Semantic Segmentation Using Multiple Graphs with Block-Diagonal Constraints

Ke Zhang, Wei Zhang, Sheng Zeng, Xiangyang Xue
Shanghai Engineering Research Center for Video Technology and System
School of Computer Science, Fudan University, China
{k_zhang,weizh,zengsheng,xyxue}@fudan.edu.cn

Abstract

In this paper we propose a novel method for image semantic segmentation using multiple graphs. The multi-view affinity graph is constructed by leveraging the consistency between semantic space and multiple visual spaces. With block-diagonal constraints, we enforce the affinity matrix to be sparse such that the pairwise potential for dissimilar superpixels is close to zero. By a divide-and-conquer strategy, the optimization for learning affinity matrix is decomposed into several subproblems that can be solved in parallel. Using the $neighborhood\ relationship$ between superpixels and the $consistency$ between affinity matrix and label-confidence matrix, we infer the semantic label for each superpixel of unlabeled images by minimizing an objective whose closed form solution can be easily obtained. Experimental results on two real-world image datasets demonstrate the effectiveness of our method.

Introduction

Image semantic segmentation is a challenging and interesting task which aims to predict a label for every pixel in the image. Semantic segmentation is usually a supervised learning problem, in contrast to low-level unsupervised segmentation which groups pixels into homogeneous regions based on features such as color or texture (Lu et al. 2011).

In the past years, semantic segmentation has attracted a lot of attention (Kohli, Ladickỳ, and Torr 2009; Ladicky et al. 2009; 2010; Shotton et al. 2006; Shotton, Johnson, and Cipolla 2008; Yang, Meer, and Foran 2007; Jain et al. 2012; Lucchi et al. 2012; Ladicky et al. 2010). Most of these methods modeled the problem with a conditional random field(CRF) with different potentials. The basic approach was formulated in (Shotton et al. 2006), where a conditional random field (CRF) was defined over image pixels with unary potentials learned by a boosted decision tree classifier over texture-layout filters. The main research direction for successive publications focused on improving the CRF structure (Verbeek and Triggs 2007b; Yang, Meer, and Foran 2007; Jain et al. 2012; Lucchi et al. 2012). (Gould and Zhang 2012) performed semantic segmentation by constructing a graph of dense overlapping patch correspondences across large image sets. However, the above algorithms are far from perfectness and the imprecision of segmentation has an influence on labeling accuracy, which motivated approaches using multiple and hierarchical segmentations (Kumar and Koller 2010; Carreira and Sminchisescu 2010; Gonfaus et al. 2010; Ladicky et al. 2009; Munoz, Bagnell, and Hebert 2010; Wang et al. 2013). Furthermore, (Kohli, Ladickỳ, and Torr 2009) introduced hierarchy with higher order potentials, (Ladicky et al. 2010) integrated label co-occurrence statistics, and (Jain et al. 2012) learned a discriminative dictionary with supervised information using latent CRFs with connected hidden variables. (Lucchi et al. 2012) proposed a kernelized method via structured learning approaches which make it possible to jointly learn these CRF model parameters. Recently, a few works have been proposed to address the weakly supervised semantic segmentation problem, for which only the image-level annotations are available(Zhang et al. 2013; Verbeek and Triggs 2007a; Vezhnevets and Buhmann 2010; Vezhnevets, Ferrari, and Buhmann 2011).

Figure 1: Illustration of visual diversity and semantic confusion: car in (a) and car in (b) look quite dissimilar to each other; car in (b) and 'boat' in (c) look similar visually. (Best viewed in color.)

In semantic segmentation, each image is divided in to several regions called superpixels. Each superpixel can be described by multiple visual features. Each kind of feature has its fair share of pros and cons; and there is not a single kind of feature suitable for all semantic categories. Since images and superpixels can be described in multiple visual feature spaces, semantic segmentation may intuitively benefit from the integration of multiple representations. Among recent works on semantic segmentation, (Shotton, Johnson, and Cipolla 2008) showed quite fast and powerful feature via random decision forests that convert heterogeneous features to similar semantic texton histograms. (Tighe and Lazebnik

