Weakly-Supervised Image Parsing via Constructing Semantic Graphs and Hypergraphs

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

ABSTRACT

In this paper, we address the problem of weakly-supervised image parsing, whose aim is to automatically determine the class labels of image regions given image-level labels only. In the literature, 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 image parsing problem cannot be handled within a single image, and that the affinity graph facilitates label propagation among multiple images. Therefore, in contrast to existing methods, we focus on how to make the affinity graph more descriptive through embedding more semantics into it. We develop two novel graphs by leveraging the weak supervision information carefully: 1) Semantic graph, which is established upon a conventional graph by utilizing the proposed weakly-supervised criteria; 2) Semantic hypergraph, which explores both intra-image and inter-image high-order semantic relevance. Experimental results on two standard datasets demonstrate that the proposed semantic graphs and hypergraphs not only capture more semantic relevance, but also perform significantly better than conventional graphs in image parsing. More remarkably, due to the complementariness among the proposed semantic graphs and hypergraphs, the combination of them shows even more promising results.

Categories and Subject Descriptors

I.4.6 [**Image Processing and Computer Vision**]: Segmentation—*pixel classification*

General Terms

Algorithms, Experimentation, Performance

*Corresponding author.

MM'14, November 3–7, 2014, Orlando, Florida, USA.

http://dx.doi.org/10.1145/2647868.2654910.

Keywords

Weakly-supervised image parsing; Semantic graph construction; Semantic hypergraph construction

1. INTRODUCTION

Image parsing, whose aim is to assign semantic labels to image regions [28], is a fundamentally challenging problem [17, 16, 36, 12]. Being a sort of fine-grained image analysis, an effective image parsing system facilitates many higher-level image understanding tasks, e.g., image editing [26] and region-based image retrieval [39]. However, although the goal of image parsing is to classify pixels, directly modeling pixels may lead to unreliable predictions due to the fact that a single pixel contains little information. In order to yield semantically consistent results, existing image parsing approaches are generally based on image regions (aka, superpixels).

In the literature, most image parsing methods 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 [35] or resort to non-parametric methods to transfer labels from training images to query images [14]. However, it is generally too laborious and time-consuming to annotate superpixel-level labels manually. Fortunately, thanks 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 visual consistency and error sparsity [43]. 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 learning problem, several approaches have been proposed in the literature. For example, [17] 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, [16] 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

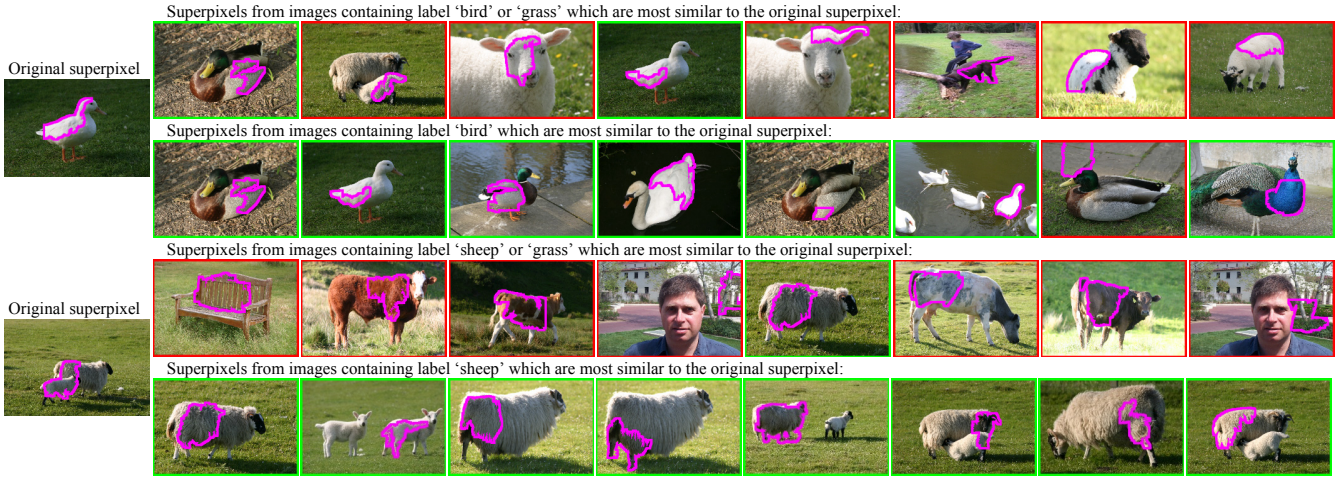


Figure 1: Illustrations of our motivation in constructing semantic graphs by reducing the number of *candidate superpixels*. An image is in a green box if its corresponding superpixel (which is bounded by a magenta closed curve) is semantically relevant to (i.e., has the same ground-truth label with)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

graphs are necessary and important to the aforementioned image parsing methods.

However, despite the effectiveness of the aforementioned approaches, the superpixel graphs are constructed in relatively simpler settings. These approaches are mainly 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.

Therefore, it is crucial to construct a superpixel graph with more semantic relevance. In order to handle this task, we start from the following three 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 a graph is 99% (where the 1% loss lies in the error of over-segmentation).
- *Reducing the number of candidate superpixels is beneficial.* As shown by the illustrative examples in Fig. 1, by reducing the number of candidate superpixels, the graph can be made more descriptive. If all the candidate superpixels are selected correctly, the parsing accuracy on MSRC-21 dataset is 89%.
- *It is important to explore high-order semantic relevance.* For example, superpixels (usually more than two) that are both visually similar and spatially adjacent within an image tend to be semantically relevant.

