Object Extraction from Bounding Box Prior with Double Sparse Reconstruction

Lingzheng Dai, Jundi Ding, Jian Yang, Fanlong Zhang and Junxia Li
Dept. of CSE, Nanjing University of Science and Technology, Nanjing 210094, PR China
lingzhengdai@163.com, {dingjundi2010,csjyang}@njust.edu.cn, csfzhang@126.com, junxia1i99@163.com

Abstract

Extracting objects from natural images has long been an active problem in image processing. Despite various attempts, it has not been completely solved up to date. Current state-of-the-art object proposal methods tend to extract a set of object segments from an image, and often these are consequential differences among these results for each image. Another type of methods strive to detect one object into a bounding box where some background parts are often covered. For these two methodologies, we observe: 1) there are generally some regions overlapped among different proposals, which are usually from one object; they could be as object ‘segment hypotheses’; 2) pixels outside the detected bounding box could be as ‘background hypotheses’ as they are with high probability from the background. With them, we formulate the object extraction as a “double” sparse reconstruction problem in terms of the bounding box results. The idea is that object regions should be with small reconstruction errors to segment hypotheses bases, simultaneously, they should have large reconstruction errors to background hypotheses bases. Comprehensive experiments and evaluations on PASCAL VOC object segmentation dataset and GrabCut-50 database demonstrate the superiority of our built method. In particular, we achieve the state-of-the-art performance for the object segmentation with bounding box prior on these two benchmark datasets.

1. Introduction

Extracting objects from natural images has long been one of the most fundamental and critical problems in computer vision and image processing [3][18][28]. It plays a key role in vision applications, including object recognition, classification [11][27][29][30], etc. However, experiments on PASCAL [32] or GrabCut-50 [12] database show that it is still an unsolved and challenging problem. This is mainly due to that photographs of natural scenes reflect real-world variations and are characterized by large ranges of color, texture and shapes.

Two paradigms have shaped this field of object extrac-
Figure 2. Framework of our double sparse reconstruction based segmentation model. Given an image and bounding box prior of the object, we first segment the image into super-pixels. Super-pixels outside the bounding box and many object proposals generated by CPMC method, respectively, are used to construct the background and segment hypotheses. Background hypotheses are used to reconstruct the background reconstruction map (BRM), and meanwhile the segment reconstruction map (SRM) is reconstructed through segment hypotheses. The output of BRM and SRM are obtained by the proposed double sparse reconstruction framework, including the reconstruction coefficients. They are then integrated to yield the final object confidence map. Region merging procedure is finally applied on the object confidence map to make the extracted object with a well-preserved boundary.
use some brush strokes, or bounding box prior to predict starting seeds or locations of objects are related to our work.

**Seed-based approaches**: The seed-based methods include: GraphCut (GC) [15], constrained parametric min-cut (CPMC) [2] and Laplacian Coordinates [24]. Considering the image as a graph, GC seeks to find the minimum cut between seeded regions, where the similarity between neighboring pixels is encoded as edges of this weighted graph. GC uses a max-flow/min-cut algorithm to find this cut in order to segment images. Using a graph-cut based model, Carreira et al. [2] seek to generate a pool of object hypotheses by hypothesizing a set of placements of fore- and background seeds. For each configuration, segmentation results are obtained by solving a constrained parametric min-cut. Particularly, in their model the smoothness term borrows the definition from gPb [10] of similarity between adjacent pixels. Recently, Laplacian Coordinates (LC)[24] is proposed to minimize a novel quadratic energy function defined from an affinity graph of pixels. In their model, the average distance of pairwise pixels is minimized and anisotropic propagation of seeds labels is controlled well. Generally, careful assignment of these seeds is a non-trivial job, which influences the segmentation performance critically.