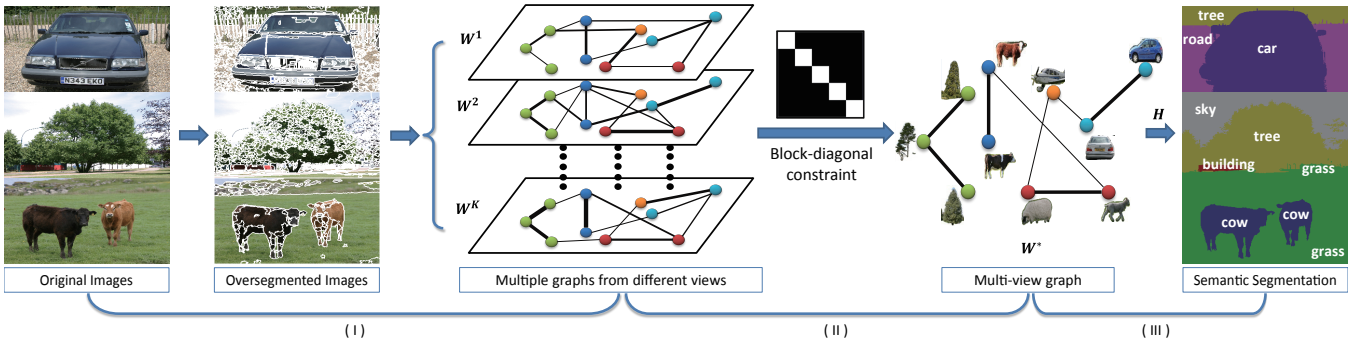

Figure 2: The overview of our framework. (I) Oversegment each image into superpixels, extract multiple features for each superpixel, and use the reconstruction weight from the neighboring superpixels as the affinity; (II) Learn the multi-view graph using the block-diagonal constraints and the consistency between semantic and visual spaces; (III) Infer superpixel labels by encouraging superpixels with similar appearance and position from images to share labels.

2010) leveraged a diverse and large set of visual features integrated in a weighted sum, where weights correspond to the usefulness of features. (Vezhnevets, Ferrari, and Buhmann 2012) introduced pairwise potentials among multi-feature images as components of CRF appearance model. However, to the best of our knowledge, there is no previous work that intensively explores relationships of multiple features in semantic segmentation.

The similarities between the same pair of superpixel may not be consistent when using different visual features; so we shall seek for an method to explore the consistency among multiple visual feature spaces. As in (Zhou and Burges 2007), one could construct an undirected (or directed) graph by inferring an affinity matrix from each type of image features, and then obtain multiple graphs of different views (there are multiple affinities between each pair of nodes). (Vedaldi et al. 2009) used multiple kernel learning to integrate diverse feature sets into one model. However, calculation of similarities solely based on visual features might lead to unsatisfying performance due to visual diversity and semantic confusion, i.e., superpixels similar in semantic space are not necessarily similar in visual feature space; on the other hand, superpixels similar in visual feature space are not always similar in semantic space, as seen in Fig.1. Like most tasks in computer vision, semantic segmentation also suffer from 'semantic gap'. The way to find a bridge over the 'semantic gap' is of significance to semantic segmentation based on visual features.

In this paper, we propose a novel method for semantic segmentation using multiple graphs with block-diagonal constraints. We perform dataset-wise segmentation using a affinity matrix which captures the similarity between every pair of superpixels. The affinity matrix is learned for different feature channels by leveraging various consistencies: (i) between semantic and visual spaces, (ii) between various features, and (iii) between weights and features. To infer semantic label for each superpixel of unlabeled images, we minimize an objective that (i) encourages the superpixels of the training images to be assigned their ground-truth labels; (ii) encourages adjacent superpixels in the same image to share a label; and (iii) encourages similar superpixels to be assigned a similar label (specifically, the distribution over the labels to be similar).

Fig.2 gives the overview of our framework. We firstly oversegment each image into superpixels, and extract multiple features for each superpixel. Secondly, we construct multi-view affinity graph whose weight measures similarity between superpixels. With block-diagonal constraints, the affinity matrix is sparse and of low rank. Finally, based on the affinity matrix and the position cue, the label for each superpixel can be inferred more precisely.

The rest of this paper is organized as follows: In the next section we firstly construct multi-view graph and learn the affinity matrix by decomposing the optimization problem into several subproblems which can be solved in parallel; secondly, we formulate the inference of superpixel label in a semi-supervised framework and obtain the closed-form solution of the optimal label-confidence matrix. We conduct experiments on MSRC and VOC2007 image datasets to demonstrate the effectiveness of our method. Finally, we give conclusions and suggestions for future work.

