Segmentation over Detection by Coupled Global and Local Sparse Representations

Wei Xia, Zheng Song, Jiashi Feng, Loong-Fah Cheong, and Shuicheng Yan

Department of ECE, National University of Singapore

Abstract. Motivated by the rising performances of object detection algorithms, we investigate how to further precisely segment out objects within the output bounding boxes. The task is formulated as a unified optimization problem, pursuing a unique latent object mask in non-parametric manner. For a given test image, the objects are first detected by detectors. Then for each detected bounding box, the objects of the same category along with their object masks are extracted from the training set. The latent mask of the object within the bounding box is inferred based on three objectives: 1) the latent mask should be coherent, subject to sparse errors caused by within-category diversities, with the global bounding-box-level mask inferred by sparse representation over the bounding boxes of the same category within the training set; 2) the latent mask should be coherent with local patch-level mask inferred by sparse representation of the individual patch over all spatially nearby (handling local deformations) patches of the same category in the training set; and 3) mask property within each sufficiently small super-pixel should be consistent. All these three objectives are integrated into a unified optimization problem, and finally the sparse representation coefficients and the latent mask are alternately optimized based on Lasso optimization and smooth approximation followed by Accelerated Proximal Gradient method, respectively. Extensive experiments on the Pascal VOC object segmentation datasets, VOC2007 and VOC2010, show that our proposed algorithm achieves competitive results with the state-of-the-art learning based algorithms, and is superior over other detection based object segmentation algorithms.

1 Introduction

Localizing and recognizing semantic objects efficiently and comprehensively in a complex visual world is one of the amazing capabilities of human visual and cognitive system. In recent years, many achievements have been witnessed in object detection and segmentation [1],[2]. While purely bottom-up segmentation based on local pixel and patch appearance is not well-posed for the object segmentation problem, integrating object detector as guidance priors has been the latest trend [3],[4].

The object detector can localize the coarse position of a certain object by a bounding box [5],[6], yet lacks the accuracy to precisely identify the object at pixel level as required in semantic object segmentation task [7],[1],[8],[2],[9],[10].

A. Fitzgibbon et al. (Eds.): ECCV 2012, Part V, LNCS 7576, pp. 662–675, 2012.

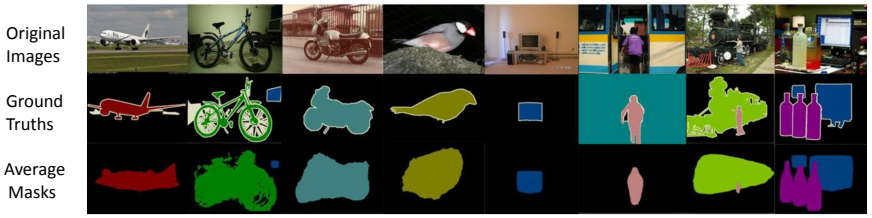

Fig. 1. Exemplar images, ground truth and corresponding average masks

An intuitive solution is to learn a coarse average mask for each object category based on the segmentation ground truths as well as the object detections on training set, and then propagate the learnt mask to each detected bounding box on the test set [11]. Some of the mask examples are shown in Figure 1.

Since the average masks learnt from detection results are far from representing the accurate position of the semantic class at pixel level, especially for those categories having large within-class variance, to achieve semantic segmentation, some compensatory information from the image itself should be integrated. Therefore we propose a unified framework of coupled global and local sparse representations to refine towards a unique latent mask for each detected bounding box in test images based on the coarse mask from object detection and the finer image details.

As in Figure 2, the proposed framework pursues the optimal latent mask as well as the optimal reconstructions for the sparse reconstruction constraints. The coarse masks from the object detectors are input to the framework as initializations. It is worth noting that the proposed framework is convex, not relying on initializations and thus global optimization can be obtained.

We introduce three objectives in our framework. Firstly, a test image and its latent mask can be sparsely reconstructed using a set of training images and their corresponding ground truth segmentation masks. Since the test and training images are the foreground objects cropped and normalized from the object detectors and encoded with visual words, such discriminative model using global and local BOW features have verified that sparse coding is better than KNN or other reconstruction methods in finding the related samples of a test sample [12].