**Bounding box based approaches**: Some work, on the other hand, use the bounding box prior to guide object segmentation, including [16] [12] [13] [36]. Compared with seed strokes, bounding box prior is intuitive to users due to its availability of taking only tow mouse clicks and the emerging of object-detection techniques. In GrabCut [16], this bounding box prior is integrated into the energy function and the model is iteratively optimized by Expectation Maximization (EM) method. Further, the object boundary is refined by border matting in order to get the final segmentation results. The authors [12] further presented a new graph-cut framework. They investigate the effectiveness of the sufficiently tight bounding box and integrate this information as a constraint into their energy function. To optimize their model, a new rounding algorithm - pinpointing is handed as the optimization strategy. In [13], segmentation task is tackled as an adaptive figure-ground classification algorithm using a user provided bounding box. It compiles various foreground priors and one common background prior seamlessly. With the different foreground priors, many hypotheses are generated with evaluation score functions. At last, the one with the maximum segmentation quality score is selected as the best segmentation. Recently, Tang et al. [36] propose an alternative approach to color clustering using kernel K-means energy. Compared with histogram or GMM fitting used by [16], they argue that the fore/background regions can be clustered better using this energy. Probably the most similar work to us is Xia [6], which proposes to generate the object shape by directly selecting the best overlapping segments that align well to the object boundary and thereafter integrate it into the subsequent graph-cut based inference algorithm to obtain the segmentation results. Segmentation performance of this method heavily relies on the shape based graph-cut process. We utilize the generated object candidates as segment hypotheses bases. However, unlike [6], we operationalize this idea by exploring the usefulness of each segment towards object extraction based on an object reconstruction model. Furthermore, the pairwise correlation information of segment hypotheses can be preserved in our method, which is crucial to produce accurate and reliable results.

**3. Object Extraction via Double Sparse Reconstruction**

In this section, we present the proposed method in detail. Given a test image, we first segment it into super-pixels. Then for each bounding box input, the super-pixels outside the bounding box are used to construct background hypotheses bases. Meanwhile, we compute a large pool of object candidates to construct segment hypotheses bases for each image, using the publicly available Constrained Parametric Min-Cuts algorithm (CPMC) [2]. These two bases are integrated into our model as reconstruction bases for predicting the object confidence map. The obtained confidence map is further refined through some techniques including multi-scale strategy and region merging for extracting object entirely.

**3.1. Background Reconstruction**

When the bounding box prior is provided, although some background pixels are covered in it, it can be observed that the pixels outside it are with high probability from the background, as shown in Fig. 3. This means that some background regions can be easily identified. We apply SLIC [7] method to the test image and segment it into many super-pixels. Then the super-pixels from outside bounding box are used to construct the background hypotheses. Intuitively, using background hypotheses as bases to reconstruct the foreground and background regions, the reconstruction errors between them shall be different. For this reason, we seek to distinguish the foreground from the background based on a sparse reconstruction model.

The first reconstruction map is designed to be reconstructed by using background hypotheses as bases. Formally, let an image $X$ formed by initial super-pixels, i.e., $X = [x_1, x_2, \ldots, x_M] \in \mathbb{R}^{N \times M}$, where $M$ is the number of super-pixels and $N$ is the feature dimension. Let us represent each super-pixel with mean color features and coordinates, i.e., $x_i = \{L, a, b, R, G, B, x, y\}$, where both Lab and RGB color spaces are used to describe its features, and $x, y$ denote its coordinates.

With the bounding box, the background hypotheses bases are formally formed as: $A = [a_1, a_2, \ldots, a_M] \in \mathbb{R}^{N \times M}$.
\[ \mathbf{a}_i = \mathbf{x}_i \text{ if } \mathbf{a}_i \text{ belongs to the bounding box outside regions, otherwise } \mathbf{a}_i = \mathbf{0}. \] For an image, we seek to represent each \( \mathbf{x} \) by using an over-complete dictionary whose vectors are background bases themselves, i.e., \( \mathbf{x} = \mathbf{A} \mathbf{y} \). It can be sought by solving the following optimization problem:

