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, Ladicky, 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 correspon-
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 to share a label; and (ii) 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 $C = \{c_1, \ldots, c_M\}$ 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, $N_I$ and $N_u$ as the number of labeled and unlabeled superpixels respec-
tively, i.e., \(N = N_L + N_R\), and \(X^k = [x^k_1, \ldots, x^k_{N_L}, \ldots, x^k_N]\), \(Y = [y_1, \ldots, y_{N_L}, \ldots, y_N]\), \(H = [h_1, \ldots, h_{N_L}, \ldots, h_N]\), where \(x^k_j \in \mathbb{R}^{N_k}\) 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 repre-

In the cost function Eq.(1), \(\min_{W^1, \ldots, W^K} f(W^1, \ldots, W^K) = \sum_{k=1}^{K} ||X^kW^k - X^k||_2 + \alpha \sum_{k=1}^{K} \sum_{i,j=1}^{N_L} (W^k_{i,j} - L_{i,j})^2 \beta \sum_{i,j=1}^{N_L} (W^k_{i,j})^2 + \gamma \sum_{k=1}^{K} ||W^k||_1 \) 

\(s.t. W^k_{i,j} \geq 0, \sum_{i=1}^{N} W^k_{i,j} = 1, (k = 1, \ldots, K)\)  

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

In the first term of Eq.(1), \(X^k = [x^k_1, \ldots, x^k_N]\) whose \(j\)-th column corresponds to the \(j\)-th superpixel represented by the \(k\)-th visual feature, and \(||X^kW^k - X^k||_2 = \sum_{j=1}^{N} ||\sum_{i=1}^{N} W^k_{i,j} \text{col}(X^k, i) - \text{col}(X^k, j)||\) which is the reconstruction error expressed in the Frobenius matrix norm. By constraining that \(W^k_{i,j} = 0(j = 1, \ldots, N)\), 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 appropri-

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:

\[
\tilde{W} = \begin{bmatrix}
W^1_{11} & W^1_{12} & \cdots & W^1_{1N} \\
W^2_{11} & W^2_{12} & \cdots & W^2_{1N} \\
\vdots & \vdots & \ddots & \vdots \\
W^K_{11} & W^K_{12} & \cdots & W^K_{1N}
\end{bmatrix}
\]

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

In the last term of Eq.(1), \(||W^kW^k\|_1 = \sum_{i,j=1}^{N} \text{col}(W^k, i)^T \text{col}(W^k, j)\), herein \(\text{col}(W^k, j)\) denotes the \(j\)-th column of \(W^k\). Since \(W^k_{i,j} \in [0,1]\), minimizing \(||W^kW^k\|_1\) encourages \(\text{col}(W^k, i)\) and \(\text{col}(W^k, j)\) to be both sparse such that their inner product tends to be zero; what’s more, minimizing \(||W^kW^k\|_1\) also enforces \(W^k_{i,j}\) 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:

\[
f(W^1, \ldots, W^K) = \sum_{k=1}^{K} \left\{ \alpha \sum_{i,j=1}^{N} \tau_{ij}((W^k_{ij})^2 - 2W^k_{ij}L_{ij} + (L_{ij})^2) + \beta \left( \sum_{i,j=1}^{N} \sum_{i,j=1}^{N} x^k_{ij}(p)W^k_{ij} \right)^2 \right\}
\]

\[
\text{where } \tau_{ij} = 1, \text{for } i, j = 1, \ldots, N, \text{ and } \tau_{ij} = 0, \text{ for the rest.}
\]

\[
x^k_{ij} \text{ denotes the } p\text{-th element of } x^k_{ij}. \text{ Like (Zhang et al. 2013), we use Cauchy-Schwarz Inequality } \left( \sum_{i=1}^{n} a_i b_i \right)^2 \leq \left( \sum_{i=1}^{n} a_i^2 \right) \left( \sum_{i=1}^{n} b_i^2 \right) \text{ to obtain the upper bound of the cost function:}
\]

\[
f(W^1, \ldots, W^K) = \sum_{k=1}^{K} \sum_{i,j=1}^{N} \left\{ \alpha \sum_{i=1}^{N} \left( L_{ij} \right)^2 \tau_{ij} + \frac{\beta}{1 + \tau_{ij}} \left( \sum_{p=1}^{P} x^k_{ij}(p) \right)^2 \right\}
\]

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

\[
T^k_{ijp} = \frac{(x^k_{ij}(p)W^k_{ij})^2}{\sum_{i,j=1}^{N} (x^k_{ij}(p)W^k_{ij})^2}, \quad P^k_{ij} = \frac{(W^k_{ij})^2}{\sum_{i,j=1}^{N} (W^k_{ij})^2}, \quad Q_{ij} = \frac{\sum_{i,j=1}^{N} \sum_{k=1}^{K} (W^k_{ij})^2}{\sum_{i,j=1}^{N} \sum_{k=1}^{K} (W^k_{ij})^2};
\]