Beyond global reconstruction constraint, local reconstruction is also applied using local image features and local masks, which are extracted on regularly partitioned patches on training and testing images as well as masks. A certain degree of spatial flexibility in the localized reconstruction is introduced using neighborhood patches. These two constraints aim to find similar samples and patches in the training set as well as a better reconstructed latent mask across different within-category instances.

Since sparse reconstruction may bring artifacts to the learnt mask, we additionally introduce a smooth objective. We first perform super-pixel segmentation [13] on the test image within the predicted bounding box area. Then we enforce

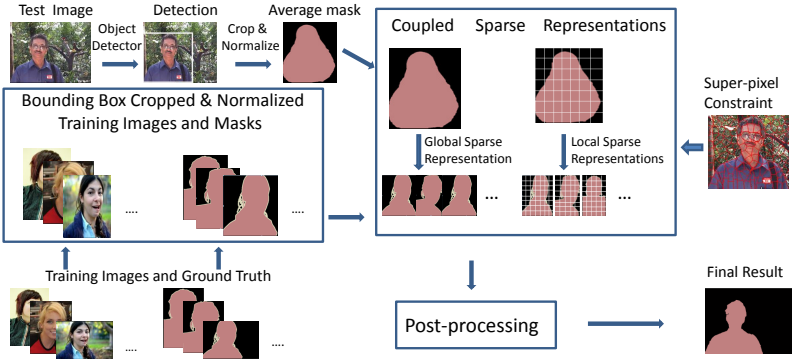

Fig. 2. Illustration on the proposed framework. Given a test image, a bounding box is predicted, then the mask is cropped and normalized within the bounding box. Based on the training images and the corresponding ground truth segmentations from the same category, the initial mask for the test image is refined through a coupled sparse representation framework. For the global part, the global representation feature for each cropped image are extracted and are sparsely reconstructed with the global masks, while for the local part, the cropped image and mask are first regularly partitioned and then reconstructed locally. Finally, the learnt mask is further refined through post-processing to obtain the final result.

the latent mask elements corresponding to the same super-pixel to have the same value, which will make the mask boundary more accurate.

2 Related Work

Image and semantic object segmentation are very important branches of research in computer vision [7],[1],[2]. Traditional image segmentation approaches mainly follow a bottom-up manner based on low-level image features such as color, texture, and shape [14]. However, different with general image segmentation problems, object segmentation need explore more information on object shape and appearance structures across the whole object category. Hence many recent works integrating top-down semantic information have been proposed [1],[10].

One popular line of research to combine top-down cues is based on Conditional Random Field (CRF) model, [9],[8],[1]. Ladicky *et al.* [9] proposed a hierarchical CRF of multiple scales, combining multi-layer image classification and contrast sensitive pairwise smooth potentials, to assign a fixed number of category labels to pixels. Ladicky *et al.* [8] introduced global co-occurrence statistics as another top-down cue in the CRF model. A drawback of such pixel or super-pixel based methods is merging many neighboring local patches from the same object without modeling the object globally. Recently through adding object detection bounding box constraints to the global potential functions [8] or to the local unary and pairwise potentials [1], such limits are partially handled. In addition, Li *et al.* [2] proposed an approach pipeline to categorize between

multiple figure-ground hypotheses with large object spatial support, generated by bottom-up computational processes. Both [2] and [1] achieved the state-of-the-art performances on PASCAL VOC segmentation benchmark [15].

Another branch of framework was proposed in [11],[4] and [3], which applied the idea of refining the masks from detectors built upon the state-of-the-art part-based object detector by Felzenszwalb *et al.* [5]. Felzenszwalb *et al.* [11] proposed the intuitive baseline to predict a coarse mask from detection. Yang *et al.* [4] refined the prediction of the detector using color and depth order of objects. Brox *et al.* [3] applied part-based poselet detector, which can predict masks for numerous parts of an object, then aligned the poselet to the object contours and aggregated them into a variational smoothing object, and finally refined the segmentation based on self-similarity defined on small image patches. This approach has achieved comparable performance on PASCAL VOC with the CRF model, however the heavy manual labeling burden to annotate the poselet samples poses a great limitation.