\[
\hat{y}_i = \arg \min_{\mathbf{y}_i} \frac{1}{2} \| \mathbf{x}_i - \mathbf{A} \mathbf{y}_i \|_2^2 + \lambda_1 \| \mathbf{y}_i \|_1 \tag{1}
\]

where \( \lambda_1 \) is the regularized parameter and \( \ell_1 \) penalty can yield a sparse solution for \( \mathbf{y}_i \). The \( \ell_2 \) norm is used to minimizing the distance between the prediction reconstruction and each super-pixel. After obtaining the solution \( \hat{y}_i \), it is easily to design a sparse representation based classifier [1] in terms of its reconstruction residual. The corresponding reconstruction residual is defined by

\[
r(\hat{y}_i) = \| \mathbf{x}_i - \mathbf{A} \hat{y}_i \|_2 \tag{2}
\]

An example of \( \mathbf{Y} = [\mathbf{\hat{y}}_1, \mathbf{\hat{y}}_2, \ldots, \mathbf{\hat{y}}_M] \) is presented in Fig. 3, from which we can see that all the representation of \( \mathbf{y}_i \) of super-pixels outside the bounding box are located in the diagonal of the affinity matrix, while the regions within the bounding box are not in the diagonal. This means that the super-pixels outside the bounding box are represented by themselves, and super-pixel within the bounding box is represented by linear combination of the background hypotheses bases \( \mathbf{A} \). The combination coefficients are the elements of \( \mathbf{y}_i \), with larger magnitude showing brighter.

After obtaining the representation \( \mathbf{Y} \), we can yield the \( r(\hat{y}_i) \) for each super-pixel by (2). Within the bounding box, the super-pixels that belong to background regions can be well reconstructed by background hypotheses bases \( \mathbf{A} \) through (1) and thus they may have small sparse reconstruction errors. On the contrary, the super-pixels that belong to object regions derive large sparse reconstruction errors, which gives us a straightforward way to express each super-pixel in the image with its reconstruction residual. We normalize the reconstruction residual value of each super-pixel into \([0, 1]\). Super-pixels that yield larger values are shown in brighter, meanwhile that have smaller values will illustrate darker, as shown in Fig. 3. We refer to it as background reconstruction map (BRM).

However, the reconstruction residual measure that drives the BRM may not be robust to images that contain multiple instances of similar objects: super-pixels outside the bounding box would look similar to super-pixels inside it and yield low residuals for foreground. In this case, the reconstruction map can benefit from object candidates.

### 3.2. Segment Reconstruction

Give an input image, object proposal methods can generate a pool of visually consistent object-level segments, as shown in Fig. 4. It can be observed that there are often some regions overlapped among different segments, which are usually from one object. We are interested in using these candidates to construct segment hypotheses bases. The CPMC method [2] is used to generate the set of object proposals. Note that the procedure of ranking or classification of generated segments of [2] is not applied in our work. The second reconstruction map is designed to be reconstructed by using segment hypotheses as bases.

Different from previous work, such as [17] [6], the way we pursue is not only to make such segments more powerful by summing them but also to exploit the cross information between them, adapted to the sparse reconstruction model. To achieve this, extracting objects from images is formulated as a segment reconstruction problem, where the reconstruction map of each super-pixel is expressed as a linear combination of generated segment hypotheses bases, referred as Segment Reconstruction Map (SRM).

Suppose \( S_1, S_2, \ldots, S_c \subset \mathbb{R}^2 \) be the regions of the remaining segments cropped by the bounding box, let \( T_i : \mathbb{R}^2 \to \{0, 1\} \) be the characteristic function of
each super-pixel $r_j$ for all $j = 1, \ldots, M$ of $S_i$. Then we use vector $S_{B_i} = [T_i(r_1), T_i(r_2), \ldots, T_i(r_M)]^T \in \mathbb{R}^M$ to represent each segment hypothesis $S_i$, and $S_B = [S_{B_1}, S_{B_2}, \ldots, S_{B_M}]^T \in \mathbb{R}^{c \times M}$ be all segment hypotheses bases. An example of generating $S_B$ is illustrated in Fig. 4.