It can be concluded from the *first* observation that, 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. Based on the *second* observation, 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 semantic graphs. Moreover, as shown by the *third* observation, we may further enrich the superpixel graph by resorting to hypergraphs [42]. In this paper, semantic hypergraphs are constructed to model both intra-image and inter-image high-order relevance by leveraging the weak supervision information.

The rest of the paper is organized as follows. A brief overview of related studies is presented in Section 2. In Section 3, the graph propagation approach to weakly-supervised image parsing is introduced as a preliminary. Then, we present the proposed semantic graph construction approach and semantic hypergraph construction approach in Section 4 and Section 5, respectively. In Section 6, we show how the aforementioned graphs and hypergraphs are combined by using random walk. The proposed methods are evaluated on two standard datasets in image parsing in Section 7. Finally, Section 8 draws the conclusions.

2. RELATED WORK

In this section, we review some studies related to the proposed approach in the following aspects: image parsing, weakly-supervised image segmentation, graph construction and hypergraph construction.

2.1 Image Parsing

The image parsing problem has received wide interests in the vision community, and numerous methods have been proposed. Earlier studies mainly focus on modeling shapes [33, 5]. These methods, however, can only handle images either with a singe object or without occlusions between objects. Some other approaches are mostly based on discrim-

inative learning techniques, e.g., conditional random field [37], dense scene alignment [14] and deep learning [8]. 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, [17] 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 [18]. In [16], a weakly-supervised graph propagation model is developed to directly infer the superpixel labels. Moreover, in [15], 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 to construct a superpixel graph with more semantic relevance by using the weak supervision information carefully.

2.2 Weakly-Supervised Image Segmentation

The weakly-supervised image segmentation task is similar to weakly-supervised image parsing, where the only difference lies in 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, [29] has proposed to handle this task by using the Markov field aspect model. In [30], multiple instance learning and multi-task learning strategies are adopted. Multi-image model [31] and criteria on multiple feature fusion [32] have also been studied.

What is more, recent approaches contain criteria on probabilistic graphlet cut [41], weakly-supervised dual clustering [20] and classifier evaluation [40]. 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.

2.3 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 [34], ϵ-ball graph and k-nearest neighbor (k-NN) graph. Recent studies are mostly based on the combinations and extensions of these graphs. As an example, [44] has proposed to handle semi-supervised learning with a non-negative low rank and sparse graph. In [13], 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 [27].

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 task to handle.

2.4 Hypergraph Construction

A hypergraph is a graph in which an edge can connect more than two vertices [4]. Similar to simple graphs, hypergraphs can also be used in learning tasks [42]. For example, [22] has proposed to handle latent semantic learning in action recognition through sparse coding and hypergraph regularization. Moreover, hypergraphs enable visual-textual joint relevance learning for tag-based social image search [11]. In [10], hypergraphs are exploited in codebook learning for image classification.

Besides the aforementioned approaches and applications, we propose to model high-order relevance among superpixels in weakly-supervised image parsing by resorting to hypergraphs in this paper.

3. WEAKLY-SUPERVISED IMAGE PARSING BY GRAPH PROPAGATION

In order to handle weakly-supervised image parsing, the proposed semantic graph and hypergraph construction approaches are based on the weakly-supervised graph propagation model in [16]. As a preliminary, we begin by formally defining the problem, and then present the formulation and solution. In this paper, we only show the key steps here. Please refer to [16] for detailed derivations.

3.1 Problem Definition

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 the image-level label collection is denoted as $Y = [y_1, \ldots, y_m, \ldots, y_M]^\top$. After image over-segmentation with a certain approach, e.g., SLIC [1], 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 total 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.

3.2 Formulation

In a traditional label propagation formulation, both the graph and the labeled information are considered; however, in a weakly-supervised label propagation formulation, both the graph and the weak supervision information are taken into account. First of all, given an $N \times N$ matrix W denoting the affinity graph of superpixels, we can obtain the smoothness regularizer [3] as follows

$$T_1 = \mathrm{tr}(F^\top L F) \tag{1}$$

where L is a Laplacian matrix defined as $L = D - W$, and D is the degree matrix of W. The smoothness regularizer enforces similar superpixels in feature space to share similar labels, which also resembles the idea of spectral clustering [23]. Furthermore, the image-level supervision information can be formulated in the following form

$$T_2 = \sum_m \sum_c |\max_{x_{mi} \in X_m} f_{mi}^c - y_m^c| \tag{2}$$

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

$$\begin{aligned} T_3 = &\sum_m \sum_c (1 - y_m^c) h_c F^\top q_m \\ &+ \sum_m \sum_c y_m^c (1 - \max_{x_{mi} \in X_m} g_{mi} F h_c^\top) \end{aligned} \tag{3}$$

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

$$\begin{aligned} \min_F \ & \lambda \operatorname{tr}(F^\top L F) + \sum_m \sum_c (1 - y_m^c) h_c F^\top q_m \\ &+ \sum_m \sum_c y_m^c (1 - \max_{x_{mi} \in X_m} g_{mi} F h_c^\top) \\ s.t. \ & F \geq 0, \quad F\mathbf{e}_1 = \mathbf{e}_2 \end{aligned} \tag{4}$$

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 $F\mathbf{e}_1 = \mathbf{e}_2$, where $\mathbf{e}_1 = \mathbf{1}_{C \times 1}$, and $\mathbf{e}_2 = \mathbf{1}_{N \times 1}$.