Our framework follows the detection-based framework, but different in its non-parametric learning philosophy based on sparse representations. Non-parametric or sample-based methods are proved as efficient as parametric models for object detection by Malisiewicz et al. [16]. Moreover, the sample-based matching tends to be more semantically meaningful in obtaining the unique object segmentations.

Sparse representation has seen significant impact in computer vision, several variants of L1 minimization technique have been applied to face recognition [17], image classification and segmentation [18],[19], motion and data segmentation [20],[21], and so on.

In the next section, we introduce how to obtain the average masks based on the state-of-the-art detectors and illustrate the details on how to formulate the mask refinement as reconstructing image features and object masks, from global and local view, respectively. Section 4 provides the optimization procedure to solve the objective function. Section 5 shows some implementing details, while Section 6 demonstrates the experimental results and the comparison analysis with the state-of-the-art algorithms. Finally, we conclude the paper and propose some discussion about possible future work in Section 7.

3 Problem Formulation

3.1 Object Masks from Detectors

In [22] [23], a latent hierarchical structural learning method for object detection was presented, in which an object is represented by a mixture of hierarchical tree models where the nodes represent object parts. The nodes can move spatially allowing for some deformation. After learning the hierarchical model through a latent SVM, object detection can be performed by scanning the sub-windows using dynamic programming.

Based on the object detector above and the detection results of the training samples with segmentation ground truths, we can learn a binary mask for each

detection model and associate the mask with the detection bounding boxes to generate segmentations for test images. Some exemplar object masks are shown in Figure 1. The predicted segmentation accuracy of the average mask depends on both the detection precision and within-category varieties. For regular and convex shape objects like TV monitor and bottle, the mask can be quite semantically meaningful, while for categories with large pose variations like birds and persons, the unique contour information for each object instance is ignored and thus the segmentation accuracy declines significantly. Therefore in order to obtain accurate object segmentation, a better solution is desirable. A framework of coupled global and local sparse representations is introduced right for such purpose in the subsequent section.

3.2 Coupled Global and Local Sparse Representations

Given a test image and a set of training image from the same object category, in order to remove the effect of the cluttered background and better utilize the results of object detectors, the test image denoted as I_t is represented as the foreground objects cropped and normalized based on the object detector bounding boxes, while the N training images are foreground objects denoted as $\{I_b\}_{b=1}^N$ with segmentation mask $\{\mathbf{m}_b\}_{b=1}^N$ cropped and normalized in the areas extended by the ground truth segmentations. All subsequent procedures like feature extraction and over-segmentation are performed within the normalized bounding box areas. Since most of the test and the training bounding boxes have enough consistency, by enforcing the sparsity of the reconstruction coefficients, only a small subset of highly correlated training samples are selected, and the linear combination of them will greatly reduce the boundary ambiguity than the average mask learnt from all samples. Thus a global image[1] and segmentation mask reconstruction framework is proposed as,

$$\min_{\mathbf{y},\mathbf{m}} \frac{1}{2}\|\mathbf{x} - B\mathbf{y}\|_2^2 + \lambda_1\|\mathbf{y}\|_1 + \lambda_2\|\mathbf{m} - M\mathbf{y}\|_1, \tag{1}$$

where $\mathbf{y}$ is the sparse reconstruction coefficient vector for both image and segmentation mask reconstructions, $\mathbf{x}$ is a global representation vector for test image I_t, $B = [\mathbf{b}_1, \ldots, \mathbf{b}_N]$ is a basis matrix consisting of the global features of the N cropped and normalized training images from the same category, and $M = [\mathbf{m}_1, \ldots, \mathbf{m}_N]$ is a mask matrix whose columns correspond to the N cropped and normalized segmentation masks of the training images. In order to ensure the reconstructed mask to be generally accurate except for few pixels caused by possible locally spatial deformations, the mask reconstruction error is enforced to be sparse.

