison with other methods. The last column shows the mean accuracy over all classes.

Methods	bldg	grass	tree	cow	sheep	sky	plane	water	face	car	bike	flower	sign	bird	book	chair	road	cat	dog	body	boat	mean
(Liu et al. 2012a)	70	92	49	10	10	83	36	**82**	62	20	52	98	88	48	98	70	75	95	76	**43**	**23**	61
L_1 OG	68	93	55	19	11	94	27	74	55	21	59	96	84	52	98	70	76	88	67	42	17	60
L_1 LIG	64	91	48	7	8	92	30	78	59	16	48	98	92	53	**99**	73	76	97	78	35	21	60
L_1 SG	**84**	**95**	42	11	13	91	26	77	54	19	59	97	87	56	98	91	53	**99**	86	39	16	62
k-NN OG	74	94	**64**	29	12	**94**	36	75	65	40	81	96	83	56	**99**	77	**78**	93	73	34	17	65
k-NN LIG	71	92	61	25	9	92	33	75	**67**	39	82	98	90	54	98	85	73	**99**	87	32	10	65
k-NN SG	49	82	45	**59**	**51**	90	**78**	68	66	**68**	**98**	**99**	**94**	**84**	**99**	**99**	48	**99**	**98**	30	20	**73**

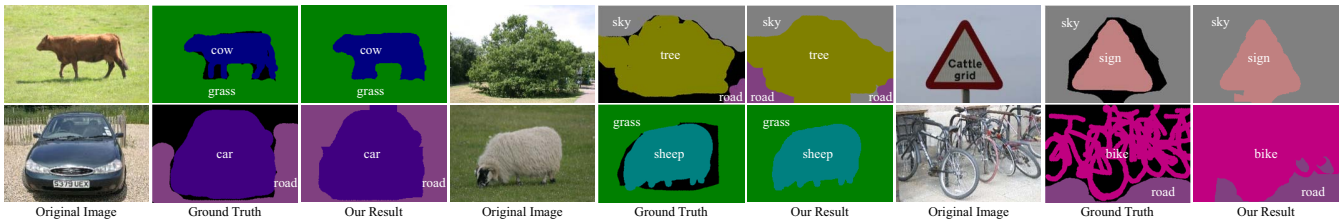

Figure 2: Some example results for image parsing by graph propagation using k-NN semantic graph (i.e., our result) in comparison with the ground-truth on MSRC-21 dataset.

Table 3: Percentages (%) of semantically relevant superpixels in different graphs along with the corresponding mean parsing accuracies (%) on PASCAL dataset.

Graphs	Percentage	Accuracy
L_1 OG	7	15
L_1 LIG	23	31
L_1 SG	25	39
k-NN OG	11	19
k-NN LIG	34	37
k-NN SG	**38**	**42**

Table 4: Percentages (%) of semantically relevant superpixels in different graphs along with the corresponding mean parsing accuracies (%) on MSRC-21 dataset.

Graphs	Percentage	Accuracy
L_1 OG	18	60
L_1 LIG	34	60
L_1 SG	35	62
k-NN OG	33	65
k-NN LIG	52	65
k-NN SG	**59**	**73**

show the effectiveness of the whole framework. Moreover, some example results for image parsing by graph propagation using k-NN SG in comparison with the ground-truth on MSRC-21 dataset are shown in Fig. 2.

Conclusion

In this paper, we focus on the graph construction in weakly-supervised image parsing. Due to the weak supervision information, the semantic relevance captured by the superpixel graph is crucial to the final performance. In order to build up graphs which can capture more semantic relevance in the weakly-supervised setting, we propose criteria in reducing the number of candidate superpixels, and develop two novel graphs: L_1 semantic graph and k-NN semantic graph. As shown in the experiments, the criteria used in superpixel graph construction yield significant performance improvement in image parsing. Moreover, as a general framework, the proposed approach is suitable for other weakly-supervised learning tasks besides image parsing.

Acknowledgments

This work was supported by National Hi-Tech Research and Development Program (863 Program) of China under Grants 2014AA015102 and 2012AA012503, National Natural Science Foundation of China under Grant 61371128, and Ph.D. Programs Foundation of Ministry of Education of China under Grant 20120001110097.

References

Achanta, R.; Shaji, A.; Smith, K.; Lucchi, A.; Fua, P.; and Süsstrunk, S. 2012. Slic superpixels compared to state-of-the-art superpixel methods. *TPAMI* 34(11):2274–2281.

Chen, Y.; Zhu, L.; Yuille, A.; and Zhang, H. 2009. Unsupervised learning of probabilistic object models (poms) for object classification, segmentation, and recognition using knowledge propagation. *TPAMI* 31(10):1747–1761.

Everingham, M.; Van Gool, L.; Williams, C.; Winn, J.; and Zisserman, A. 2010. The PASCAL visual object classes (VOC) challenge. *IJCV* 88(2):303–338.

Farabet, C.; Couprie, C.; Najman, L.; and LeCun, Y. 2013. Learning hierarchical features for scene labeling. *TPAMI* 35(8):1915–1929.

