Tensor-based High-order Semantic Relation Transfer for Semantic Scene Segmentation

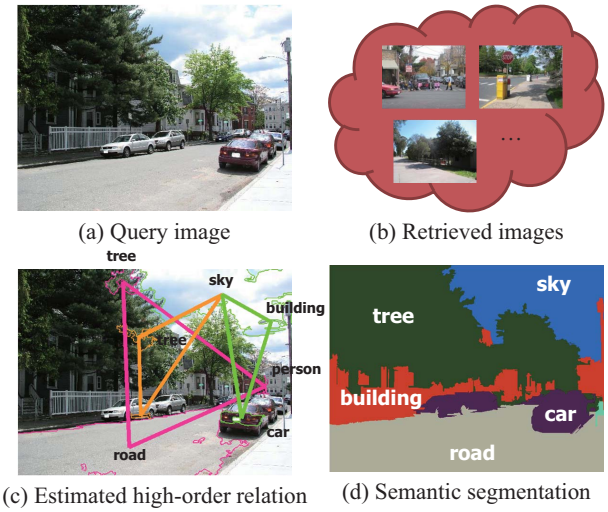
Heesoo Myeong and Kyoung Mu Lee
Department of EECS, ASRI, Seoul National University, 151-742, Seoul, Korea
heesoo.myeong@gmail.com, kyoungmu@snu.ac.kr
http://cv.snu.ac.kr

Abstract

We propose a novel nonparametric approach for semantic segmentation using high-order semantic relations. Conventional context models mainly focus on learning pairwise relationships between objects. Pairwise relations, however, are not enough to represent high-level contextual knowledge within images. In this paper, we propose semantic relation transfer, a method to transfer high-order semantic relations of objects from annotated images to unlabeled images analogous to label transfer techniques where label information are transferred. We first define semantic tensors representing high-order relations of objects. Semantic relation transfer problem is then formulated as semi-supervised learning using a quadratic objective function of the semantic tensors. By exploiting low-rank property of the semantic tensors and employing Kronecker sum similarity, an efficient approximation algorithm is developed. Based on the predicted high-order semantic relations, we reason semantic segmentation and evaluate the performance on several challenging datasets.

1. Introduction

Semantic segmentation, segmenting all the objects and identifying their categories, is fundamental and important problem in computer vision. Recently, with the increasing availability of large image collections of hand-labeled images, nonparametric label transfer approaches for this problem have attracted many computer vision researchers and shows very good performance [2, 3, 16, 23, 24, 25, 26]. Compared to conventional parametric semantic segmentation methods [1, 6, 14, 22], these approaches do not need training model parameters, hence, they can be scalable to large datasets with an unknown number of object categories. Typical label transfer approaches start by retrieving similar images for a given test image. After that, they establish dense correspondence between two images and then warp labels from the matched annotated images to the test image. In spite of good performances, these approaches sometimes produce unsatisfactory results because they do not explore high-level contextual knowledge within the annotated images. Obviously, high-level semantic relationships between objects within the annotated image are very important cues to successful semantic segmentation.



(a) Query image (b) Retrieved images (c) Estimated high-order relation (d) Semantic segmentation

Figure 1. For a query image (a), our system finds the matched similar images (b) from a large dataset using global scene descriptors. The high-order semantic relations are transferred from the annotated images (b) to the query image (a). (We densely estimate high-order semantic relation across the image, but this figure displays only a few top scored relations for visualization purposes.) We then infer semantic segmentation (d) using estimated semantic relation (c).

To this end, recent approaches have advocated the use of nonparametric context models [10, 19]. These learn pairwise relationships between objects using global scene features and local features. However, these methods use only pairwise relationships to model high-level semantic rela-

tionships. Since natural images typically contain more than three object categories, pairwise relations are not enough to represents high-level information within images.

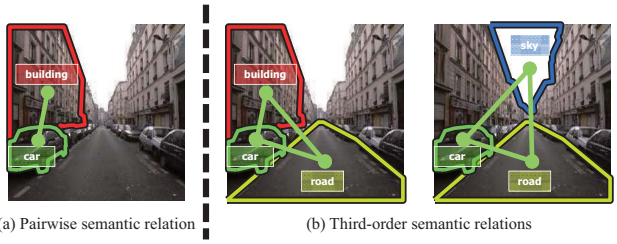
In this paper, we develop a novel nonparametric approach for semantic segmentation by incorporating high-order semantic relations. Specifically, similar to several label transfer methods [3, 16, 23, 25], we first find a set of small retrieved images from training images. Our goal is to transfer high-order semantic relations of annotated objects from each matched image to the query image. Since it is not feasible to obtain dense pixel-wise high-order semantic relations, we utilize "superpixel" regions obtained by oversegmentation of the query image. We define semantic tensors to represent the higher-order semantic relations of regions. We approach the problem of transferring the high-order semantic relations by defining a quadratic objective function of the semantic tensors. To optimize our objective function, we develop an efficient approximate algorithm based on Kronecker sum similarity and low-rank property of semantic tensors. To integrate our predicted semantic tensor into a semantic segmentation system, a fully connected Markov random field optimization is employed.

The key contributions of this paper include: (1) The use of high-order semantic relations for semantic segmentation; (2) A novel tensor-based representation of high-order semantic relations; and (3) A quadratic objective function for learning the semantic tensor and an efficient approximate algorithm.