After obtaining the $S_B$, we can yield the averaging map of it. With them, $d_i$ that represents each super-pixel of $S$-RM is predicted by finding the best linear combination of segment hypotheses bases. The SRM of all super-pixels are denoted as $D = [d_1, d_2, \ldots, d_M] \in \mathbb{R}^{c \times M}$, and we can obtain it by solving the following optimization problem:

$$
\{\hat{y}_i, \hat{d}_i\} = \arg \min_{d_i, y_i} \left\| d_i - S_B y_i \right\|_2 + \lambda_1 \left\| y_i \right\|_1 + \lambda_2 \left\| d_i - S_B (i) \right\|_2^2
$$

where $\lambda_3$ is the regularized parameter and $S_B(i) \in \mathbb{R}^c$ represents the average map of $S_B$, its each element is the average magnitude of $i$-th column vector of $S_B$.

Within the bounding box, the super-pixels that belong to the object regions can be well reconstructed by segment hypotheses bases $S_B$ and thus have small sparse reconstruction errors. The elements of $y_i$ corresponding to the object can be identified by (3). We predict the SRM not only by minimizing $\ell_2$ distance between the prediction reconstruction map and the average map, but also finding the best linear combination of segment hypotheses bases to build it. The SRM of all super-pixels are together predicted by using the designed construction framework which is capable of capturing the correlations among all segment hypotheses. Fig. 5 depicts some SRM results, from which we can see that super-pixels from object regions exhibit brighter color than that of background components, indicating that they can be reconstructed well by segment hypotheses bases.

### 3.3. Double Sparse Reconstruction

To make full use of all the information produced by the set of generated segment hypotheses and extracted background regions seamlessly, the derived $S_B$ and $A$ should be integrated into an unified framework of object reconstruction. Here, our consideration for formulating the inference process is two-side: to inherit the advantages of sparse representation, the representation of the background and segment hypotheses bases is encouraged to be sparse; to make use of the cross-information of segment hypotheses, the segment reconstruction error of each super-pixel should be enforced to be sparsity-consistent simultaneously. By considering both sides, the joint reconstruction is achieved via the following problem:

$$
\{\hat{y}_i, \hat{d}_i\} = \arg \min_{y_i, d_i} \frac{1}{2} ||x_i - Ay_i||_2^2 + \lambda_1 ||y_i||_1 + \\
\lambda_2 ||d_i - S_B y_i||_1 + \lambda_3 ||d_i - S_B(i)||_2^2
$$

where, $\alpha$ is a parameter to balance the effect of two reconstruction map. Since the confidence map learnt from the reconstruction framework is defined at super-pixel level, objects are tend to be fragmented with heterogeneous parts and strong internal contours. In order to cope with this issue, we compute the super-pixels of the image at a multi-scale strategy. Then, the final object confidence map is calculated by averaging them in order to tackle the large ranges of object color, texture, shape, or other attributes.

After obtaining the object confidence map for an image, the result is projected back to the image. To obtain the object extraction result, the most easily choice is to directly use the map with above a threshold as object and the residual as the background. However, it is hard to set a fixed threshold to find the object for each image due to natural images are very complex.

As proposed in [17], self-similarly can be used to refine the segmentation. In this part, the top high scores serve as object seeds and the segmentation result can be then obtained by a region merging strategy. To achieve this, we use the method of [26] to merge the initial regions for more precise object extraction result.

### 3.4. Optimization Process

We aim to optimize Eq. (4). Obviously, there are only two parameters to optimize. We propose to optimize it w.r.t.
representation coefficient $y_i$ and $d_i$ alternatively. The optimization procedure is keeping $d_i$ fixed to optimize $y_i$, and keeping $y_i$ fixed to optimize $d_i$ iteratively, repeating until convergence.