Han, Y.; Wu, F.; Shao, J.; Tian, Q.; and Zhuang, Y. 2012. Graph-guided sparse reconstruction for region tagging. In *CVPR*, 2981–2988.

He, R.; Zheng, W.-S.; Hu, B.-G.; and Kong, X.-W. 2013. Two-stage nonnegative sparse representation for large-scale face recognition. *TNNLS* 24(1):35–46.

Liu, X.; Cheng, B.; Yan, S.; Tang, J.; Chua, T. S.; and Jin, H. 2009a. Label to region by bi-layer sparsity priors. In *ACM MM*, 115–124.

Liu, X.; Yan, S.; Yan, J.; and Jin, H. 2009b. Unified solution to nonnegative data factorization problems. In *ICDM*, 307–316.

Liu, D.; Yan, S.; Rui, Y.; and Zhang, H.-J. 2010. Unified tag analysis with multi-edge graph. In *ACM MM*, 25–34.

Liu, S.; Yan, S.; Zhang, T.; Xu, C.; Liu, J.; and Lu, H. 2012a. Weakly supervised graph propagation towards collective image parsing. *TMM* 14(2):361–373.

Liu, X.; Yan, S.; Cheng, B.; Tang, J.; Chua, T.-S.; and Jin, H. 2012b. Label-to-region with continuity-biased bi-layer sparsity priors. *ACM TOMCCAP* 8(4):50.

Liu, Y.; Liu, J.; Li, Z.; Tang, J.; and Lu, H. 2013. Weakly-supervised dual clustering for image semantic segmentation. In *CVPR*, 2075–2082.

Liu, C.; Yuen, J.; and Torralba, A. 2009. Nonparametric scene parsing: Label transfer via dense scene alignment. In *CVPR*, 1972–1979.

Lowe, D. G. 2004. Distinctive image features from scale-invariant keypoints. *IJCV* 60(2):91–110.

Lu, Z., and Peng, Y. 2013. Latent semantic learning with structured sparse representation for human action recognition. *PR* 46(7):1799–1809.

Ng, A. Y.; Jordan, M. I.; and Weiss, Y. 2001. On spectral clustering: Analysis and an algorithm. *NIPS* 14:849–856.

Shotton, J.; Winn, J.; Rother, C.; and Criminisi, A. 2009. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *IJCV* 81(1):2–23.

Shotton, J.; Johnson, M.; and Cipolla, R. 2008. Semantic texton forests for image categorization and segmentation. In *CVPR*, 1–8.

Tang, J.; Hong, R.; Yan, S.; Chua, T.-S.; Qi, G.-J.; and Jain, R. 2011. Image annotation by knn-sparse graph-based label propagation over noisily tagged web images. *ACM TIST* 2(2):14.

Tighe, J., and Lazebnik, S. 2013. Superparsing: scalable nonparametric image parsing with superpixels. *IJCV* 101(2):329–349.

Verbeek, J., and Triggs, B. 2007. Region classification with markov field aspect models. In *CVPR*, 1–8.

Vezhnevets, A., and Buhmann, J. M. 2010. Towards weakly supervised semantic segmentation by means of multiple instance and multitask learning. In *CVPR*, 3249–3256.

Vezhnevets, A.; Ferrari, V.; and Buhmann, J. M. 2011. Weakly supervised semantic segmentation with a multi-image model. In *ICCV*, 643–650.

Vezhnevets, A.; Ferrari, V.; and Buhmann, J. M. 2012. Weakly supervised structured output learning for semantic segmentation. In *CVPR*, 845–852.

Winn, J., and Jojic, N. 2005. Locus: Learning object classes with unsupervised segmentation. In *ICCV*, 756–763.

Yan, S., and Wang, H. 2009. Semi-supervised learning by sparse representation. In *SDM*, 792–801.

Yang, Y.; Yang, Y.; Huang, Z.; Shen, H. T.; and Nie, F. 2011. Tag localization with spatial correlations and joint group sparsity. In *CVPR*, 881–888.

Yang, Y.; Huang, Z.; Yang, Y.; Liu, J.; Shen, H. T.; and Luo, J. 2013. Local image tagging via graph regularized joint group sparsity. *PR* 46(5):1358–1368.

Yuan, J.; Li, J.; and Zhang, B. 2008. Scene understanding with discriminative structured prediction. In *CVPR*, 1–8.

Yuille, A. L., and Rangarajan, A. 2003. The concave-convex procedure. *Neural Computation* 15(4):915–936.

Zhang, D.; Islam, M. M.; Lu, G.; and Sumana, I. J. 2012. Rotation invariant curvelet features for region based image retrieval. *IJCV* 98(2):187–201.

Zhang, K.; Zhang, W.; Zheng, Y.; and Xue, X. 2013a. Sparse reconstruction for weakly supervised semantic segmentation. In *IJCAI*, 1889–1895.

Zhang, L.; Song, M.; Liu, Z.; Liu, X.; Bu, J.; and Chen, C. 2013b. Probabilistic graphlet cut: Exploiting spatial structure cue for weakly supervised image segmentation. In *CVPR*, 1908–1915.

Zhu, G.; Yan, S.; and Ma, Y. 2010. Image tag refinement towards low-rank, content-tag prior and error sparsity. In