To reduce the artifacts brought from the sparse reconstruction, an additional smooth objective that enforces the consistency within each small super-pixels

[1] Note that here after, the image reconstruction means the reconstruction of the normalized bounding box within the test image, the same for the training images and masks.

is integrated. We first over-segment I_t into several visually coherent super-pixels as in [13]. Assuming that all the pixels within a super-pixel share the same mask label, the global segmentation mask $\mathbf{m}$ admits following specific form,$\mathbf{m} = [\mathbf{s}_1, \ldots, \mathbf{s}_p]\,[v_1, \ldots, v_p]^T$, where $\mathbf{s}_i$ is the binary value mask for the i^{th} superpixel, yet the latent mask is relaxed from binary values to continuous real values between 0 and 1 to make the problem tractable. $\mathbf{v} = [v_1, \ldots, v_p]^T$ is the selection vector, and the value of 1 means the super-pixel is selected as the corresponding foreground object.

To further handle the local deformation, beyond such global reconstruction consistency, the local patch reconstruction is also pursued as follows,

$$\min_{\{\mathbf{y}_k, \mathbf{m}_k\}} \sum_k \frac{1}{2}|\mathbf{x}_k - B^k\mathbf{y}_k\|_2^2 + \gamma_1\|\mathbf{y}_k\|_1 + \frac{1}{2}\gamma_2\|\mathbf{m}_k - M^k\mathbf{y}_k\|_2^2. \tag{2}$$

The test and training images I_t and I_b are regularly partitioned into several patches, indexed by k, $\mathbf{x}_k$ and $\mathbf{m}_k$ are the feature vector and patch mask for the k^{th} patch of the test image, respectively. B^k is a patch feature matrix where each column is a feature vector corresponding to a patch in the training images. In order to handle within-category varieties, some local spatial flexibility is allowed to some extent, and $B^k = [\mathbf{B}_1^{\mathcal{N}_k}, \ldots, \mathbf{B}_N^{\mathcal{N}_k}]$, where N is the number of the cropped and normalized training images from the same category. $\mathcal{N}_k$ means the neighborhood of the k^{th} patch, e.g., four nearest neighbors. M^k consists of segmentation masks for the patches in $\mathcal{N}_k$ across all the training images from the same category. By combining the Eqn. 1 and 2, we obtain the following unified objective function,

$$\begin{aligned}\min_{\mathbf{y},\mathbf{m},\{\mathbf{y}_k\}} \quad & \frac{1}{2}\|\mathbf{x} - B\mathbf{y}\|_2^2 + \lambda_1\|\mathbf{y}\|_1 + \lambda_2\|\mathbf{m} - M\mathbf{y}\|_1 \\ & + \beta\left[\sum_k \frac{1}{2}\|\mathbf{x}_k - B^k\mathbf{y}_k\|_2^2 + \gamma_1\|\mathbf{y}_k\|_1 + \frac{1}{2}\gamma_2\|\mathbf{m}_k - M^k\mathbf{y}_k\|_2^2\right]. \end{aligned} \tag{3}$$

Note that the above framework is convex, thus global optimization can be obtained. Therefore, as long as the object area are accurately detected, by alternatively optimizing the proposed framework, the object mask could be efficiently located regardless of the initial predicted mask.

4 Optimization Procedure

Although the above objective function in Eqn. (3) is convex, it is not smooth and thus very difficult to obtain a closed-form solution. Instead, we propose to alternatively optimize w.r.t reconstruction coefficients $\{\mathbf{y}, \mathbf{y}_k\}$ and optimize w.r.t the latent mask $\mathbf{m}$. The optimization procedure is iterated between optimizing $\mathbf{y}, \mathbf{y}_k$ while fixing $\mathbf{m}$, and optimizing $\mathbf{m}$ while fixing $\mathbf{y}, \mathbf{y}_k$.