The paper is organized as follows. We review some relevant works in Section 2. In Section 3, we introduce high-order semantic relation transfer algorithm and explain in detail. Section 3.3 presents a semantic segmentation method through semantic relation transfer. The experimental results are given in Section 5. Finally, in Section 6, we discuss our approach.

2. Related work

We now review related works on label transfer approaches and nonparametric context models. The problem of label transfer was first addressed recently by Liu *et al*. [16]. They first retrieved similar images using GIST matching [20] and constructed pixel-wise dense correspondence between each retrieved image and test image using SIFT flow [17]. They then transferred the annotations based on dense correspondence and reasoned semantic segmentation. Following the idea of label transfer [16], Zhang *et al*. [25] employed partial matching between the test image and the retrieved images to use partial similarity between images. Gould and Zhang [7] constructed PatchMatchGraph to reduce complexity of retrieval step. Chen *et al*. [3] proposed supervised geodesic propagation to guide label transfer. Tighe and Lazebnik [23, 24] considered superpixel-level matching to transfer label information. However, all of these approaches are restricted to transferring label information from matched images. Although Liu *et al*. [16] claimed that the label transfer approach naturally embeds contextual information in the retrieval/alignment procedure, it is hard to tell how much contextual knowledge will help or what the effects will be.



(a) Pairwise semantic relation (b) Third-order semantic relations

Figure 2. An example of pairwise and high-order semantic relations. The third-order semantic relations (b) can model complicated high-level semantic knowledges within an image compared with the pairwise semantic relation (a).

On the other hand, recent nonparametric context models [10, 19] for semantic segmentation employed contextual relationships between objects to achieve more accurate results. Jain *et al*. [10] learned which contextual relationships should be considered and predicted features weight for each relation in a nonparametric manner. Myeong *et al*. [19] formulated a data-driven context learning problem as a graph-based context link prediction problem. Since our semantic tensor can be viewed as a generalization of the context link [19], their work is most similar to our own. However, there are several important differences with respect to our work. First, they only considered pairwise object relationships. On the contrary, our method focuses on high-order (mostly third-order) semantic relations, allowing us to model complex contextual relationships. For example, triplet-wise semantic relations can be found such as (*sky*,*car*,*road*) by our method as illustrated in Figure 2. These relations become important when considering complicated scenes with many object classes. Second, we develop a quadratic objective function for the high-order semantic relation transfer problem. However, Myeong *et al*. [19] did not show how their context link prediction works mathematically.

High-order models are not well studied in the context of semantic segmentation. Kohli *et al*. [12] introduced high-order model to enforce label consistency among regions. However, their high-order model is not related to high-level semantic knowledge. To our knowledge, there are no prior works explicitly considering high-order contextual relationships between objects in the literature on semantic segmentation.

3. The high-order semantic relation transfer algorithm

3.1. Problem statement

We consider two images I^1 and I^2; the first one is not annotated whereas the second one is densely labeled with the corresponding object class. We assume that two images are closely-related in which the similar objects are present and that objects roughly maintain their high-order relation. We define high-order semantic relation transfer problem as a task to predict high-order relation between unlabeled regions in I^1 based on annotated regions in I^2. For simplicity, we will focus on third-order relations from now.

Let $\mathcal{S} = \{S^1, S^2\}$ be a set of superpixels generated by segmenting the respective images. n^1 and n^2 is the number of segments in S^1 and S^2, respectively, and $N = n^1 + n^2$ is the total number of segments. $\mathcal{C} = \{c_1, c_2, ..., c_K\}$ is a given set of object classes. Third-order semantic relations among region triplets $(s_i, s_j, s_k) \in \mathcal{S} \times \mathcal{S} \times \mathcal{S}$ is defined as a set of $N \times N \times N$ third-order tensors $\mathbb{X} = \{\mathcal{X}^{111}, \mathcal{X}^{112}, \mathcal{X}^{113}, ..., \mathcal{X}^{KKK}\}$. We refer to each tensor $\mathcal{X}^{\alpha\beta\gamma} \in \mathbb{X}$ as a *semantic tensor*. A semantic tensor $\mathcal{X}^{\alpha\beta\gamma}$ denotes third-order semantic relations among region triplets on object class triplet $(c_\alpha, c_\beta, c_\gamma)$. Each element of $\mathcal{X}^{\alpha\beta\gamma}$ is defined as

$$[\mathcal{X}^{\alpha\beta\gamma}]_{ijk} = x_{ijk}^{\alpha\beta\gamma}. \tag{1}$$

The variable $x_{ijk}^{\alpha\beta\gamma}$ indicates confidence score of how likely the region triplet (s_i, s_j, s_k) would be labeled as $(c_\alpha, c_\beta, c_\gamma)$, respectively. $x_{ijk}^{\alpha\beta\gamma}$ is close to 1 if the assigned object class triplet $(c_\alpha, c_\beta, c_\gamma)$ is reliable. On the other hand, $x_{ijk}^{\alpha\beta\gamma}$ is close to 0 if the assigned object class triplet $(c_\alpha, c_\beta, c_\gamma)$ is unreliable.