3.3 Solution

Eq. 4 can be efficiently solved via concave-convex programming [38] 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 [38], Eq. 4 can be derived and further relaxed as the following quadratic programming problem

$$\begin{aligned} \min_F \ & \lambda \operatorname{tr}(F^\top L F) + \sum_m \sum_c (1 - y_m^c) h_c F^\top q_m \\ &+ \sum_m \sum_c y_m^c (1 - h_c \beta U_m F h_c^\top) + \gamma \| F\mathbf{e}_1 - \mathbf{e}_2 \|^2 \\ s.t. \ & F \geq 0 \end{aligned} \tag{6}$$

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, γ is a weighting parameter. To efficiently solve Eq. 6, the non-negative multiplicative updating procedure in [19] is adopted, which facilitates the following element-wise updating rule

$$F_{ij} = F_{ij} \times \frac{[2\lambda W F + 2\gamma \mathbf{e}_2 \mathbf{e}_1^\top + \sum_m \sum_c y_m^c U_m^\top \beta^\top h_c^\top h_c]_{ij}}{[2\lambda D F + 2\gamma \mathbf{e}_1 \mathbf{e}_1^\top + \sum_m \sum_c (1 - y_m^c) q_m h_c]_{ij}} \tag{7}$$

Therefore, Eq. 4 can be solved by alternatively updating β and F according to Eq. 5 and Eq. 7, respectively. After convergence, the superpixel labels F are obtained as the final results of image parsing.

4. 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 [16] 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 weakly-supervised image parsing (as shown by the *first* observation in Section 1), the superpixel graph W can be made more descriptive by leveraging the weak supervision information carefully. In this section, we focus on the construction process of a novel superpixel graph, i.e., k-NN semantic graph.

4.1 Preliminaries

Based on the *second* observation 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 $j \neq c$.

The aforementioned 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 represent 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 a descriptive graph. However, the superpixel label c is to be inferred and unknown in advance. In order to handle this problem, we develop k-NN semantic graphs, which is based on the proposed criteria in selecting Z_c in conventional k-NN graphs.

4.2 k-NN Semantic Graph

In this subsection, we present the construction process of k-NN semantic graphs. 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.

$$\begin{aligned} &\min_{j, S_j} \sum_{a=1}^{k} \sum_{b=1}^{k} \operatorname{sim}(S_{ja}, S_{jb}^{cp}) \\ &s.t. \ \ S_{ja} \in S_j, \ S_{jb}^{cp} \in S_j^{cp}, \ y_m^j = 1, \ j \in \{1, \ldots, C\} \end{aligned} \tag{8}$$

where S_{ja} and S_{jb}^{cp} are superpixels belonging to sets S_j and S_j^{cp}, respectively. Moreover, $\operatorname{sim}(\cdot, \cdot)$ denotes a similarity measure of two feature vectors of superpixels. According to Eq. 8, 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. 8 is optimized in two steps: 1) Enumerate S_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 S_j which minimizes Eq. 8 (i.e., the sum of pairwise similarity) as the final result. Solving Eq. 8 requires $O(CN + Ck^2)$ for a single superpixel. Since C and k are much smaller than N, the complexity is linear with respect to N (i.e., the total number of superpixels).

As a consequence, after selecting neighbors for each superpixel in the entire image repository by reducing the number of candidate superpixels based on Eq. 8, the affinity graph W is constructed. We further assign $W = \frac{1}{2}(W + W^\top)$ to ensure its symmetry, and use it as the k-NN semantic graph in this paper.

4.3 Interpretation

Eq. 8 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.

It should be noted that, although we focus only on constructing k-NN semantic graphs in this paper, the idea of constructing a superpixel graph with more semantic relevance by reducing the number of candidate superpixels is also applicable to other types of graphs, e.g., L_1 graphs. What is more, Eq. 8 is not the only choice for constructing k-NN semantic graphs, and can be readily substituted by other feasible criteria.

More notably, our aim is to show that reducing the number of candidate superpixels is beneficial. We have conducted experiments and observed that L_1 semantic graph (with some criteria) performs better than original L_1 graph. However, constructing L_1 graph is time-consuming. Therefore, our experiments are mainly based on k-NN (semantic) graphs, which are more efficient to construct.

5. SEMANTIC HYPERGRAPH CONSTRUCTION

Besides constructing semantic graphs which only contain second-order relevance, we also model high-order semantic relevance by resorting to hypergraphs (as shown by the *third* observation in Section 1). We begin by presenting some notations along with our motivation in exploiting hypergraphs. After that, we introduce the proposed approaches to establishing intra-image hyperedges and inter-image hyperedges, respectively. Finally, we discuss the complementariness among the semantic graphs and hypergraphs.