4.1 Optimization w.r.t. $\{\mathbf{y}, \mathbf{y}_k\}$

Firstly, the objective function can be simplified as follows when fixing $\mathbf{m}$,

$$\min_{\mathbf{y},\{\mathbf{y}_k\}} \frac{1}{2}\|\mathbf{x} - B\mathbf{y}\|_2^2 + \lambda_1\|\mathbf{y}\|_1 + \lambda_2\|\mathbf{m} - M\mathbf{y}\|_1 + \beta\left[\sum_k \frac{1}{2}\|\tilde{\mathbf{x}}_k - \tilde{B}^k\mathbf{y}_k\|_2^2 + \gamma_1\|\mathbf{y}_k\|_1\right], \tag{4}$$

where $\tilde{\mathbf{x}}_k = \left[\mathbf{x}_k^T, \sqrt{\gamma_2}\mathbf{m}_k^T\right]^T$ and $\tilde{B}_k = \left[{B^k}^T, \sqrt{\gamma_2}{M^k}^T\right]^T$. Since $\mathbf{y}$ and $\mathbf{y}_k$ are independent, the problem is decomposed into the following two subproblems.

Subproblem 1: y. The first subproblem is as follows,

$$\min_{\mathbf{y}} \frac{1}{2}\|\mathbf{x} - B\mathbf{y}\|_2^2 + \lambda_1\|\mathbf{y}\|_1 + \lambda_2\|\mathbf{m} - M\mathbf{y}\|_1. \tag{5}$$

Let $\mathbf{z} = \frac{\lambda_2}{\lambda_1}(\mathbf{m} - M\mathbf{y})$, namely $[\lambda_1 I, \lambda_2 M][\mathbf{z}; \mathbf{y}] = \lambda_2\mathbf{m}$. Let $\mathbf{u} = [\mathbf{z}; \mathbf{y}]$, then the objective function in 5 can be reformulated as,

$$\min_{\mathbf{u}} \frac{1}{2}\|\mathbf{x} - [\mathbf{0}, B]\mathbf{u}\|_2^2 + \lambda_1\|\mathbf{u}\|_1 + \rho\|[\lambda_1 I, \lambda_2 M]\mathbf{u} - \lambda_2\mathbf{m}\|_2^2. \tag{6}$$

Through some algebraic manipulation, the objective function can be reformulated as

$$\min_{\mathbf{u}} \frac{1}{2}\|\tilde{\mathbf{v}} - \tilde{D}\mathbf{u}\|_2^2 + \lambda_1\|\mathbf{u}\|_1, \tag{7}$$

where $\tilde{\mathbf{v}} = [\mathbf{x}^T, \sqrt{\rho}\lambda_2\mathbf{m}^T]^T$, and $\tilde{D} = \begin{bmatrix} 0 & B \\ \sqrt{\rho}\lambda_1 I & \sqrt{\rho}\lambda_2 M \end{bmatrix}$. The subproblem is a standard sparse regularized optimization problem, which can be solved through many software packages, like L1-magic toolbox [24].

Subproblem 2: $\{\mathbf{y}_k\}$. The second subproblem is a also a standard sparse regularized optimization problem, which can be easily solved as Eqn. 7.

$$\min_{\{\mathbf{y}_k\}} \sum_k \frac{1}{2}\|\tilde{\mathbf{x}}_k - \tilde{B}^k\mathbf{y}_k\|_2^2 + \gamma_1\|\mathbf{y}_k\|_1. \tag{8}$$

4.2 Optimization w.r.t. m

By fixing $\mathbf{y}, \mathbf{y}_k$, the objective function w.r.t. $\mathbf{m}$ is

$$\min_{\mathbf{m}} \frac{1}{2}\sum_k \|\mathbf{m}_k - M^k\mathbf{y}_k\|_2^2 + \frac{\lambda_2}{\gamma_2\beta}\|\mathbf{m} - M\mathbf{y}\|_1. \tag{9}$$