Next, we define another set of $N \times N \times N$ tensor representing the observed third-order semantic relations within the image I^2. Similar to $\mathbb{X}$, we define $\mathbb{Y} = \{\mathcal{Y}^{111}, \mathcal{Y}^{112}, \mathcal{Y}^{113}, ..., \mathcal{Y}^{KKK}\}$, and represent each element of $\mathcal{Y}^{\alpha\beta\gamma}$ as

$$y_{ijk}^{\alpha\beta\gamma} = \begin{cases} 1 & \text{if } G(s_i) = c_\alpha, G(s_j) = c_\beta, G(s_k) = c_\gamma, \\ & (s_i, s_j, s_k) \in S^2 \\ 0 & \text{otherwise} \end{cases}, \tag{2}$$

where $G(s_i)$ denotes the ground truth class of region s_i and $(s_i, s_j, s_k) \in S^2$ indicates that three regions s_i, s_j, and s_k are from the same image I^2. Since there are no semantic relations within S^1 and across images, all $y_{ijk}^{\alpha\beta\gamma}$ is 0 for $(s_i, s_j, s_k) \notin S^2$. In practice, each $\mathcal{Y}^{\alpha\beta\gamma}$ can be compactly generated from label vectors. Let $\mathbf{y}^\alpha$ be a column vector of length N, where $[\mathbf{y}^\alpha]_i = y_i^\alpha$ is 1 if region s_i belongs to object class c_α; and 0 otherwise. Then each element of $\mathcal{Y}^{\alpha\beta\gamma}$ can be generated by

$$y_{ijk}^{\alpha\beta\gamma} = y_i^\alpha y_j^\beta y_k^\gamma. \tag{3}$$

Eq. (3) can be rewritten as

$$\mathcal{Y}^{\alpha\beta\gamma} = \mathbf{y}^\alpha \circ \mathbf{y}^\beta \circ \mathbf{y}^\gamma. \tag{4}$$

The symbol "$\circ$" denotes the vector outer product. Since $\mathcal{Y}^{\alpha\beta\gamma}$ can be represented as the outer product of three vectors, $\mathcal{Y}^{\alpha\beta\gamma}$ is a *rank-one* tensor [13]. This rank-one property of $\mathbb{Y}$ is one of key aspects to approximate the following objective function.

3.2. Objective function

Now, the third-order semantic relation transfer problem can be regarded as the problem of estimating the magnitude of confidence scores $x_{ijk}^{\alpha\beta\gamma}$ for all superpixel triplets (s_i, s_j, s_k) and for all object class triplets $(c_\alpha, c_\beta, c_\gamma)$ based on $\mathbb{Y}$. We assume that there is no interaction between the semantic tensors. Hence, we separately deal with the third-order semantic relations transfer problem with respect to $\mathcal{Y}^{\alpha\beta\gamma}$. For simplicity, we drop the $\alpha\beta\gamma$ suffix from now.

Following the idea of link propagation [11], we want to enforce that two similar region triplets are likely to have the same confidence score. Thus, we design the quadratic objective function with respect to $\mathcal{Y}$ as

$$\begin{aligned} F(\mathcal{X}) =& \frac{1}{2} \sum_{i,j,k,l,m,n}^{N} w_{ijk,lmn}(x_{ijk} - x_{lmn})^2 \\ &+ \lambda \sum_{i,j,k}^{N} (x_{ijk} - y_{ijk})^2, \end{aligned} \tag{5}$$

where $w_{ijk,lmn}$ is the triplet-wise similarity between two region triplets (s_i, s_j, s_k) and (s_l, s_m, s_n) and $\lambda > 0$ is the regularization parameter. The first term of Eq. (5) is the continuity constraint that two triplets should have the same confidence score if two triplets are similar. The second term is the unary constraint that each region triplet x_{ijk} tends to have their target values y_{ijk}. The cost function defined as pairwise and unary terms is a generalization of the cost function for label propagation [27].

Now, we rewrite Eq. (5) using tensors. For that, let $\mathbf{L}$ be an $N^3 \times N^3$ matrix called a *Laplacian matrix* defined as

$$\mathbf{L} = \mathbf{D} - \mathbf{W}, \tag{6}$$

where $w_{ijk,lmn}$ is rewritten as similarity matrix $\mathbf{W}$ of size $N^3 \times N^3$ and $\mathbf{D}$ is a diagonal matrix whose diagonal elements are $[\mathbf{D}]_i = \sum_j^{N^3} [\mathbf{W}]_{ij}$. Using $\mathbf{L}$, Eq. (5) can be reformulated as

$$F(\mathcal{X}) = \frac{1}{2} \mathbf{vec}(\mathcal{X})^T \, \mathbf{L} \, \mathbf{vec}(\mathcal{X}) + \lambda (\mathbf{vec}(\mathcal{X}) - \mathbf{vec}(\mathcal{Y}))^2, \tag{7}$$

where $\mathbf{vec}(\mathcal{X})$ is the vector constructed by concatenating the mode-1 fibers of the tensor $\mathcal{X}$ [13].

Differentiating Eq. (7) with respect to $\mathbf{vec}(\mathcal{X})$, and set to 0, we can get $\mathcal{X}$ that minimizes Eq. (7),

$$\frac{\partial F(\mathcal{X})}{\partial \mathbf{vec}(\mathcal{X})} = \mathbf{L}\,\mathbf{vec}(\mathcal{X}) + \lambda \mathbf{vec}(\mathcal{X}) - \lambda \mathbf{vec}(\mathcal{Y}) = 0 \quad (8)$$

It can be transformed into

$$(\mathbf{L} + \lambda \mathbf{I})\mathbf{vec}(\mathcal{X}) = \lambda \mathbf{vec}(\mathcal{Y}), \quad (9)$$

where $\mathbf{I}$ indicates identity matrix of size $N^3 \times N^3$. Since $\mathbf{L} + \lambda \mathbf{I}$ is positive definite, the linear equation (9) can be solved by matrix inversion. However, computing inverse matrix of size $N^3 \times N^3$ is not realistic in practice.