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:

\[
\min_w \frac{1}{2} \sum_{i,j=1}^{N} \sum_{k=1}^{K} (x^k_{ij})^2 + \frac{\beta}{1 + \tau_{ij}} \left( \sum_{p=1}^{P} x^k_{ij}(p) \right)^2 \text{ s.t. } W^k_{ij} \succeq 0, \omega^T W^k_{ij} = 1;
\]

where \( W^k_{ij} \) denotes the \( i\)-th column of \( W^k \) whose element is non-negative, and \( I \) denotes an all-one vector. \( \Lambda^k_{ij} \in \mathbb{R}^{N \times N} \) is a diagonal matrix whose \( i\)-th element on the diagonal \( \lambda_{ii} = 2(\beta_{ij} + \gamma_{ij}) \). \( B^k_{ij} \in \mathbb{R}^{N \times 1} \),

with the \( i\)-th element \( b_{ij} = -2(\alpha_{ij}^k + \alpha_{ij}) \), \( i, j = 1, \ldots, N \). 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 \( T^k_{ijp}, P^k_{ij}, Q_{ij} \) with Eq.(4) and updating \( W^k_{ij} \) 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 \( h_i \) corresponds to the label confidence vector for superpixel \( x_i \). 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:

\[
\min_H Q(H) = \sum_{i=1}^{N} \| h_i - y_i \|^2 + \theta_1 \sum_{i,j=1}^{N} S_{ij} \| h_i - h_j \|^2 + \theta_2 \sum_{i,j=1}^{N} W^*_{ij} \left( \frac{h_i}{\sqrt{D_{ii}}} - \frac{h_j}{\sqrt{D_{jj}}} \right)^2
\]

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 spatial 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:

\[
H^* = \frac{1}{1 + \theta_1 + \theta_2} \left( I - \frac{\theta_1}{1 + \theta_1 + \theta_2} S - \frac{\theta_2}{1 + \theta_1 + \theta_2} D^{-1/2} W^* D^{-1/2} \right) - 1 Y
\]

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

\(^1\text{MOSEK: http://www.mosek.com}\)
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.

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

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
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.

<table>
<thead>
<tr>
<th></th>
<th>aeroplane</th>
<th>bicycle</th>
<th>bird</th>
<th>boat</th>
<th>bottle</th>
<th>bus</th>
<th>car</th>
<th>cat</th>
<th>chair</th>
<th>cow</th>
<th>dining table</th>
<th>dog</th>
<th>horse</th>
<th>motorbike</th>
<th>person</th>
<th>plant</th>
<th>sheep</th>
<th>sofa</th>
<th>train</th>
<th>tvmonitor</th>
<th>average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brookes</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>9</td>
<td>5</td>
<td>10</td>
<td>1</td>
<td>2</td>
<td>11</td>
<td>0</td>
<td>6</td>
<td>6</td>
<td>29</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>11</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>(Shotton, Johnson, and Cipolla 2008)</td>
<td>66</td>
<td>15</td>
<td>15</td>
<td>6</td>
<td>13</td>
<td>32</td>
<td>19</td>
<td>7</td>
<td>7</td>
<td>13</td>
<td>44</td>
<td>44</td>
<td>27</td>
<td>39</td>
<td>35</td>
<td>12</td>
<td>7</td>
<td>39</td>
<td>23</td>
<td>24</td>
<td></td>
</tr>
<tr>
<td>Ladicky et al. 2009</td>
<td>27</td>
<td>12</td>
<td>36</td>
<td>30</td>
<td>31</td>
<td>6</td>
<td>50</td>
<td>24</td>
<td>38</td>
<td>32</td>
<td>29</td>
<td>28</td>
<td>45</td>
<td>46</td>
<td>30</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Csurka and Perronnin 2011)</td>
<td>73</td>
<td>26</td>
<td>31</td>
<td>4</td>
<td>6</td>
<td>26</td>
<td>41</td>
<td>7</td>
<td>31</td>
<td>34</td>
<td>30</td>
<td>11</td>
<td>28</td>
<td>5</td>
<td>50</td>
<td>25</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TKK</td>
<td>19</td>
<td>21</td>
<td>15</td>
<td>7</td>
<td>1</td>
<td>17</td>
<td>23</td>
<td>44</td>
<td>31</td>
<td>34</td>
<td>30</td>
<td>11</td>
<td>28</td>
<td>35</td>
<td>23</td>
<td>31</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ours</td>
<td>65</td>
<td>25</td>
<td>39</td>
<td>8</td>
<td>17</td>
<td>38</td>
<td>17</td>
<td>26</td>
<td>25</td>
<td>17</td>
<td>47</td>
<td>41</td>
<td>44</td>
<td>32</td>
<td>59</td>
<td>34</td>
<td>36</td>
<td>23</td>
<td>35</td>
<td>33</td>
<td></td>
</tr>
</tbody>
</table>

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


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


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


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.


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.


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


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, 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.