5.1 Notation and Motivation

A hypergraph $G(A, \mathcal{E})$ consists of a vertex set A and a hyperedge set $\mathcal{E}$ [42]. Each hyperedge e is a subset of the vertex set A, where the weight corresponding to the hyperedge e is denoted as $w(e)$. The degree of a vertex a is defined as $d(a) = \sum_{\{e \in \mathcal{E} | a \in e\}} w(e)$. Moreover, the incidence matrix H is an $|A| \times |\mathcal{E}|$ matrix, whose entry $H(a, e) = 1$ if $a \in e$, and $H(a, e) = 0$ otherwise. The degree of a hyperedge e is defined as $\delta(e) = |e|$. Therefore, we have $d(a) = \sum_{e \in \mathcal{E}} w(e) H(a, e)$ and $\delta(e) = \sum_{a \in A} H(a, e)$. Note that, the hyperedge weight $w(e)$ in this paper is defined as

$$w(e) = \frac{1}{|e|} \sum_{a_1 \in e, a_2 \in e} \text{sim}(a_1, a_2) \tag{9}$$

where $\text{sim}(\cdot, \cdot)$ denotes a predefined similarity measure of two vertices. Based on these definitions, we define the Laplacian matrix of the aforementioned hypergraph as follows

$$L^h = D_a - H W_e D_e^{-1} H^\top \tag{10}$$

where D_a, D_e, and W_e denote the diagonal matrices of the vertex degrees, the hyperedge degrees, and the hyperedge weights of the hypergraph, respectively.

According to Eq. 10, the affinity graph can be defined as $W^h = H W_e D_e^{-1} H^\top$, and thus a hypergraph can be viewed as a graph with pairwise similarities. It can be observed after some derivation that, for any two vertices $a_1 \in e$ and $a_2 \in e$, the affinity between them in W^h is $\frac{w(e)}{\delta(e)}$. Consequently, the high-order relevance among vertices in a hypergraph can be decomposed into a set of second-order relevance (as shown in W^h). With this in mind, the hypergraph can be viewed to have a cluster-based structure partitioned by hyperedges (if these hyperedges do not overlap), and we present two types of cluster-based structures in weakly-supervised image parsing in the next two subsections.

5.2 Intra-Image Hyperedges

We begin by analyzing the cluster-based structure within a single image. Intuitively, superpixels that are both visually similar and spatially adjacent tend to be semantically relevant. In order to achieve this goal, we establish a graph W_m for the image X_m whose vertices correspond to superpixels of X_m, and only retain the pairwise visual similarities of spatially adjacent superpixels on W_m. Then, the cluster-based structure is learned on W_m by spectral clustering [23], which can be viewed as a graph partitioning task [24]. The spectral clustering algorithm is summarized as follows:

1. Find K smallest nontrivial eigenvectors $v_1, \ldots, v_K$ of the Laplacian matrix $L_m = D_m - W_m$, where D_m is the degree matrix of W_m.
2. Form $E = [v_1, \ldots, v_K]$, and normalize each row of E to have unit length.
3. Perform K-means clustering on the vectors $E_{i\cdot}$ (where $i = 1, \ldots, n_m$) to partition the n_m superpixels into K clusters.

After performing spectral clustering for each image, superpixels belonging to the same cluster form an intra-image hyperedge. By stacking all the hyperedges into a single incidence matrix H, we can obtain the Laplacian matrix according to Eq. 10. In order to state conveniently, we denote

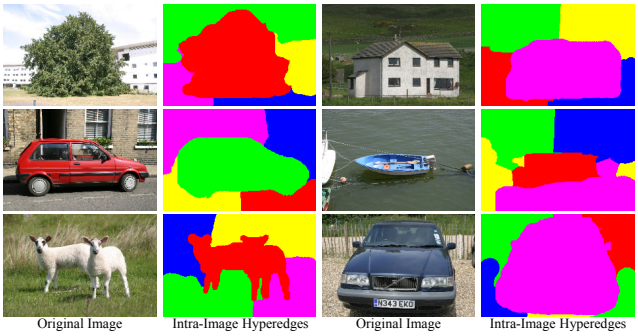


Figure 2: Illustration of some intra-image hyperedges. Superpixels in the area of the same color share an intra-image hyperedge. Please note that an area generally contains more than one superpixel, and that the number of clusters here is set to 5.

the resultant hypergraph as HG_{intra}. Some example results are shown in Fig. 2.

It should be noted that, spatial adjacency has already been utilized in image segmentation in the literature. For example, all pairwise similarities between spatially adjacent superpixels are retained in [20]. However, such a setting may cause all the superpixel labels to be similar, which may not be a good choice.

5.3 Inter-Image Hyperedges

Besides intra-image hyperedges, there is a second type of cluster-based structures, i.e., inter-image hyperedges. Recall that in Subsection 4.1, we define $Z_j \in \mathbb{R}^{d \times N_j}$ to be all the superpixels belonging to images with the j-th label. Moreover, the necessary condition of two superpixels being semantically relevant is that, the two images to which they belong have at least one label in common. Therefore, instead of discovering the cluster-based structure in Z (i.e., all the superpixels), we focus on the subsets Z_j, where $j \in \{1, \ldots, C\}$.

The approach to learning inter-image hyperedges consists of two steps: 1) Establish an $N_j \times N_j$ visual similarity graph W_j for all superpixels in Z_j; 2) Patition W_j by spectral clustering [23].

Similarly with Subsection 5.2, after performing spectral clustering for each category, superpixels belonging to the same cluster form an inter-image hyperedge. By stacking all the hyperedges into a single incidence matrix H, we can obtain the Laplacian matrix according to Eq. 10. We denote the resultant hypergraph as HG_{inter}.

Please note that, rather than using the full graph W_j, we observe in the experiments that the k-NN version of W_j results in better performance.

5.4 Discussion

We discuss the complementariness between HG_{intra} and HG_{inter} here. Moreover, the k-NN semantic graph proposed in Section 4 is also taken into account.