3.3. Approximate algorithm

In this section, we present an efficient optimization scheme for the proposed objective function. Since providing all of the N^6 elements of the triplet-wise similarity matrix $\mathbf{W}$ is intractable, we consider constructing $\mathbf{W}$ using the segments-wise similarity matrix $\mathbf{W}_S$ the same as [11]. As described in Section 5, $\mathbf{W}_S$ is defined as similarity between two superpixels. Recommended by [11], we define $\mathbf{W}$ based on *Kronecker sum similarity*. Hence, $\mathbf{L}$ can be re-represented as

$$\mathbf{L} = \mathbf{L}_S \oplus \mathbf{L}_S \oplus \mathbf{L}_S, \quad (10)$$

where $\oplus$ indicates the Kronecker sum and $\mathbf{L}_S$ is defined as $\mathbf{L}_S = \mathbf{D}_S - \mathbf{W}_S$ and $\mathbf{D}_S$ is a diagonal matrix whose diagonal elements are $[\mathbf{D}_S]_i = \sum_j^N [\mathbf{W}_S]_{ij}$. Using Eq. (10), the objective function (5) can be expressed as

$$F(\mathcal{X}) = \frac{1}{2}\mathbf{vec}(\mathcal{X})^T \mathbf{vec}(\mathcal{X} \times_1 \mathbf{L}_S + \mathcal{X} \times_2 \mathbf{L}_S + \mathcal{X} \times_3 \mathbf{L}_S) + \lambda(\mathbf{vec}(\mathcal{X}) - \mathbf{vec}(\mathcal{Y}))^2, \quad (11)$$

where $\times_n$ represents n-mode product of tensor [13]. Inspired by [5, 18], we approximate the objective function in three optimization steps:

$$\dot{\mathcal{X}} = \arg\min_{\mathcal{X}} \frac{1}{2}\mathbf{vec}(\mathcal{X})^T \mathbf{vec}(\mathcal{X} \times_1 \mathbf{L}_S) + \lambda(\mathbf{vec}(\mathcal{X}) - \mathbf{vec}(\mathcal{Y}))^2 \quad (12)$$

$$\ddot{\mathcal{X}} = \arg\min_{\mathcal{X}} \frac{1}{2}\mathbf{vec}(\mathcal{X})^T \mathbf{vec}(\mathcal{X} \times_2 \mathbf{L}_S) + \lambda(\mathbf{vec}(\mathcal{X}) - \mathbf{vec}(\dot{\mathcal{X}}))^2 \quad (13)$$

$$\hat{\mathcal{X}} = \arg\min_{\mathcal{X}} \frac{1}{2}\mathbf{vec}(\mathcal{X})^T \mathbf{vec}(\mathcal{X} \times_3 \mathbf{L}_S) + \lambda(\mathbf{vec}(\mathcal{X}) - \mathbf{vec}(\ddot{\mathcal{X}}))^2. \quad (14)$$

That is, we sequentially estimate the semantic tensor for each mode product term. In a similar way to Eq. (9), we can obtain linear system equation for each optimization step.

$$\mathcal{X} \times_1 (\mathbf{L}_S + \lambda \mathbf{I}_S) = \lambda \mathcal{Y} \quad (15)$$

$$\mathcal{X} \times_2 (\mathbf{L}_S + \lambda \mathbf{I}_S) = \lambda \dot{\mathcal{X}} \quad (16)$$

$$\mathcal{X} \times_3 (\mathbf{L}_S + \lambda \mathbf{I}_S) = \lambda \ddot{\mathcal{X}}, \quad (17)$$

where $\mathbf{I}_S$ indicates identity matrix of size $N \times N$. For solving each linear equation, let us consider Eq. (15), 1-mode tensor product of Eq. (15) can be expressed in terms of unfolded tensors.

$$(\mathbf{L}_S + \lambda \mathbf{I}_S)\mathbf{X}_{(1)} = \lambda \mathbf{Y}_{(1)}, \quad (18)$$

where $\mathbf{X}_{(1)}$ denotes the mode 1 matricization of a tensor $\mathcal{X}$ (see [13] for more details). Remind that $\mathcal{Y}$ is rank-one, $\mathcal{Y}$ can be written as in matricized form [13],

$$\mathbf{Y}_{(1)} = \mathbf{y}^\alpha (\mathbf{y}^\gamma \circ \mathbf{y}^\beta)^T. \quad (19)$$

Hence, $\dot{\mathcal{X}}$ can be efficiently computed by

$$\dot{\mathbf{X}}_{(1)} = (\mathbf{L}_S + \lambda \mathbf{I}_S)^{-1} \lambda \mathbf{y}^\alpha (\mathbf{y}^\gamma \circ \mathbf{y}^\beta)^T. \quad (20)$$

We continue to solve for $\ddot{\mathcal{X}}$ and $\hat{\mathcal{X}}$ similarly. Then we can obtain the approximate solution of the objective function (5) as follows.

$$\hat{\mathcal{X}} = [(\mathbf{L}_S + \lambda \mathbf{I}_S)^{-1} \lambda \mathbf{y}^\alpha] \circ [(\mathbf{L}_S + \lambda \mathbf{I}_S)^{-1} \lambda \mathbf{y}^\beta] \circ [(\mathbf{L}_S + \lambda \mathbf{I}_S)^{-1} \lambda \mathbf{y}^\gamma]. \quad (21)$$

Note that $\hat{\mathcal{X}}$ also can be represented as the outer product of three vectors, $\hat{\mathcal{X}}$ is a *rank-one* tensor. In Figure 3, this procedure summarizes schematically. We independently transfer each $\mathcal{Y}^{\alpha\beta\gamma}$, hence, this procedure repeats K^3 times. Finally, we can get the set of the predicted semantic tensors $\hat{\mathbb{X}} = \{\hat{\mathcal{X}}^{111}, \hat{\mathcal{X}}^{112}, \hat{\mathcal{X}}^{113}, ..., \hat{\mathcal{X}}^{KKK}\}$.