The Proposed Approach

Each image is represented as a set of superpixels, obtained by the existing oversegmentation algorithm (Comaniciu and Meer 2002). Suppose that the $i-$th image consists of N_i superpixels $I_i = \{x_{i,j}, y_{i,j}\}_{j=1}^{N_i}$, where $x_{i,j}$ denotes the $j-$th superpixel of $i-$th image, and $y_{i,j}$ denotes the corresponding labels $y_{i,j} = [y_{i,j}^1, \ldots, y_{i,j}^M]^\top \in \{0,1\}^M$. K kinds of features are extracted for each superpixel as $\{x_{i,j}^k\}_{k=1}^K$. Let $\mathcal{C} = \{c_1, \ldots, c_M\}$ be the semantic lexicon of M categories, and if the category c_m is associated with $x_{i,j}$, then $y_{i,j}^m = 1(m = 1, \ldots, M)$; otherwise, $y_{i,j}^m = 0$. Let $h_{i,j} \in [0,1]^M$ denote the label confidence vector for the superpixel $x_{i,j}$, and the $m-$th element of $h_{i,j}$ measures the probability that the superpixel $x_{i,j}$ belongs to the category c_m.

For the purpose of clarity, we further denote N as the total number of superpixels from all images, N_l and N_u as the number of labeled and unlabeled superpixels respec-

tively, i.e.,$N = N_l + N_u$, and $X^k = [x_1^k, \ldots, x_{N_l}^k, \ldots, x_N^k]$, $Y = [y_1, \ldots, y_{N_l}, \ldots, y_N]$, $H = [h_1, \ldots, h_{N_l}, \ldots, h_N]$, where $x_j^k \in \mathbb{R}^{P^k}$ is the $k-$th visual feature for superpixel x_j, y_j is the semantic label vector for x_j, and h_j is the label confidence vector for x_j.

Multi-View Affinity Graph Construction

In the task of semantic segmentation, each superpixel can be represented by multiple features (e.g., color, texture, and shape) which are heterogeneous although they are all visual descriptors. Each kind of visual feature describes the superpixel from a certain view, and heterogeneous features play different roles in describing various patterns, e.g., color and texture features for the concept 'water' while the shape feature for 'book'. We should consider learning from data with multiple views to effectively explore and exploit multiple representations simultaneously. For the same pair of superpixels, similarities measured by different visual features may not be consistent. Our goal is to learn an appropriate multi-view similarity which is as consistent with all similarities measured in different visual spaces as possible.

Inspired by (Roweis and Saul 2000), we assume that all superpixels lie on a locally linear embedding such that each superpixel can be approximately reconstructed by a linear combination of its neighbors. Intuitively, for a certain superpixel, those more similar samples will contribute more in reconstructing it; therefore, it is reasonable to look on reconstructing weights as the affinities between superpixels. Thus, we learn the multi-view affinity graph via an optimization problem formulated as follows:

$$\begin{aligned}
&\min_{W^1,\ldots W^K} \quad f(W^1,\ldots,W^K) = \\
&\sum_{k=1}^{K} \|X^k W^k - X^k\|_2 + \alpha \sum_{k=1}^{K} \sum_{i,j=1}^{N_l} (W_{i,j}^k - L_{i,j})^2 \\
&+ \beta \left(\sum_{i,j=1}^{N} \sqrt{\sum_{k=1}^{K} (W_{i,j}^k)^2} \right)^2 + \gamma \sum_{k=1}^{K} \|W^{k\top} W^k\|_1 \\
&s.t.\ W_{i,j}^k \geq 0, \sum_{i=1}^{N} W_{i,j}^k = 1,\ (k = 1 \ldots, K)
\end{aligned} \tag{1}$$

where $W^k \in [0,1]^{N \times N} (k = 1, \ldots, K)$ denotes the adjacency matrix of affinity graph whose entry $W_{i,j}^k$ measures pairwise similarity between superpixels represented by the $k-$th visual feature.

In the first term of Eq.(1), $X^k = [x_1^k, \ldots, x_N^k]$ whose $j-$th column corresponds to the $j-$th superpixel represented by the $k-$th visual feature, and $\|X^k W^k - X^k\|_2 = \sum_{j=1}^{N} \| \sum_{i=1}^{N} W_{ij}^k col(X^k, i) - col(X^k, j)\|$ which is the reconstruction error expressed in the Frobenius matrix norm. By constraining that $W_{j,j}^k = 0 (j = 1, \ldots, N)$, each superpixel can be estimated as a linear combination of other superpixels, which also avoids the case that the optimal W^k collapses to the identity matrix. As mentioned before, we learn the affinities between superpixels by using the reconstructing weights.

In the second term, $L_{i,j} \in \{1, 0\}$ measures the similarities between superpixels in the semantic space. More specifically, for those labeled superpixels, if superpixel i has the same category as superpixel j then $L_{i,j} = 1$ otherwise $L_{i,j} = 0$. Therefore, it is of significance to learn the appropriate W^k such that the gap $\sum_{i,j=1}^{N_l} (W_{i,j}^k - L_{i,j})^2$ becomes narrow. Minimizing the second term of Eq.(1) helps to reduce the semantic gap by achieving the consistency of similarities between semantic space and visual space.