- HG_{intra} vs HG_{inter}. HG_{inter} is based on visual appearance only, while HG_{intra} leverages the spatial adjacency information in a single image.
- k-NN semantic graph vs HG_{intra}. Similarly with the above, k-NN semantic graph and HG_{intra} exploit visual appearance and spatial adjacency, respectively.
- k-NN semantic graph vs HG_{inter}. There may be errors in predicting candidate superpixels in k-NN semantic graph, and thus the neighbors of a given superpixel are not really semantic relevant ones. Fortunately, the cluster-based structure learned in HG_{inter} may help reduce the errors caused by k-NN semantic graph.

Therefore, the aforementioned three graphs (hypergraphs) are complementary with each other, and thus the combination of them can yield better parsing performance.

6. GRAPH COMBINATION BY RANDOM WALK

Currently, we have proposed three semantic graphs (hypergraphs) in total, i.e., k-NN semantic graph, HG_{intra} and HG_{inter}. As discussed in the previous section, these graphs are complementary with each other, and thus it is beneficial to combine them for image parsing. Hence, in this section, we present an approach to combine these graphs based on random walk. Recall that according to Eq. 10, the affinity graph is defined as $W^h = HW_eD_e^{-1}H^\top$, and thus a hypergraph can be viewed as a graph with pairwise similarities.

In order to state clearly, assume that we have S undirected and symmetric graphs $G_s = (A, W_s)$, where $s = 1, \ldots, S$. Each element $W_s(a, b)$ in the $N \times N$ similarity matrix W_s measures the similarity between superpixels x_a and x_b on the s-th graph. It should be noted that, these graphs share the same set of vertices while having different similarity matrices. With respect to each graph G_s, as in [6], for the vertex $a \in A$,we denote its degree on G_s as $d_s(a) = \sum_b W_s(a, b)$. Furthermore, the volume of graph G_s is defined as $\mathrm{vol}_s\, A = \sum_{a \in A} d_s(a) = \sum_{a \in A, b \in A} W_s(a, b)$. The natural random walk on G_s can be defined as follows. That is, for any two vertices a and b on G_s, their transition probability on G_s is

$$p(a \to b|G_s) = W_s(a, b)/d_s(a) \tag{11}$$

and the stationary probability of a on G_s is

$$p(a|G_s) = d_s(a)/\,\mathrm{vol}_s\, A \tag{12}$$

Denote $p(G_s)$ (where $s = 1, \ldots, S$) as the prior probabilities of the random walker choosing the graph G_s, and we have $p(G_s) \geq 0$ and $\sum_s p(G_s) = 1$. Therefore, the posterior probability in selecting the graph G_s at vertex a is

$$p(G_s|a) = \frac{p(G_s, a)}{\sum_s p(G_s, a)} = \frac{p(a|G_s)p(G_s)}{\sum_s p(a|G_s)p(G_s)} \tag{13}$$

For any two vertices a and b, their transition probability on multiple graphs can be computed as

$$p(a \to b) = \sum_s p(a \to b|G_s)p(G_s|a) \tag{14}$$

In addition, the stationary probability of vertex a on multiple graphs is computed as

$$p(a) = \sum_s p(a|G_s)p(G_s) \tag{15}$$

Finally, according to [9], the combined Laplacian matrix $\hat{L}$ of multiple graphs is defined as follows

$$\hat{L} = \Pi - \frac{\Pi P + P^\top \Pi}{2} \tag{16}$$

where P denotes the transition probability matrix with its elements being $p(a \to b)$, and Π is the diagonal stationary probability matrix with its diagonal elements being $p(a)$.

7. EXPERIMENTS

In this section, we evaluate the proposed semantic graphs and hypergraphs in weakly-supervised image parsing. We begin by describing the experimental setup and then compare our method with other closely related methods. Moreover, the parameter setting details are presented.

7.1 Experimental Setup

We conduct experiments on two standard datasets: PASCAL VOC'07 (PASCAL for short) [7] and MSRC-21 [26]. 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 weakly-supervised image parsing, 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 [25]. Moreover, we adopt SLIC [1] to obtain superpixels for each image, and represent each superpixel by the bag-of-words model while using SIFT [21] as the local descriptor. Histogram intersection kernel [2] is adopted to measure the similarity between two feature vectors of superpixels. To present fair comparisons, we adopt the same parameters for the graph propagation model shown in Eq. 4. Furthermore, the setting of parameters in constructing semantic graphs and hypergraphs will be investigated in Subsection 7.3.

Notably, besides computing the parsing accuracy on the entire image repository, we also measure the semantic relevance captured by a graph with a percentage value

$$\text{percentage} = \frac{\#(\text{adjacent superpixels with the same label})}{\#(\text{adjacent superpixels})} \tag{17}$$

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

Apart from comparing with the state-of-the-arts [17, 16], we mainly focus on the comparisons with some closely related baselines. For example, to demonstrate the effectiveness of the proposed k-NN semantic graph (k-NN SG for short), we compare with the following two baselines: 1) k-NN original graph (k-NN OG), where all superpixels are candidates for a given superpixel; 2) k-NN label intersection graph (k-NN LIG), where all the candidate superpixels belong to images which have at least one common label with the image containing the given superpixel. Our aim is to show that it is beneficial to reduce the number of candidate superpixels in establishing superpixel graphs.

In addition, since HG_{intra} exploits the spatial adjacency information, we compare it with an adjacency graph G_{adj}, where all pairwise similarities between spatially adjacent superpixels are retained [20]. We also compare HG_{inter} with baseline k-NN graphs, since HG_{inter} is based on a partition of k-NN visual similarity graphs as shown at the end of Subsection 5.3. Finally, we enumerate all the possible graph combination strategies to demonstrate the complementariness among the proposed semantic graphs and hypergraphs.