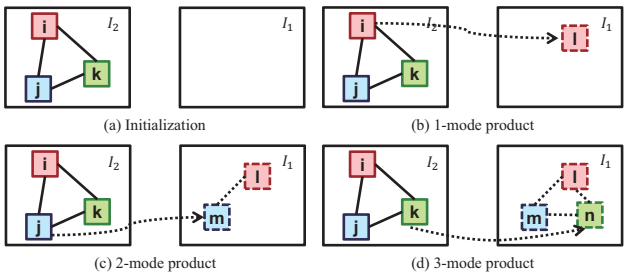


Figure 3. Illustration of the proposed approximate algorithm. The algorithm (b) first find similar region s_l with respect to s_i while fixing s_j and s_k, (c) then find similar region s_m with respect to s_j while fixing s_l and s_k, (d) and finally find similar region s_n with respect to s_k while fixing s_l and s_m.

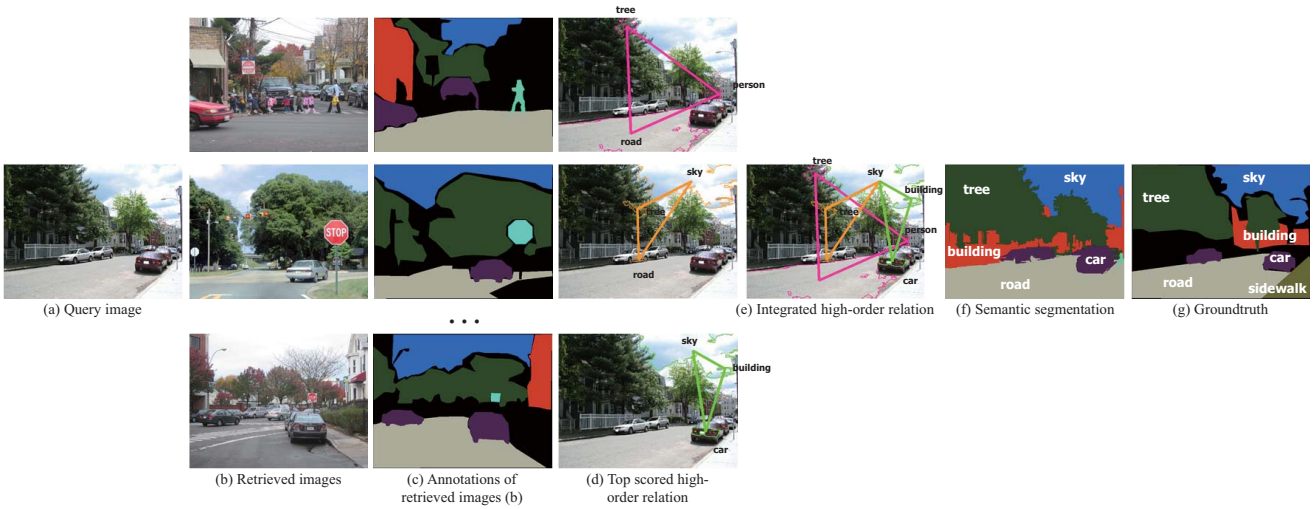


Figure 4. System overview. For a query image (a), we first retrieve the matched similar scenes (b). We predict the third-order semantic relations (d) by transferring semantic relations from each annotated image (c) to the query image (a). We aggregate semantic relations (e) from multiple semantic relation candidates (d) and generate semantic segmentation (f). (g) is the ground-truth annotation of (a).

4. Semantic segmentation through semantic relation transfer

Now that we have the semantic relation transfer algorithm from annotated images to unlabeled images, we can infer semantic segmentation using estimated semantic tensors.

4.1. Scene retrieval

Recall that we assume that each pair of images I^1 and I^2 roughly agree on the spatial layout of objects. Hence, it is essential to to extract closely-related images from large dataset with respect to a query image for successful semantic relation transfer. Unreliable semantic tensors can be predicted between two unrelated images. To find similar images, we first retrieve M candidate images by color histogram, GIST matching [20], and spatial pyramid [15] from the training dataset. This candidate image set will be used to transfer its high-order semantic relations into the query image.