Specifically: it consists of the following iterations:
(i) Given $d_i = d_i^k$, optimize Eq. (4), update $y_i^{k+1} \leftarrow y_i^k$.
(ii) Given $y_i = y_i^k$, optimize Eq. (4), update $d_i^{k+1} \leftarrow d_i^k$.

The detailed Optimization Process of Eq. (4) is presented in supplementary material.

4. Experiments

In this section, we study the quality of the proposed double sparse reconstruction method for object extraction. We conduct comprehensive experiments on two publicly available datasets: PASCAL VOC object segmentation dataset [32] and GrabCut-50 dataset [12].

Confidence Map Details: This part presents the implementation details of final confidence map generation. As the confidence map learnt from Eq. (4) is processed at super-pixel level, objects are tend to be fragmented with heterogeneous parts and strong internal contours. In order to cope with this, we compute the super-pixels of an image at a multi-scale strategy. For each image, we first perform over-segmentation by SLIC [7] at eight different scales, with super-pixel number set respectively from 50 to 400. Then we run the double sparse reconstruction method eight times, and obtain the object confidence map by averaging the eight results. We set $\lambda_1 = 0.01$, $\lambda_2 = 0.01$, $\lambda_3 = 0.01$ and $\mu = 1$ in all experiments.

Some final object confidence maps are displayed in Fig. 6. It can be observed that pixels from object regions exhibit brighter color, further advocating the value of our method compared with directing merging segment hypotheses.

4.1. GrabCut-50

Comparison with Bounding Box based Methods: In this part, we provide comparisons against various bounding box based segmentation methods. We first conduct experiments on the popular interaction image segmentation dataset GrabCut-50, which is provided by [12] and includes 50 images with ground truth bounding boxes.

We use error-rate to evaluate the segmentation performance for different approaches, which is defined as the percentage of mislabeled pixels inside the bounding box. We compare our double sparse reconstruction method with following state-of-the-art: [16] [12] [13] [36] [6]. Results on this benchmark are reported on Table 1. Our proposed method achieves the mean score 3.2% on this database, significantly outperforming the two bounding box based methods (GrabCut [16] and Tang et al. [36]). When more interactive priors are used, GrabCut-Pinpoint [12] and F-G Classification [13] on this dataset can achieve better performance. Our method also obtain lower error score (3.2% vs 3.7% for [12] and 5.4% for [13]). This verifies the effectiveness of our method for extracting object from image with bounding box prior. For fair comparison, we also report the results of method [6] that use graph-cut based procedure to increase the final segmentation performance (3.3%) by integrating the shape prior generated by CPMC proposals. Our method does not use this procedure and achieve competitive results with [6]. This demonstrates that our double sparse reconstruction model has the capability of directly predicting the reasonable object confidence map for object extracting without graph-cut based procedure which is commonly used by many interactive segmentation methods.

Some qualitative segmentation results are displayed in Fig. 7. The extracted objects by our method are highlighted by color mask. Our method successfully predicts the masks of objects that are in complex background. For example, for the banana in the clustered flowers, we can detect it entirely with well-preserved boundaries. It demonstrates that our framework performs well for extracting objects from natural images.

Comparison with Seed based Methods: As the procedure of region merging servers as a seed-based region merging strategy, in this part we additionally compare the proposed method to the state-of-the-art seed based segmentation approaches, including Laplacian Coordinates (LC) approach [24], Graph Cuts (GC), Power Watershed (PWS) [21], Maximum Spanning Forest with Kruskal’s (MSFK) and Prim’s (MSFP) algorithm [22] and Random Walker (RW) [23]. As these do not use the bounding box prior and for fair comparison, we directly use the published results.
Table 1. Error-rate of bounding box prior based algorithms GrabCut, GrabCut-Pinpoint, Adaptive Kernel Segmentation, F-G Classification, Xia, and our method on the Grabcut-50 dataset.