Since $\mathbf{m} = [\mathbf{s}_1, \ldots, \mathbf{s}_p]\,\mathbf{v} = S\mathbf{v}$ and $\mathbf{m}_k = \left[\mathbf{s}_1^k, \ldots, \mathbf{s}_p^k\right]\mathbf{v} = S^k\mathbf{v}$, where S^k is the matrix corresponding to the superpixel mask on the k^{th} patch, the above objective function is equivalent to,

$$\min_{\mathbf{v}} \frac{1}{2}\sum_k \|S^k\mathbf{v} - M^k\mathbf{y}_k\|_2^2 + \eta\|S\mathbf{v} - M\mathbf{y}\|_1, \tag{10}$$

where $\eta = \frac{\lambda_2}{\gamma_2\beta}$. Let $\tilde{S} = [S^1, S^2, \ldots, S^K]^T$ and $\tilde{\mathbf{y}} = [M^1\mathbf{y}_1, M^2\mathbf{y}_2, \ldots, M^K\mathbf{y}_K]^T$, then the objective function can be simplified as $\min_{\mathbf{v}} \frac{1}{2}\|\tilde{S}\mathbf{v} - \tilde{\mathbf{y}}\|_2^2 + \eta\|S\mathbf{v} - M\mathbf{y}\|_1$.

Since the above sparse regularization optimization problem is non-smooth, according to [25], the subgradient methods to solve non-smooth convex problem have efficiency estimate of the order $O(\frac{1}{\xi^2})$, where ξ is the desired absolute accuracy of the approximate solution. Thus to improve the rate of convergence, the non-smooth term of $\eta\|S\mathbf{v} - M\mathbf{y}\|_1$ can be approximated by following smooth function,

$$\widehat{f}_\mu(\mathbf{v}) = \max_{\|\mathbf{w}\|_\infty \le 1} \langle S\mathbf{v} - M\mathbf{y}, \mathbf{w}\rangle - \frac{\mu}{2}\|\mathbf{w}\|_2^2, \tag{11}$$

where μ is a parameter to control the approximation accuracy.

For a fixed $\mathbf{v}$, let $\mathbf{w}(\mathbf{v})$ denote the unique maximizer of Eqn 11. Then $\mathbf{w}(\mathbf{v}) = \min\{1, \max\{-1, (S\mathbf{v} - M\mathbf{y})/\mu\}\}$, where operators $\min\{\cdot,\cdot\}$ and $\max\{\cdot,\cdot\}$ are performed in element-wise for the involved vector. Moreover, the smooth approximation in 11 is differentiable and its gradient $S\mathbf{w}(\mathbf{v})$ is Lipschitz continuous with the constant $1/\mu\|S\|_2$.

The entire gradient for the smooth approximation function $f_\mu(\mathbf{v}) = \frac{1}{2}\|\tilde{S}\mathbf{v} - \tilde{\mathbf{y}}\|_2^2 + \eta\widehat{f}_\mu(\mathbf{v})$ is:$\nabla f_\mu(\mathbf{v}) = \tilde{S}^T\left(\tilde{S}\mathbf{v} - \tilde{\mathbf{y}}\right) + \eta S\mathbf{w}$, and its Lipschitz constant is $L_{f_\mu} = \|\tilde{S}^T\tilde{S}\|_2 + \eta/\mu\|S\|_2$, where $\|\cdot\|_2$ denotes the spectral norm for a matrix.

The detailed smooth minimization procedure is shown in Algorithm 1 and the entire optimization procedure is also demonstrated in Algorithm 2.

Algorithm 1. Smooth Minimization Procedure.