Minimizing the third term of Eq.(1) is equivalent to encouraging that affinities across different graphs should be consistent to the largest extent. Actually, if $W^1, W^2, \ldots, W^K$ are concatenated together in the following form:

$$\widetilde{W} = \begin{bmatrix} W_{11}^1 & W_{12}^1 & \ldots & W_{NN}^1 \\ W_{11}^2 & W_{12}^2 & \ldots & W_{NN}^2 \\ \vdots & \vdots & \ddots & \vdots \\ W_{11}^K & W_{12}^K & \ldots & W_{NN}^K \end{bmatrix}$$

then the third term of Eq.(1) is just the $L_{2,1} - norm$ of $\widetilde{W}$, denoted by $\|\widetilde{W}\|_{2,1}$, i.e., $L_2 - norm$ for column firstly, and $L_1 - norm$ for row secondly. Minimizing the $L_2 - norm$ for each column makes the elements in the same column as equal as possible, while minimizing $L_1 - norm$ results in sparsity of $\widetilde{W}$, and then, all $W^k (k = 1, \ldots, K)$ are sparse consequently.

In the last term of Eq.(1), $\|W^{k\top} W^k\|_1 = \sum_{i,j=1}^{N} col(W^k, i)^\top col(W^k, j)$, herein $col(W^k, j)$ denotes the $j-$th column of W^k. Since $W_{i,j}^k \in [0, 1]$, minimizing $\|W^{k\top} W^k\|_1$ encourages $col(W^k, i)$ and $col(W^k, j)$ to be both sparse such that their inner product tends to be zero; what's more, minimizing $\|W^{k\top} W^k\|_1$ also enforces $W_{i,j}^k$ to be zero if the similarity between superpixels is too small such that W^k is block-diagonal when the superpixels are re-ordered(Wang et al. 2011).

Optimization

In the cost function Eq.(1), $W^k (k = 1, 2, \ldots, K)$ are all $N \times N$ matrices, thus the computational complexity in optimization is $O(K \times N^2)$. Fortunately, it can be converted into $K \times N$ sub-problems each of which operates on a single column of W^k with the complexity of $O(N)$. Since these sub-problems are independent of each other after conversion, parallel computation is carried out to accelerate the op-

timization process. Eq.(1) can also be expressed as follow:

$$
\begin{aligned}
&f(W^1,\ldots,W^K) = \\
&\sum_{k=1}^{K}\Bigg\{\alpha\sum_{i,j=1}^{N}\tau_{ij}((W_{ij}^k)^2-2W_{ij}^kL_{ij}+(L_{ij})^2)+ \\
&\Bigg(\sum_{j=1}^{N}x_j^{k\top}x_j^k\ -\ \sum_{j=1}^{N}2x_j^{k\top}\sum_{i=1}^{N}x_i^kW_{ij}^k+ \\
&\sum_{j=1}^{N}\sum_{p=1}^{P^k}(\sum_{i=1}^{N}x_i^k(p)W_{ij}^k)^2\Bigg)+\gamma\sum_{i=1}^{N}(\sum_{j=1}^{N}\ W_{ij}^k)^2\Bigg\}+ \\
&\beta\left(\sum_{i,j=1}^{N}\sqrt{\sum_{k=1}^{K}(W_{ij}^k)^2}\right)^2
\end{aligned}
\tag{2}
$$

where $\tau_{ij}=1$, for $i,j=1,\ldots,N_l$, and $\tau_{ij}=0$, for the rest. $x_i^k(p)$ denotes the $p-$th element of x_i^k. Like (Zhang et al. 2013), we use Cauchy-Schwarz Inequality $\left(\sum_{i=1}^n a_ib_i\right)^2 \le (\sum_{i=1}^n a_i^2)(\sum_{i=1}^n b_i^2)$ to obtain the upper bound of the cost function:

$$
\begin{aligned}
&f(W^1,\ldots,W^K) \le \\
&\sum_{k=1}^{K}\sum_{j=1}^{N}\Bigg\{x_j^{k\top}x_j^k+\alpha\sum_{i=1}^{N}(L_{ij})^2\tau_{ij}+ \\
&\qquad\sum_{i=1}^{N}\Bigg\{-2(x_i^{k\top}x_j^k+\alpha L_{ij}\tau_{ij})W_{ij}^k+ \\
&\qquad\left(\beta\frac{1}{Q_{ij}}+\sum_{p=1}\frac{(x_i^k(p))^2}{T_{ijp}^k}+\gamma\frac{1}{P_{ij}^k}+\alpha\tau_{ij}\right)(W_{ij}^k)^2\Bigg\}\Bigg\}
\end{aligned}
\tag{3}
$$