7.2 Empirical Results

The per-class accuracies on PASCAL dataset and MSRC-21 dataset are listed in Table 1 and Table 2, respectively.

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

Graphs	Percentage	Accuracy
k-NN OG	11	19
k-NN LIG	34	37
k-NN SG	38	42
HG_{inter}	32	41
G_{adj}	43	n/a
HG_{intra}	49	n/a

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

Graphs	Percentage	Accuracy
k-NN OG	33	65
k-NN LIG	52	65
k-NN SG	59	73
HG_{inter}	51	70
G_{adj}	50	n/a
HG_{intra}	66	n/a

Note that we have numbered the compared methods in these two tables for convenience.

It can be observed that trends on both datasets are similar, and that the results achieved by (11) are significantly better than the state-of-the-arts [17, 16]. In addition, we have the following findings based on the results in Table 1 and Table 2:

- (3) vs {(1), (2)}: k-NN SG outperforms the other two baselines, and thus it is beneficial to reduce the number of candidate superpixels in graph construction.
- (5) vs (1), (8) vs (6), (10) vs (3): HG_{intra} takes into account the spatial adjacency information of superpixels, which is complementary to the visual appearance. Therefore, combining HG_{intra} with graphs considering only visual appearance can improve the performance.
- (5) vs (4), (8) vs (7): G_{adj} retains the similarity of all spatially adjacent superpixels, which may cause all the superpixel labels to be similar after label propagation. In contrast to such an undesirable setting, HG_{intra} only impose regularization on superpixels which are *both spatially adjacent and visually similar*.
- (6) vs (1): Although HG_{inter} is based on a partition of k-NN visual similarity graphs, the high-order semantic relevance is a key factor to the parsing performance.
- (8) vs (6), (9) vs {(3), (6)}, (10) vs (3), (11) vs {(8), (9), (10)}: It can be observed that the proposed k-NN SG, HG_{intra} and HG_{inter} are complementary with each other, which is in correspondence with the discussion in Subsection 5.4.

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. By comparing among k-NN OG, k-NN LIG and k-NN SG, we observe that, generally, the more semantic relevance captured by the graph, the better the parsing accuracy is. However, although the

Table 1: Accuracies (%) of the proposed semantic graphs and hypergraphs 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
bi-layer sparse coding [17]	24	25	40	25	32	35	27	45	16	49	24	32	13	25	56	28	17	16	33	18	**82**	32
L_1 graph+χ^2 distance graph [16]	28	20	52	28	46	41	39	60	25	68	25	35	17	**35**	56	36	46	17	31	20	65	38
(1) 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
(2) 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
(3) 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
(4) k-NN OG+G_{adj}	28	24	9	14	13	19	22	12	18	14	8	12	9	12	**59**	14	17	15	12	27	76	21
(5) k-NN OG+HG_{intra}	30	35	11	16	15	21	26	11	17	20	13	16	9	12	57	22	18	16	16	26	74	23
(6) HG_{inter}	56	21	71	**57**	44	40	43	**62**	16	76	19	**45**	17	23	46	30	48	10	49	38	53	41
(7) HG_{inter}+G_{adj}	57	17	72	49	53	42	**55**	54	18	64	23	38	23	30	45	24	48	17	41	27	54	41
(8) HG_{inter}+HG_{intra}	62	18	77	56	**58**	44	54	51	15	65	26	41	**28**	**35**	46	29	46	22	48	34	50	43
(9) k-NN SG+HG_{inter}	76	34	**89**	55	36	41	41	**62**	21	80	20	25	17	21	48	19	**77**	20	**58**	34	46	44
(10) k-NN SG+HG_{intra}	76	**62**	79	49	49	39	49	53	26	73	23	23	12	15	48	**40**	55	**32**	47	**42**	38	44
(11) k-NN SG+HG_{inter}+HG_{intra}	77	48	87	50	56	**48**	44	60	27	76	18	38	25	31	52	38	59	31	51	34	41	**47**

Table 2: Accuracies (%) of the proposed semantic graphs and hypergraphs 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
L_1 graph+χ^2 distance graph [16]	70	92	49	10	10	83	36	**82**	62	20	52	98	88	48	98	70	75	95	76	43	23	61
(1) 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
(2) 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
(3) k-NN SG	49	82	45	59	51	90	78	68	66	68	**98**	**99**	94	84	**99**	**99**	48	**99**	98	30	20	72
(4) k-NN OG+G_{adj}	74	92	67	36	18	92	47	74	71	50	84	96	88	57	**99**	81	77	96	78	32	**25**	68
(5) k-NN OG+HG_{intra}	**75**	**94**	63	55	28	93	45	74	72	59	80	96	86	63	98	86	77	97	84	37	18	70
(6) HG_{inter}	70	87	72	25	13	91	81	79	54	70	80	97	86	61	98	92	68	**99**	92	**50**	9	70
(7) HG_{inter}+G_{adj}	70	88	74	20	14	91	82	78	77	69	82	97	88	58	98	93	65	98	93	44	15	71
(8) HG_{inter}+HG_{intra}	72	88	**76**	30	18	93	84	80	80	75	81	97	90	61	98	92	67	98	94	49	19	73
(9) k-NN SG+HG_{inter}	67	88	59	45	38	93	82	75	81	78	95	98	**95**	**85**	**99**	98	62	**99**	98	34	16	75
(10) k-NN SG+HG_{intra}	53	84	46	63	**62**	89	72	67	78	69	96	**99**	90	84	98	97	53	**99**	98	27	20	74
(11) k-NN SG+HG_{inter}+HG_{intra}	71	89	60	**64**	57	93	**90**	76	**90**	**85**	95	**99**	**95**	83	**99**	**99**	66	**99**	**99**	34	**25**	**80**

semantic relevance captured by k-NN LIG is much more than k-NN OG, there is nearly no improvement in parsing accuracy on MSRC-21 dataset. It may be due to the fact that k-NN LIG only discards 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, k-NN SG further improves the percentage of semantically adjacent superpixels, which is beneficial for the final performance.