<table>
<thead>
<tr>
<th>Method</th>
<th>Error-rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>GrabCut [16]</td>
<td>8.1 %</td>
</tr>
<tr>
<td>Kernel Segmentation [36]</td>
<td>9.7 %</td>
</tr>
<tr>
<td>Adaptive Kernel Segmentation [36]</td>
<td>7.1 %</td>
</tr>
<tr>
<td>GrabCut-Pinpoint [12]</td>
<td>3.7 %</td>
</tr>
<tr>
<td>F-G Classification [13]</td>
<td>5.4 %</td>
</tr>
<tr>
<td>Xia [6]</td>
<td>3.3 %</td>
</tr>
<tr>
<td>Ours</td>
<td>3.2 %</td>
</tr>
</tbody>
</table>

Table 2. PRI, GCE and Vol of seed based prior based algorithms Graph-cut, MSFK, MSFP, PWS, RW, LC and our method on the Grabcut-50 dataset.

<table>
<thead>
<tr>
<th>Method</th>
<th>PRI</th>
<th>GCE</th>
<th>Vol</th>
</tr>
</thead>
<tbody>
<tr>
<td>GC</td>
<td>0.9714</td>
<td>0.0268</td>
<td>0.1877</td>
</tr>
<tr>
<td>MSFK [22]</td>
<td>0.9690</td>
<td>0.0293</td>
<td>0.2013</td>
</tr>
<tr>
<td>MSFP</td>
<td>0.9689</td>
<td>0.0278</td>
<td>0.1931</td>
</tr>
<tr>
<td>PWS [21]</td>
<td>0.9700</td>
<td>0.0280</td>
<td>0.1934</td>
</tr>
<tr>
<td>RW [23]</td>
<td>0.9715</td>
<td>0.0262</td>
<td>0.1836</td>
</tr>
<tr>
<td>LC [24]</td>
<td>0.9765</td>
<td>0.0262</td>
<td>0.1654</td>
</tr>
<tr>
<td>Ours</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3. IoU of bounding box prior based algorithms GraphCut, SegmentsSum, Xia [6], and our method on the PASCAL VOC 2011 validation set.

<table>
<thead>
<tr>
<th>Method</th>
<th>IoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>GraphCut</td>
<td>63.1 %</td>
</tr>
<tr>
<td>SegmentsSum</td>
<td>56.7 %</td>
</tr>
<tr>
<td>Xia [6]</td>
<td>72.6 %</td>
</tr>
<tr>
<td>Ours</td>
<td>73.2 %</td>
</tr>
</tbody>
</table>

Table 4. Jaccard similarity (%) of our method vs oracle scores of object proposals methods on VOC 2012 validation dataset [32].

<table>
<thead>
<tr>
<th>Method</th>
<th>0</th>
<th>1</th>
<th>c</th>
<th>avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>GraphCut</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>SegmentsSum</td>
<td>1100</td>
<td>1100</td>
<td>1100</td>
<td>1100</td>
</tr>
<tr>
<td>Xia [6]</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Ours</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

4.2. PASCAL VOC

In order to more thoroughly evaluate segmentation performance, we have experimented with our method on the PASCAL VOC object segmentation dataset [32]. We use images from the validation dataset to evaluate the method performance, where the bounding box for each object is provided. We use Intersection over Union (IoU) [32] to evaluate the performance of comparing methods. The weight $\alpha$ in Eq. (5) is set over the interval $[0.8, 2]$ for all experiments on VOC validation dataset. We vary the parameter $\alpha$ with a step size of 0.1. Each specific category shares a fixed $\alpha$.