Eq.(3) holds for any $T_{ijp}^k, P_{ij}^k, Q_{ij} \in (0,1)$ satisfying $\sum_{i=1}^N T_{ijp}^k = 1$, $\sum_{j=1}^N P_{ij}^k = 1$, $\sum_{i,j=1}^N Q_{ij} = 1$. Specifically, the equality in Eq.(3) holds if and only if

$$
\begin{aligned}
T_{ijp}^k &= \frac{(x_i^k(p)W_{ij}^k)^2}{\sum_j^N(x_i^k(p)W_{ij}^k)^2};\quad P_{ij}^k = \frac{(W_{ij}^k)^2}{\sum_i^N(W_{ij}^k)^2}; \\
Q_{ij} &= \frac{\sum_{k=1}^K(W_{ij}^k)^2}{\sum_i^N\sum_j^N\sum_k^K(W_{ij}^k)^2};
\end{aligned}
\tag{4}
$$

Therefore, under the condition of Eq.(4), the original optimization problem is equivalent to minimizing the right side of Eq.(3), which can be furthermore divided into $K\times N$ independent quadratic programming sub-problems:

$$
\begin{aligned}
&\min_{W_{\cdot j}^k}\ \frac{1}{2}W_{\cdot j}^{k\top}\Lambda_j^kW_{\cdot j}^k+B_j^{k\top}W_{\cdot j}^k \\
&s.t.\ W_{\cdot j}^k\succeq 0,\ \mathbf{1}^\top W_{\cdot j}^k=1;
\end{aligned}
\tag{5}
$$

where $W_{\cdot j}^k$ denotes the $i-$th column of W^k whose element is non-negative, and $\mathbf{1}$ denotes an all-one vector. $\Lambda_j^k\in R^{N\times N}$ is a diagonal matrix whose $i-$th element on the diagonal $\lambda_{ii}=2\big(\beta\frac{1}{Q_{ij}}+\sum_p\frac{(x_i^k(p))^2}{T_{ijp}^k}+\gamma\frac{1}{P_{ij}^k}+\alpha\tau_{ij}\big)$. $B_j^k\in R^{N\times 1}$, with the $i-$th element $b_i=-2\left(x_i^{k\top}x_j^k+\alpha\, L_{ij}\tau_{ij}\right)$, $i,j=1,\ldots,N$. Such quadratic programming problem can be easily solved via the existing software solver MOSEK[1]. By iteratively solving the optimization problem in a flip-flop manner, i.e., updating $T_{ijp}^k, P_{ij}^k, Q_{ij}$ with Eq.(4) and updating W_{ij}^k with Eq.(5) alternatively until converge, we obtain the optimal affinity matrices: $W_k, k=1,2,\ldots,K$, then compute multi-view affinity graph as the average: $W^*=\frac{1}{K}\sum_{k=1}^K(W^k)$.

Label Inference

Based on the learned multi-view affinity graph, we can infer label for each superpixel of unlabeled images by estimating a label confidence matrix H, whose column h_j corresponds to the label confidence vector for superpixel x_j. The label confidence matrix H should be consistent with the learned multi-view affinity graph W^*, which encourages similar patches to take the same label over the entire dataset. At the same time, spatial relationship between superpixels should be leveraged as well. If two superpixels x_i and x_j are spatially adjacent in the same image, we define $S_{ij}=1$; otherwise $S_{ij}=0$. By using W^* and $S\in\{0,1\}^{N\times N}$ together, both appearance similarity and spatial neighborhood are taken into account in superpixel label inference, which is formulated as a semi-supervised framework:

$$
\begin{aligned}
\min_H\ \mathcal{Q}(H) =& \sum_{i=1}^{N_l}\|h_i-y_i\|^2+\theta_1\sum_{i,j=1}^{N}S_{ij}\|h_i-h_j\|^2 \\
&+\theta_2\sum_{i,j=1}^{N}W_{ij}^*\|\frac{h_i}{\sqrt{D_{ii}}}-\frac{h_j}{\sqrt{D_{jj}}}\|^2
\end{aligned}
\tag{6}
$$

where D is a diagonal matrix with $D_{ii}=\sum_{j=1}^N W_{ij}^*$, and $\theta_1,\theta_2>0$ are the trade-off parameters. The first term of Eq.(6) is the fitting constraint, which means a good label confidence matrix should be compatible with the ground-truth of the labeled samples. The second term is to encourage spatially smooth labelings. The third term is also smoothness constraint, which contains labeled as well as unlabeled superpixels. The second and the third terms indicate that superpixels with neighborhood relationship or similar appearance tend to share a label. The closed-form of optimal solution can be obtained as follows:

$$
\begin{aligned}
H^* = \frac{1}{1+\theta_1+\theta_2}(&I-\frac{\theta_1}{1+\theta_1+\theta_2}S \\
&-\frac{\theta_2}{1+\theta_1+\theta_2}D^{-1/2}W^*D^{-1/2})^{-1}Y
\end{aligned}
\tag{7}
$$

Once the optimal label confidence matrix H^* is estimated, the label for each superpixel can be easily inferred via a threshold.