4.2. Inference

After performing the scene retrieval in section 4.1, we transfer high-order semantic relations from each candidate image to the query image and obtain multiple sets of predicted semantic tensors $\{\mathbb{X}\}_{u=1:M}$. Our goal is to assign object class for each region in the query image. To integrate the sets of predicted semantic tensors with a conventional unary and pairwise potential, we build high-order fully connected Markov random field model. The energy function is defined as

$$E(\{l_i\}) = \sum_{i}^{n^1} E^D(l_i) + \sum_{(i,j)\in E} E^P(l_i, l_j) + \sum_{i,j,k}^{n^1} E^H(l_i, l_j, l_k), \quad (22)$$

where $l_i \in \{1, ..., K\}$ is the index of object class for region s_i. Since we want to label the regions in the query image, the energy function is only defined on the regions of image I^1. The first term is data term which represents the negative logarithm of the probability of class l_i given the region s_i. The second term is smoothness term which encourage two neighboring regions to have the same label. These two terms are typically used to conventional nonparametric scene parsing approaches [16, 23, 24].

However, it is nontrivial how to integrate the sets of predicted semantic tensors to semantic segmentation framework. Hence we develop two third-order clique potential E^H_{max} and E^H_{sum}. The first high-order potential E^H_{max} take the form

$$E^H_{max}(l_i = c_\alpha, l_j = c_\beta, l_k = c_\gamma) = -\log(\max_u \{\hat{x}^{\alpha\beta\gamma}_{ijk}\}_u). \quad (23)$$

The first clique potential E^H_{max} take maximum confidence score among M number of candidate scores for region triplet (s_i, s_j, s_k) and for object triplet $(c_\alpha, c_\beta, c_\gamma)$. This means that we only consider the strongest one from the set of relation candidates. The second high-order potential

Table 1. Performance comparison of our algorithm on the three challenging datasets. Per-pixel recognition rates and average per-class recognition rates in parentheses are presented.

	Jain *et al.* [10]	LMO [16]	Polo [25]
Jain *et al.* [10]	59.0 (-)	-	-
Liu *et al.* [16]	-	74.8 (-)	-
Tighe and Lazebnik [23]	-	76.8 (29.4)	87.9 (76.1) [25]
Zhang and Quan [25]	-	-	89.8 (82.5)
Chen *et al.* [2]	75.6 (45)	-	-
Myeong *et al.* [19]	80.1 (53.3)	**77.1 (32.3)**	-
Gould and Zhang [7]	-	-	**94.2 (91.7)**
Proposed (max)	81.5 (51.2)	76.1 (28.9)	89.1 (80.6)
Proposed (sum)	**81.8** (**54.4**)	76.2 (29.6)	88.3 (79.3)

E^H_{sum} have the form

$$E^H_{sum}(l_i = c_\alpha, l_j = c_\beta, l_k = c_\gamma) = -\log(\sum_u^M \{\hat{x}^{\alpha\beta\gamma}_{ijk}\}_u). \quad (24)$$

Meanwhile, the second clique potential E^H_{max} takes summation of M number of confidence scores. This potential picks average scores from the set of relation candidates. These two potential will be examined in the experimental section.

It is very important to effectively minimize the energy function (22), but efficient order reduction techniques such as [9] cannot be used due to space and time complexity. Hence, we apply multistart simulated annealing algorithm.

5. Experiements

In this section, we (1) evaluate our method's semantic segmentation performance and compare against pairwise semantic segmentation [19] and (2) analyze integration of our predicted semantic tensors. We validate our approach with three challenging datasets: the dataset of Jain *et al.* [10], LabelMe Outdoor (LMO) dataset [16], and Polo dataset [25]. We evaluate on all sets, but focus additional analysis on the LMO dataset since it has the largest number of categories. Table 1 summarizes our semantic segmentation accuracy compared with the state-of-the-art methods. Proposed (max) indicates the accuracy of the semantic segmentation with the max high-order term Eq. (23). Proposed (sum) represents performance with the sum high-order term Eq. (24).

Implementation details. Our implementation is based on the framework of Tighe and Lazebnik [23, 24]. We use the algorithm of Felzenszwalb and Huttenlocher [4] for segmentation, and fix the parameters $\sigma = 0.8, K = 200, min = 100$ on all sets. To form superpixel-wise weight $\mathbf{W}_S$, we use several types of descriptors $a_k(s_i)$ for regions s_i: shape, texture, color, and appearance from [23].

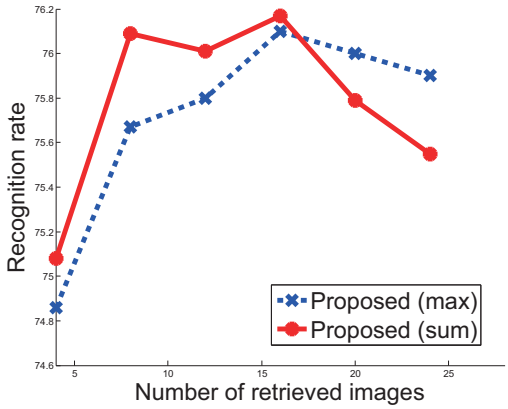


Figure 5. Recognition rate of two different high-order potential as a function of the number of the retrieved images M on the LMO dataset.

Along with appearance features, we integrate geometric position $g(s_i)$ (row+column) of the center of the region s_i. Hence, each elements of W_S are computed as

$$[\mathbf{W}_S]_{ij} = e^{-\sum_k \frac{\|a_k(s_i)-a_k(s_j)\|}{\sigma_{a_k}} - \frac{\|g(s_i)-g(s_j)\|}{\sigma_g}} \quad (25)$$

where $a_k(s_i)$ is the feature vector of the k-th type for s_i and σ_{a_k} denotes the standard deviation of a_k. Note that we densely obtain the weight between regions, it means that a region is connected to all the other regions with the corresponding weights. We fix the parameter of the objective function $\lambda = 10$. To compute E^D, we employ the nonparametric superpixel parsing [23] for the LMO dataset and the boosted decision tree classifier [8] for the other datasets. As a pairwise term E^P, we adopt simple Potts model.