Input $\tilde{S}$, $\tilde{\mathbf{y}}$, η. **Output** $\mathbf{v}$.
Initialization Calculate L_{f_μ}. Initialize $\mathbf{v}^{(0)}$, $\boldsymbol{\gamma}^{(0)}$, and let $\theta^{(0)} \leftarrow 0, t \leftarrow 0$.
Repeat until convergence
$\boldsymbol{\alpha}^{(t)} = (1-\theta^{(t)})\mathbf{v}^{(t)} + \theta^{(t)}\boldsymbol{\gamma}^{(t)}$, Calculate $\nabla f_\mu(\alpha^{(t)})$,
$\boldsymbol{\gamma}^{(t+1)} = \boldsymbol{\gamma}^{(t)} - \frac{1}{\theta^{(t)}L_{f_\mu}}\nabla f_\mu(\alpha^{(t)})$, $\mathbf{v}^{(t+1)} = (1-\theta^{(t)})\mathbf{v}^{(t)} + \theta^{(t)}\boldsymbol{\gamma}^{(t+1)}$,
$\theta^{(t+1)} = \frac{2}{t+1}, t \leftarrow t+1$.

5 Implementation Details

In this section, we introduce some implementation details. For the detected bounding boxes of each category, over-segmentation and feature extraction are performed within the bounding box area. In the over-segmentation step, the

Algorithm 2. Full Alternative Optimization Procedure.

Input $\mathbf{x}$, $\mathbf{m}$ **Output** $\mathbf{y}$, $\mathbf{m}$
Initialization Initialize λ_1, λ_2, γ_1, γ_2, β and let $t \leftarrow 0$.
Repeat until convergence
Fix $\mathbf{m}^{(t)}$, optimize Eqn. 4 , **update** $\mathbf{y}^{(t+1)} \leftarrow \mathbf{y}^{(t)}, \mathbf{y_k}^{(t+1)} \leftarrow \mathbf{y_k}^{(t)}$
Fix $\mathbf{y}^{(t+1)}, \mathbf{y_k}^{(t+1)}$, optimize Eqn. 9, **update** $\mathbf{m}^{(t+1)} \leftarrow \mathbf{m}^{(t)}, t \leftarrow t+1$

number of the superpixels is set to be 150. For the global image representation, according to [26], LLC (Locality-constrained Linear Coding)[26] is applied to extract the Spatial Pyramid Representation (SPM) [27] vectors within the bounding box area. For the local patch representation, the patch size are set to be 16 by 16 pixels. Dense SIFT [28] features are extracted regularly at every other pixel. Then LLC is also applied to code the SIFT features into a Bag-of-Word patch representation.

The parameters $\lambda_1, \lambda_2, \gamma_1, \gamma_2, \beta, \mu$ in the optimization framework are learned empirically with the validation set. On a PC with dual-core of 2.99 GHz Intel CPU and 8GB memory, for a given test image, it takes on average around 1 minute to do the over-segmentation and about 3 - 10 minutes for the alternative optimization framework to converge using the current un-optimized Matlab code, further speedup could be expected if applying C++ implementation.

Post-processing. Since the latent mask learnt from the coupled sparse representations framework is defined at super-pixel level, some post-processings are proposed to further refine the mask to reduce the errors introduced by super-pixels that are not well aligned to the object contours. For a given cropped and normalized test bounding box, we first perform over-segmentation at different scales, with super-pixel number set as 150, 200, and 250, respectively, then re-run the coupled sparse representation framework three times, and finally calculate the ultimate mask as the intersection of the three outputs.

6 Experiments

The experiments are divided into two parts. The first part demonstrates the proof-of-concept studies, where several important aspects of the algorithm are evaluated. In the second part, we show the results on the benchmark databases with the comparison to other state-of–the-art algorithms.

6.1 Proof-of-Concept Experiments

We first evaluate the effects of different algorithm parts on the VOC2010 training vs. validation sets. Table 1 shows the main results. It can be observed that each part of the model improves the average performance, from the baseline performance of 31.7% to the final full model performance of 36.4%. The baseline model is calculated directly from the average masks predicted from the object

Table 1. Study of the effects of different algorithm parts of the proposed method on the VOC2010 training vs. validation sets. The BA is the average masks predicted from the object detectors. +GL only applies the global sparse representation, +LO applies the coupled global and local sparse representation framework, while the Full model adds the extra post-processing to the coupled sparse representation framework.