[1]MOSEK: http://www.mosek.com

	building	grass	tree	cow	sheep	sky	aeroplane	water	face	car	bicycle	flower	sign	bird	book	chair	road	cat	dog	body	boat	average
(Shotton et al. 2006)	62	**98**	86	58	50	83	60	53	74	63	75	63	35	19	92	15	86	54	19	62	7	58
(Yang, Meer, and Foran 2007)	63	**98**	90	66	54	86	63	71	83	71	80	71	38	23	88	23	88	33	34	43	**32**	62
(Verbeek and Triggs 2007a)	52	87	68	73	84	94	88	73	70	68	74	89	33	19	78	34	89	46	49	54	31	64
(Shotton, Johnson, and Cipolla 2008)	49	88	79	**97**	**97**	78	82	54	**87**	74	72	74	36	24	93	51	78	75	35	66	18	67
(Ladicky et al. 2009)	**80**	96	86	74	87	**99**	74	**87**	86	87	82	**97**	**95**	30	86	31	**95**	51	69	66	9	75
(Csurka and Perronnin 2011)	75	93	78	70	79	88	66	63	75	76	81	74	44	25	75	24	79	54	55	43	18	64
(Lucchi et al. 2012)	59	90	**92**	82	83	94	91	80	85	**88**	**96**	89	73	**48**	96	**62**	81	**87**	33	44	30	76
Ours	68	**98**	**92**	86	82	96	**95**	84	85	86	89	94	73	32	**99**	58	90	82	**72**	**75**	26	**79**

Table 1: The accuracy of our method in comparison with other related competitive algorithms for individual labels on the MSRC-21 dataset. The last column is the average accuracy over all labels.

Figure 3: Semantic segmentation results of our method in comparison with the ground truth for some exemplary images from MSRC.

Experiments

We conduct the experiments on two real-world image datasets MSRC (Shotton et al. 2006) and VOC2007 (Everingham et al. 2007). On both datasets, we employ the Edge Detection and Image Segmentation (EDISON) system(Comaniciu and Meer 2002) to obtain the low-level segmentations. To get results from different quantization of images, 9 sets of parameters of the mean-shift kernels were randomly chosen as (5;5); (5;7); (5;9); (8;7); (8;9.5); (8;11); (12;10); (12; 15); (12;18). Then the final label prediction for each pixel can be computed as the harmonic mean of label confidences for multiple superpixels. Parameters α, β, γ are set by 10-fold cross-validation on the training set of each dataset for different segmentations. We extract the same visual features as in (Ladicky et al. 2009), i.e., Semantic Texton Forest(STF), color with 128 clusters, location with 144 clusters, and HOG descriptor(Dalal and Triggs 2005) with 150 clusters.

On MSRC-21 Dataset

The MSRC image dataset contains 591 samples of resolution 320×213 pixels, accompanied with a labeled object segmentation of 21 object classes. The training, validation and test subsets are 45%, 10%, and 45% of the whole image dataset, respectively.

Some examples of the segmentation results of our method in comparison with the ground-truth are given in Fig.3. Note that pixels on the boundaries of objects are usually labeled as background in the ground-truth. Table 1 shows the average accuracy of our method in compared with the state-of-the-art methods in (Shotton et al. 2006), (Yang, Meer, and Foran 2007), (Verbeek and Triggs 2007a), (Shotton, Johnson, and Cipolla 2008), (Ladicky et al. 2009), (Csurka and Perronnin 2011), and (Lucchi et al. 2012). For each category, the best result is highlighted in boldface. Our method performs better than other methods in most cases. Besides the best average performance, our method achieves the best performance for some categories, and keeps the second best for many of the rest. The results in Fig.3 and Table 1 both demonstrate the effectiveness of our method. In particular, due that our method learns an appropriate multi-view similarity consistent with various similarities computed by multiple visual features, it can adaptively select discriminant features, especially for those categories whose instances are similar in certain features. For example, the instances of $water$ are more similar in color and texture, the instances of $book$ are more similar in shape and texture, and the instances of $glass$ are more similar in color and texture. It can be seen that our method achieves more promising results especially on some categories such as $water$, sky, $book$, and $glass$.