**Compared with State-of-the-art Methods:** A series of experiments have been conducted on the VOC 2011 validation database which contains 1,112 images. For these images, the bounding boxes are provided by ground truth. To demonstrate the effectiveness of the proposed method for utilizing the segment hypotheses to extract objects, we first compare our method with the result that directly merges segments generated by CPMC [2], named as SegmentsSum. The results are reported in Table 3. SegmentsSum achieves the Jaccard score of 56.7% on this database. Our method gives a huge boost in segmentation accuracy. It obtains 73.2% by leveraging the segment hypotheses based on a sparse reconstruction framework. For fair comparison, we further compare our method against the state-of-the-art methods GraphCut and [6]. With the bounding box prior, the method of GraphCut and [6] achieve 63.1% and 72.6% of average Jaccard score on this database, respectively. Our method also outperforms these two baselines GraphCut and [6] which integrates object shape guidance generated by CPMC method into their graph-cut-based optimization. This further demonstrates that the image objects can be effectively reconstructed by our proposed double sparse reconstruction method. And our methodology is more powerful than the segment hypothesis integrated approaches for extracting objects.

Some qualitative extraction results of our method are visualized in Fig. 8. Many objects from different categories are included which are often with intrinsic inhomogeneity. The extracted objects of our method are highlighted by color mask. It is visually clear that our method can produce satisfactory results for extracting the objects of large appearance or pose variations in natural images. However, in some cases, our method will fail, as shown in Fig. 9.
ods: As the segment hypotheses are generated by the object proposals method [2] in our method, in this part, we will compare our proposed method to the state-of-the-art object proposals methods to demonstrate the effectiveness of our strategy for object extraction. Note that, to measure the quality of proposals, candidates generation methods report an “oracle” score that selected best candidate for each object (also Best Spatial Support score (BSS) [35]) among the pool with respect to the number of candidates. For example, when the number of object candidates is 100, each one is evaluated with the ground truth and then reports the best score among them. Obviously, the Jaccard score of our method is not an oracle score.

Table 4 reports the detailed comparison of Jaccard similarity of our method with the oracle scores of object proposal generation methods on VOC 2012 validation dataset. These methods include [2][3][18][11][25][27]. On this database, when the total number of candidates of these methods is 100, CPMC [2] achieves 59% of oracle score and the state-of-the-art object proposal method MCG [18] achieves 63.7% of oracle score. The proposed methodology achieves 73.3% of the mean Jaccard score (which is not an oracle score) on this dataset, clearly outperforming the state-of-the-arts. This verifies the effectiveness of our algorithm for obtaining accurate extraction results: the performance is even comparable with the oracle scores of the state-of-the-art object proposals methods.

Note that in this work, we do not mean to claim that our method is always superior over MCG method. It is predicted that in [18], MCG can achieve better results with generating more object candidates. For example, when they generate about 1000 object candidates for each image then report the best overlap ones, the oracle score increases to 76.0%. It is reasonable that they can achieve better results when the number of object candidates among the pool is getting larger. However it makes more difficult to pick out the best proposal.

5. Conclusion

This paper presents a double sparse reconstruction method to extract objects from images with the bounding box prior. Object regions can be well reconstructed by our model since they are with small reconstruction errors to segment hypotheses bases, simultaneously, large reconstruction errors to background hypotheses bases. Region merging procedure is finally used to make the reconstructed object with a well-preserved boundary. The proposed object extraction method is examined on two popular segmentation databases: PASCAL VOC object segmentation dataset and GrabCut-50 database, and experimental results indicate that (i) the proposed method is more robust than state-of-the-art semi-supervised methods for object extraction, and (ii) the proposed double sparse reconstruction scheme is more powerful than the segments integrated approaches for characterizing the correlation information between regions. Our future work will focus on how to obtain bounding box for natural image with more robustness.

6. Acknowledgments

This work was partially supported by the National Science Fund for Distinguished Young Scholars under Grant Nos. 61125305, 91420201, 61472187, 61233011 and 61373063, the National Natural Science Fund of China(Grant Nos. 61103058), the Key Project of Chinese Ministry of Education under Grant No. 313030, the 973 Program No. 2014CB349303, Fundamental Research Funds for the Central Universities No. 3092014012005, and Program for Changjiang Scholars and Innovative Research Team in University No. IRT13072.
References


[33] Emmanuel Candes, J. Romberg L1-magic: Recovery of sparse signals via convex programming. 2005