Moreover, although HG_{inter} is based on a partition of k-NN visual similarity graphs, it contains more semantic relevance than k-NN OG and thus performs better. Besides, by learning the cluster-based structure within each image, HG_{intra} obtains more semantic relevance than G_{adj}, where some semantically irrelevant adjacencies are discarded. It should be noted that, since neither HG_{intra} nor G_{adj} models the relevance among images, they cannot be directly used in image parsing.

Notably, the criterion shown in Eq. 8 is 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 k-NN SG is 64%, whereas the accuracy of label propagation with k-NN SG is 73%. The result shows the effectiveness of the whole framework. Moreover, some example results for image parsing by graph propagation using (11) in comparison with the ground-truth on MSRC-21 dataset are shown in Fig. 3.

However, the proposed method may fail in several cases. For example, the 'boat' category in MSRC-21 dataset performs relatively poor, since the number of 'boat' superpixels is small, and the intra-class variance of the 'boat' category is large. For another example, it is difficult to distinguish between the 'face' category and the 'body' category in MSRC-21 dataset, since images containing label 'face' generally contains label 'body'. As we only know the image-level label, the difference between these two labels is small. Moreover, the visual appearance of these two categories is similar (since both of them contain human skins), and thus the selection of candidate superpixels may fail sometimes.

Besides, we also conduct qualitative experiments on NUS-WIDE dataset. It should be noted that, NUS-WIDE dataset does not contain pixel-level labels for quantitative evaluation. [16] only evaluates the parsing results qualitatively. We have conducted experiments on a subset of NUS-WIDE dataset and qualitatively observed that our methods generally perform well. In order to allow for quantitative evaluation, we plan to manually annotate some pixel-level labels and conduct thorough experiments in the future.

7.3 Parameter Setting

There are four parameters in constructing the proposed semantic graphs and hypergraphs: 1) the number of neighbors k_1 in k-NN SG; 2) the number of clusters K_2 in HG_{intra}; 3) the number of neighbors k_3 in visual similarity graph W_j for constructing HG_{inter}; 4) the number of clusters K_4 in HG_{inter}. Note that subscripts are added to distinguish among these parameters. Moreover, since the sizes of visual similarity graphs $W_j \in \mathbb{R}^{|Z_j|\times|Z_j|}$ (where $j \in \{1,\ldots,C\}$) vary a lot among different classes in constructing HG_{inter}, we empirically set $K_4 = |Z_j|/k_3$ to allow for a flexible choice

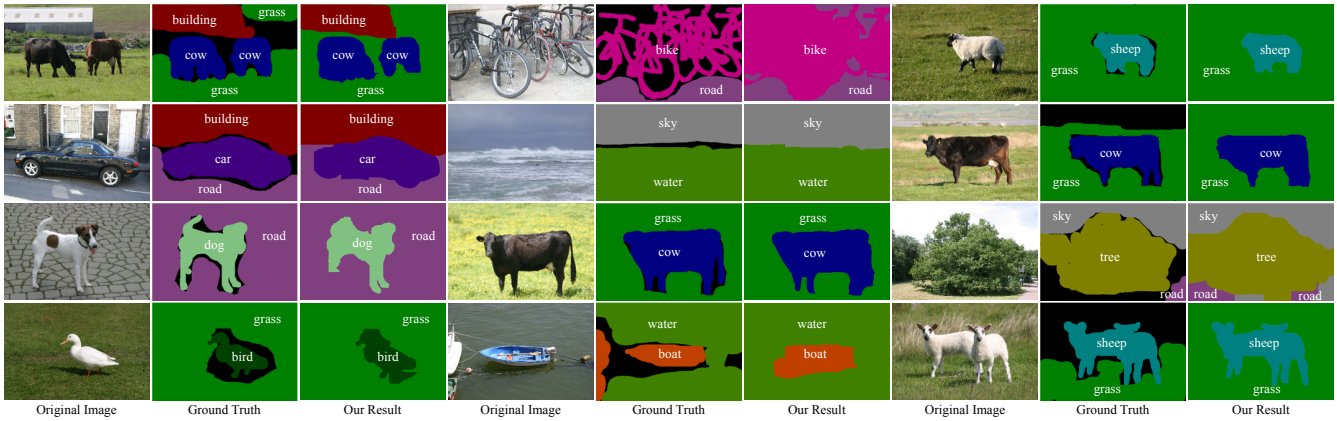


Figure 3: Some example results for image parsing by graph propagation using (11) k-NN SG+HG_{inter}+HG_{intra} (i.e., our result) in comparison with the ground-truth on MSRC-21 dataset.

of K_4 for different classes. Hence, we focus on the setting of k_1, K_2 and k_3, where the parsing accuracies by varying these parameters on MSRC-21 dataset are shown in Fig. 4.