On VOC-2007 Dataset

PASCAL VOC 2007 data set was used for the PASCAL Visual Object Category segmentation contest 2007. It contains 5011 training and 4952 testing images where only the bounding boxes of the objects present in the image are marked, and 20 object classes are given for the task of classification, detection, and segmentation. Rather on the 5011 annotated training images with bounding box indicating object location and rough boundary, we conduct experiments

	aeroplane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	dining table	dog	horse	motorbike	person	potted plant	sheep	sofa	train	tvmonitor	average
Brookes	6	0	0	0	0	9	5	10	1	2	11	0	6	6	29	2	2	0	11	1	6
(Shotton, Johnson, and Cipolla 2008)	66	6	15	6	15	32	19	7	7	13	44	31	**44**	27	39	35	12	7	39	23	24
(Ladicky et al. 2009)	27	**33**	**44**	11	14	36	30	**31**	27	6	**50**	28	24	38	52	29	28	12	45	46	30
(Csurka and Perronnin 2011)	**73**	12	26	**21**	**20**	0	17	**31**	**34**	6	26	41	7	31	34	30	11	28	5	50	25
TKK	19	21	5	16	3	1	**78**	1	3	1	23	**69**	**44**	**42**	0	**65**	30	**35**	**89**	**71**	31
$Ours$	65	25	39	8	17	**38**	17	26	25	**17**	47	41	**44**	32	**59**	34	**36**	23	35	31	**33**

Table 2: The accuracy of our method in comparison with other related competitive algorithms for individual labels on the VOC2007 dataset. The last column is the average accuracy over all labels.

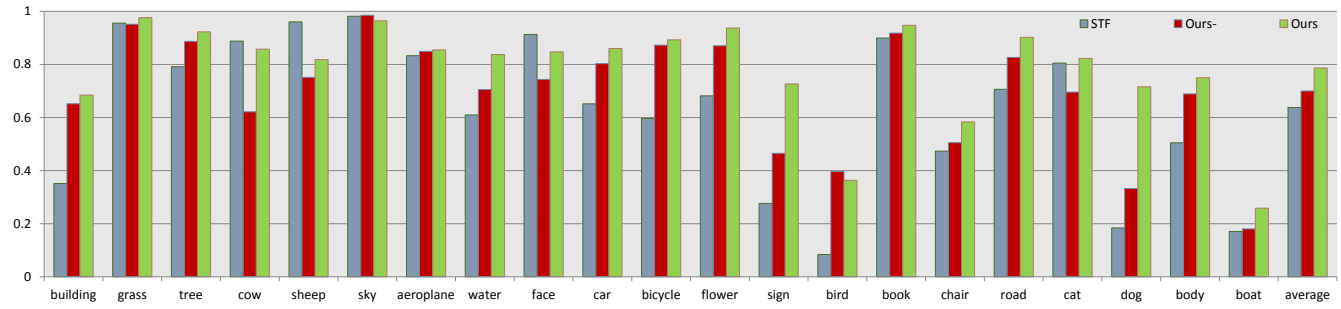

Figure 4: Comparison of our method $'Ours'$ with its degenerated variations of our method denoted by STF and $'Ours-'$ on MSRC-21 dataset. STF uses STF feature only; $'Ours-'$ uses the concatenation of all low level-features.

on the segmentation set with the 'train-val' split including 422 training-validation images and 210 test images, which are well segmented and thus are suitable for evaluation of the segmentation task.

The experimental results of our method compared with other related works are given in Table 2. The last column of 2 shows that the average accuracy of our method is better than all the others. For individual concepts, the performance of our method is better than or comparable to the state-of-art methods in most cases. Our method performs far better than the only segmentation entry (Brookes)(Everingham et al. 2007). Although our method uses much fewer training images than TKK(Everingham et al. 2007) which is trained by 422 training-validation images as well as a large number of annotated images with semantic bounding boxes from 5011 training sample, our method outperforms TKK in average. Evaluations on both MSRC and VOC2007 datasets sufficiently demonstrate the effectiveness of our method.

Multi-Graph Consistency Evaluation

To illustrate the significance of our method in capturing consistency among multiple visual feature spaces, we also evaluate two degenerated variations of our method denoted by STF and $'Ours-'$:

- STF: our method using Semantic Texton Forest(STF) feature only;
- $'Ours-'$: our method using a simple concatenation of all low level-features without capturing inter-feature consistency.