Evaluation metric. We use both pixel-wise measure and class-wise measure to quantify the accuracy. The former rates total proportion of correctly labeled pixels, while the latter indicates the average proportion of correctly labeled pixels in each object class.

19-Class Jain *et al.* [10] dataset. Jain *et al*. [10] randomly collects 350 images of size 640×480 from LabelMe [21] with 19 classes. This dataset is splitted into 250 training images and 100 test images. The number of similar images M is set to be 16. The semantic segmentation accuracy on this dataset is 81.8%.

This is relatively good dataset to evaluate high-order semantic relations. The size of the images is large enough and there are a lot of objects within an image. We achieve the state-of-the-art performance on this dataset and obtain promising results.

33-Class LabelMe Outdoor (LMO) dataset. This dataset provided by Liu *et al*. [16] contains total 2,688 images of size 256×256 from LabelME [21] with 33 object categories. Liu *et al*. [16] randomly split this dataset into 2,488 training images and 200 test images. For qualitative comparison with [16, 19, 23], we use the same training/test split. We set the number of similar images M to 16. The semantic segmentation accuracy of the proposed method on this dataset is 76.2%.

Our results are below the state-of-the-art methods. We think that this is due to many images from this dataset with one or two object classes. The number of test images containing less than two object classes is 43 out of 200. It seems that complex contextual models such as the proposed method are not crucial to improve performance on this dataset.

6-Class Polo dataset. The polo dataset consists of 320 images from the web with keyword polo. Zhang *et al*. [25] annotated each image into six categories:*sky, horse, person, ground, tree, grass*. We set the number of similar images M to 20.

Our results are under the state-of-the-art methods. One reason is that context is not much important since all images have almost the same object classes. The other reason is the state-of-the-art method use complex pixel-wise model, on the other hand, we works on relatively simple region level.

Max vs. Sum. We design two different high-order potential for incorporating the set of the predicted semantic tensors. As shown in Figure 5, sum potential, taking summarization of candidates confidence scores, provides more better semantic segmentation results at some point. On the other hand, max potential, taking maximum of candidates confidence scores, is more robust to the number of retrieved images M. As gradually adding retrieved images, wrong matched images become larger and the performance of sum potential decreases faster.

6. Conclusion

We have presented a novel approach to learn high-order semantic relations of regions in a nonparametric manner. We cast the high-order semantic relation transfer problem as a quadratic objective function of semantic tensors and propose an efficient approximate algorithm. We develop a novel semantic tensor representation of the high-order semantic relations. While we have presented this representation in the context of semantic segmentation, it can be applicable to various computer vision problem including object detection, scene classification, and total scene understanding.

Acknowledgments

This research was supported in part by the MKE, Korea and Microsoft Research, under IT/SW Creative research program supervised by the NIPA (NIPA-2012-H0503-12-1035).

References

[1] J. Carreira, R. Caseiro, J. Batista, and C. Sminchisescu. Semantic segmentation with second-order pooling. In *ECCV*, 2012. 1

[2] X. Chen, A. Jain, A. Gupta, and L. S. Davis. Piecing together the segmentation jigsaw using context. In *CVPR*, 2011. 1, 6

[3] X. Chen, Q. Li, Y. Song, X. Jin, and Q. Zhao. Supervised geodesic propagation for semantic label transfer. In *ECCV*, 2012. 1, 2

[4] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient graph-based image segmentation. *Int. J. Comput. Vision*, 59:167–181, 2004. 6

[5] Z. Fu, Z. Lu, H. H.-S. Ip, Y. Peng, and H. Lu. Symmetric graph regularized constraint propagation. In *AAAI*, 2011. 4

[6] S. Gould, R. Fulton, and D. Koller. Decomposing a scene into geometric and semantically consistent regions. In *ICCV*, 2009. 1

[7] S. Gould and Y. Zhang. PatchMatchGraph: Building a graph of dense patch correspondences for label transfer. In *ECCV*, 2012. 2, 6

[8] D. Hoiem, A. A. Efros, and M. Hebert. Recovering surface layout from an image. *Int. J. Comput. Vision*, 75:151–172, 2007. 6

[9] H. Ishikawa. Higher-order clique reduction in binary graph cut. In *CVPR*, 2009. 6

[10] A. Jain, A. Gupta, and L. S. Davis. Learning what and how of contextual models for scene labeling. In *ECCV*, 2010. 1, 2, 6, 7

[11] H. Kashima, T. Katoy, Y. Yamanishiz, and M. Sugiyama. Link propagation: A fast semi-supervised learning algorithm for link prediction. In *SIAM International Conference on Data Mining*, 2009. 3, 4

[12] P. Kohli, L. Ladický, and P. H. Torr. Robust higher order potentials for enforcing label consistency. *Int. J. Comput. Vision*, 82(3):302–324, May 2009. 2

[13] T. G. Kolda and B. W. Bader. Tensor decompositions and applications. *SIAM Rev.*, 51(3):455–500, Aug. 2009. 3, 4

[14] L. Ladicky, C. Russell, P. Kohli, and P. Torr. Graph cut based inference with co-occurrence statistics. In *ECCV*, 2010. 1