We have the following observations according to Fig. 4:

- As k_1 increases, the performance generally decreases, which may be due to the fact that taking into account more neighbors also incurs more semantically irrelevant superpixels. Hence, we set k_1 to a relatively small value in the experiments, i.e., $k_1 = 20$.
- K_2 is the number of partitions in an image, as shown in Fig. 2. A small K_2 enforces the labels of most superpixels to be similar, whereas a large K_2 makes superpixels within an image disjoint with each other, neither of which are good ideas. Consequently, we set $K_2 = 5$ to exploit the spatial adjacency information effectively.
- Due to similar reasons with k_1, k_3 should not be too large, either. Although there are better results when k_3 is set to some large value (e.g., $k_3 = 80$), the results are unstable as k_3 goes large. In the experiments, we adopt $k_3 = 20$ in order to obtain reliable results.

Please note that we use the same parameters as discussed above on PASCAL dataset.

8. CONCLUSIONS

Based on the observations listed in Section 1, it is important to construct superpixel graphs with more semantic relevance in order to achieve better results in weakly-supervised image parsing. Since it is difficult to directly train a classifier (i.e., modeling the relationship between visual features and labels of superpixels) in weakly-supervised learning problems, it is an interesting and important issue to establish descriptive graphs in weakly-supervised setting by exploiting the weak supervision information carefully.

Moveover, we empirically observe that it is beneficial to reduce the number of candidate superpixels and to explore high-order semantic relevance. Therefore, we investigate in this paper the semantic graph and hypergraph construction. As shown in the experiments, the proposed semantic graphs and hypergraphs perform significantly better than conventional graphs. What is more, due to the complementariness among these graphs, the combination of them yields even more promising results. More notably, as a general framework, the proposed approach is suitable for other weakly-supervised learning tasks besides image parsing.

9. 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.

10. REFERENCES

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

[2] A. Barla, F. Odone, and A. Verri. Histogram intersection kernel for image classification. In *ICIP*, pages 513–516, 2003.

[3] M. Belkin and P. Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. *NIPS*, 14:585–591, 2001.

[4] C. Berge. *Hypergraphs*. North-Holland, Amsterdam, 1989.

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

[6] F. R. Chung. *Spectral graph theory*. American Mathematical Society, 1997.

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

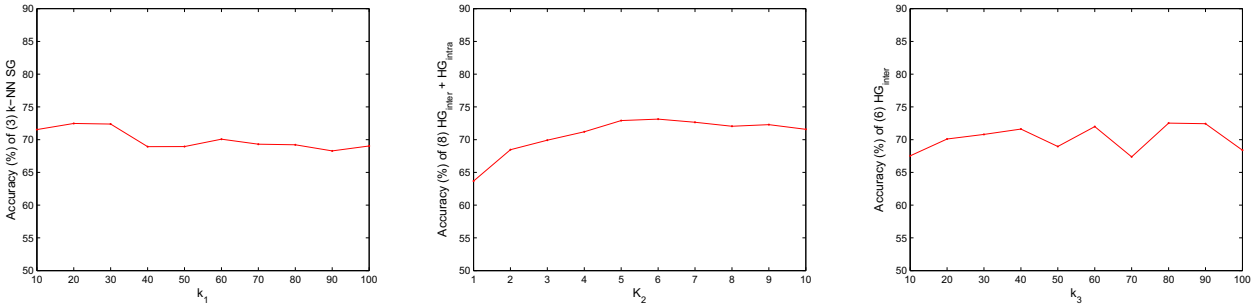
[8] C. Farabet, C. Couprie, L. Najman, and Y. LeCun. Learning hierarchical features for scene labeling. *TPAMI*, 35(8):1915–1929, 2013.

[9] Z. Fu, H. H. Ip, H. Lu, and Z. Lu. Multi-modal constraint propagation for heterogeneous image clustering. In *ACM MM*, pages 143–152, 2011.

[10] S. Gao, I.-H. Tsang, and L.-T. Chia. Laplacian sparse coding, hypergraph laplacian sparse coding, and applications. *TPAMI*, 35(1):92–104, 2013.

[11] Y. Gao, M. Wang, Z.-J. Zha, J. Shen, X. Li, and X. Wu. Visual-textual joint relevance learning for tag-based social image search. *TIP*, 22(1):363–376, 2013.

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



(a) (b) (c)

Figure 4: Mean accuracies (%) by varying parameters in graph construction on MSRC-21 dataset: (a) the number of neighbors k_1 in k-NN SG; (b) the number of clusters K_2 in HG_{intra}; (c) the number of neighbors k_3 in visual similarity graph W_j for constructing HG_{inter}.

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

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

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

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

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

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

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

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

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

[22] Z. Lu and Y. Peng. Latent semantic learning by efficient sparse coding with hypergraph regularization. In *AAAI*, pages 411–416, 2011.

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

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

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

[26] 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(1):2–23, 2009.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[42] D. Zhou, J. Huang, and B. Schölkopf. Learning with hypergraphs: Clustering, classification, and embedding. *NIPS*, 19:1601–1608, 2007.

[43] G. Zhu, S. Yan, and Y. Ma. Image tag refinement towards low-rank, content-tag prior and error sparsity. In *ACM MM*, pages 461–470, 2010.

[44] L. Zhuang, H. Gao, Z. Lin, Y. Ma, X. Zhang, and N. Yu. Non-negative low rank and sparse graph for semi-supervised learning. In *CVPR*, pages 2328–2335, 2012.