The comparison of performance is shown in Fig.4. In most cases, $'Ours-'$ outperforms STF by combining multiple features; $'Ours'$ outperforms both STF and $'Ours-'$ by effectively leveraging consistency of similarities across multiple visual feature spaces. In 16 out of 21 categories, 'Ours' achieves the best accuracy.

Conclusion

We address the problem of image semantic segmentation by encouraging superpixels with similar appearance or neighboring position to share a label. For each superpixel, different kinds of features are extracted. The sparse affinity matrix measuring similarity between superpixels for multiple feature channels can be learned by capturing the consistency between semantic space and multiple visual spaces. As for the future work, we plan to extend the proposed method to hierarchical segmentation, which might be another interesting direction of research.

Acknowledgement

We would like to thank the anonymous reviewers for their helpful comments. We would also like to thank Mr. Ruiqi Zhang for his help in experiments. This work was supported in part by the Shanghai Leading Academic Discipline Project (No.B114), the STCSM's Programs (No. 12XD1400900), the NSF of China (No.60903077), and the 973 Program (No.2010CB327906).

References

Carreira, J., and Sminchisescu, C. 2010. Constrained parametric min-cuts for automatic object segmentation. In *CVPR*.

Comaniciu, D., and Meer, P. 2002. Mean shift: A robust approach toward feature space analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 24(5):603–619.

Csurka, G., and Perronnin, F. 2011. An efficient approach to semantic segmentation. *International Journal of Computer Vision* 95(2):198–212.

Dalal, N., and Triggs, B. 2005. Histograms of oriented gradients for human detection. In *CVPR*.

Everingham, M.; Van Gool, L.; Williams, C.; Winn, J.; and Zisserman, A. 2007. The pascal visual object classes challenge 2007. In *http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html*.

Gonfaus, J.; Boix, X.; Van De Weijer, J.; Bagdanov, A.; Serrat, J.; and Gonzalez, J. 2010. Harmony potentials for joint classification and segmentation. In *CVPR*.

Gould, S., and Zhang, Y. 2012. Patchmatchgraph: Building a graph of dense patch correspondences for label transfer. In *ECCV*.

Jain, A.; Zappella, L.; McClure, P.; and Vidal, R. 2012. Visual dictionary learning for joint object categorization and segmentation. *ECCV*.

Kohli, P.; Ladickỳ, L.; and Torr, P. 2009. Robust higher order potentials for enforcing label consistency. *International Journal of Computer Vision* 82(3):302–324.

Kumar, M. P., and Koller, D. 2010. Efficiently selecting regions for scene understanding. In *CVPR*.

Ladicky, L.; Russell, C.; Kohli, P.; and Torr, P. 2009. Associative hierarchical crfs for object class image segmentation. In *ICCV*.

Ladicky, L.; Russell, C.; Kohli, P.; and Torr, P. 2010. Graph cut based inference with co-occurrence statistics. *ECCV*.

Lu, Y.; Zhang, W.; Lu, H.; and Xue, X. 2011. Salient object detection using concavity context. In *ICCV*.

Lucchi, A.; Li, Y.; Smith, K.; and Fua, P. 2012. Structured image segmentation using kernelized features. *ECCV*.

Munoz, D.; Bagnell, J. A.; and Hebert, M. 2010. Stacked hierarchical labeling. In *ECCV*.

Roweis, S. T., and Saul, L. K. 2000. Nonlinear dimensionality reduction by locally linear embedding. *Science* 290(5500):2323–2326.

Shotton, J.; Winn, J.; Rother, C.; and Criminisi, A. 2006. Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. *ECCV*.

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

Tighe, J., and Lazebnik, S. 2010. Superparsing: Scalable nonparametric image parsing with superpixels. In *ECCV*.

Vedaldi, A.; Gulshan, V.; Varma, M.; and Zisserman, A. 2009. Multiple kernels for object detection. In *ICCV*.

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

Verbeek, J., and Triggs, W. 2007b. Scene segmentation with crfs learned from partially labeled images. In *NIPS*.

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

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

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

Wang, S.; Yuan, X.; Yao, T.; Yan, S.; and Shen, J. 2011. Efficient subspace segmentation via quadratic programming. *AAAI*.

Wang, X.; Lin, L.; Huang, L.; and Yan, S. 2013. Incorporating structural alternatives and sharing into hierarchy for multiclass object recognition and detection. In *CVPR*.

Yang, L.; Meer, P.; and Foran, D. 2007. Multiple class segmentation using a unified framework over mean-shift patches. In *CVPR*.

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

Zhou, D., and Burges, C. 2007. Spectral clustering and transductive learning with multiple views. In *ICML*.