[15] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006. 5

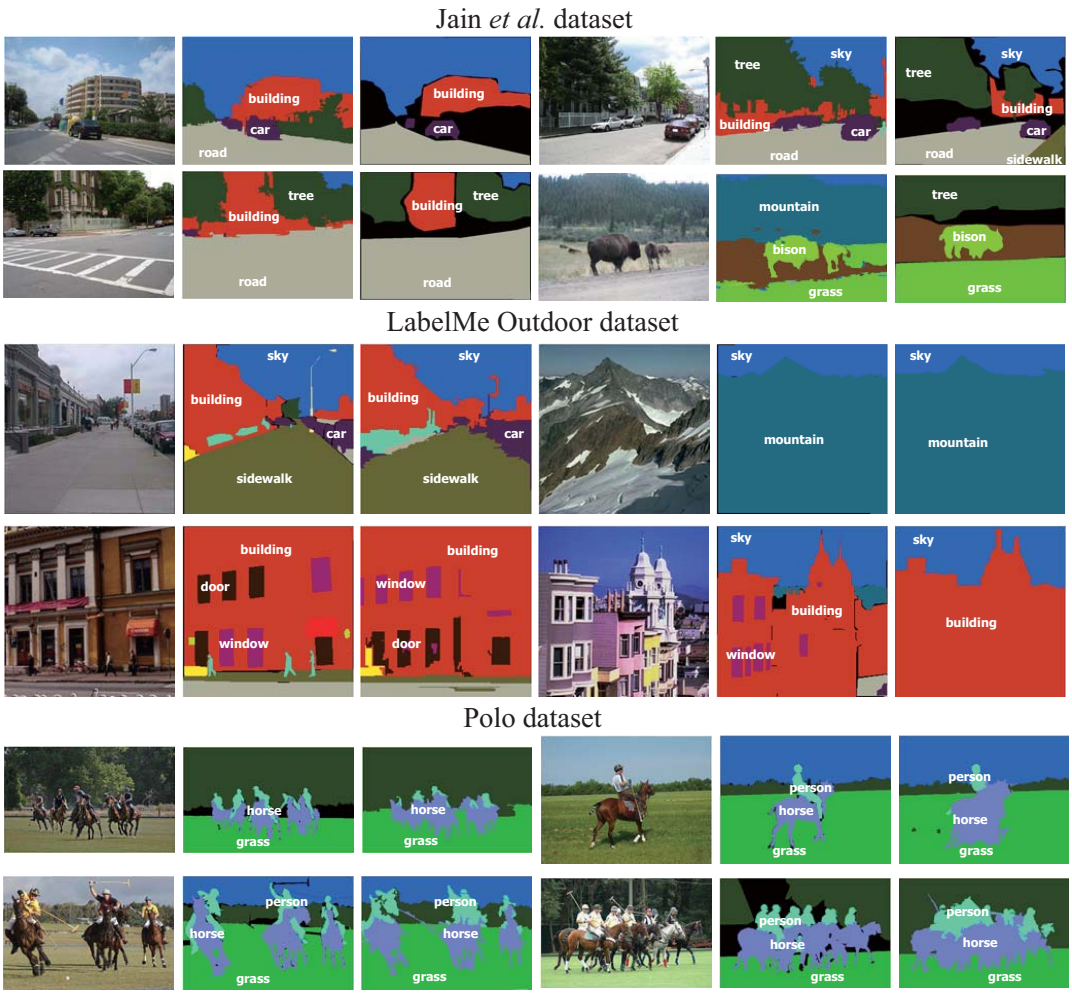


Figure 6. Example results from different datasets. The query images, ground truth, and results from our proposed (sum) are shown. Best viewed in color.

[16] C. Liu, J. Yuen, and A. Torralba. Nonparametric scene parsing: Label transfer via dense scene alignment. In *CVPR*, 2009. 1, 2, 5, 6, 7

[17] C. Liu, J. Yuen, A. Torralba, J. Sivic, and W. T. Freeman. Sift flow: Dense correspondence across different scenes. In *ECCV*, 2008. 2

[18] Z. Lu and H. H. Ip. Constrained spectral clustering via exhaustive and efficient constraint propagation. In *ECCV*, 2010. 4

[19] H. Myeong, J. Y. Chang, and K. M. Lee. Learning object relationships via graph-based context model. In *CVPR*, 2012. 1, 2, 6, 7

[20] A. Oliva and A. Torralba. Building the gist of a scene: the role of global image features in recognition. In *Progress in Brain Research*, 2006. 2, 5

[21] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. Labelme: A database and web-based tool for image annotation. *Int. J. Comput. Vision*, 77:157–173, May 2008. 7

[22] 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. *Int. J. Comput. Vision*, 81(1):2–23, Jan. 2009. 1

[23] J. Tighe and S. Lazebnik. Superparsing: Scalable nonparametric image parsing with superpixels. In *ECCV*, 2010. 1, 2, 5, 6, 7

[24] J. Tighe and S. Lazebnik. Understanding scenes on many levels. In *ICCV*, 2011. 1, 2, 5, 6

[25] H. Zhang, T. Fang, X. Chen, Q. Zhao, and L. Quan. Partial similarity based nonparametric scene parsing in certain environment. In *CVPR*, 2011. 1, 2, 6, 7

[26] H. Zhang, J. Xiao, and L. Quan. Supervised label transfer for semantic segmentation of street scenes. In *ECCV*, 2010. 1

[27] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf. Learning with local and global consistency. In *Advances in Neural Information Processing Systems*, 2004. 3