Mthd	bg	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	average
BA	78.5	35.6	14.6	25.8	22.6	39.1	54.3	46.1	24.8	5.2	33.5	9.8	24.5	29.2	42.1	33.7	17.6	30.8	19.2	39.1	40.1	31.7
+GL	80.1	39.3	19.7	31.9	27.1	41.0	58.5	47.8	28.4	5.5	37.1	12.7	28.1	32.9	46.2	37.4	20.3	34.5	23.5	40.7	40.8	34.9
+LO	80.5	41.9	20.9	33.8	28.5	41.2	58.7	49.1	29.2	6.8	38.7	13.9	29.0	34.4	48.7	39.2	22.1	36.7	23.8	41.2	40.5	36.1
Full	80.6	42.6	20.7	34.2	28.9	41.7	58.6	49.1	29.8	7.6	39.1	14.4	29.4	34.6	48.9	39.5	21.9	36.8	23.7	41.4	40.5	36.4

detectors. The best improvement is seen from the baseline to the global only reconstruction, which makes sense because the sparse reconstruction framework selects the most similar training masks and significantly refines the initial test mask in the bounding box. In comparison, local reconstruction gets a relatively lower improvement since it is introduced to handle local spatial deformations, which only covers a small portion of the entire segmented masks.

6.2 Performance Comparison

We evaluate the proposed framework on two challenging datasets for object class segmentation: PASCAL VOC2007 [11] and VOC2010 [15] segmentation challenge. VOC2007 contains 632 images in total, with 20 foreground (objects) classes and 1 background class, while VOC2010 extends to 2892 images. Quantitative analysis of VOC results is based on intersection vs. union measure [2]. In VOC2007, we compare the performance gain of our mask refinement technique with the state-of-the-art algorithm of Thomas Brox's poselet alignment [3], which also applies a segmentation from detection framework. In VOC2010, we mainly compare with all the state-of-the-art methods ever published (submitted to the VOC2010 challenge or already published papers).

Performance Gain from Mask Refinement. In this subsection, we mainly compare the performance gain obtained through the mask refine optimization with Brox's algorithm [3], which is based on poselet alignment and smoothing from detection. To make it fair, we adopt the same experiment setting as proposed in [3], which is evaluated on the combined Pascal VOC 2007 training, validation and test set (632 images). In [3], the poselet classifiers are trained on the *training* + *validation* set of the whole challenge excluding images from VOC2007 segmentation challenge, which is the same training set for our object detectors and coupled sparse representation framework. The detailed comparison are shown in Table 2. From the table, we can observe that although our baseline performance is a little bit higher than Brox's, the improvement from the baseline to the full model is higher. Furthermore, compared with Brox's algorithm,

[2] Defined as $accuracy = \frac{TruePositives}{TruePositive+FalseNegative+FalsePositive}$.

Table 2. Accuracy comparison at the same database setting in [3] for VOC2007. The first two columns are performances before and after poselet alignment and smoothing from Brox's algorithm [3]. Our baseline is the results directly from the predicted average masks, and our full model is the result after the mask refinement. The numbers in the brackets are the performance gain from the baseline model to the full model. Our model achieves higher performance gain in 13 classes among the 21 classes (including background), and a higher overall gain.

Method	Brox baseline	Brox full model	Our baseline	Full model
background	78.58	79.23 (**+0.65**)	79.42	80.05 (+0.63)
aeroplane	26.63	36.26 (**+9.63**)	34.87	40.37 (+5.50)
bicycle	32.14	38.54 (**+6.40**)	16.72	18.69 (+1.97)
bird	12.70	16.57 (+3.87)	27.11	31.23 (**+4.12**)
boat	12.74	12.14 (-0.60)	18.77	26.67 (**+7.90**)
bottle	31.40	30.40 (-1.00)	34.11	38.68 (**+4.57**)
bus	29.24	33.20 (+3.96)	45.63	56.26 (**+10.63**)
car	39.25	42.15 (+2.9)	39.01	47.21 (**+8.20**)
cat	38.19	44.99 (**+6.8**)	22.69	27.54 (+4.85)
chair	7.89	10.33 (**+2.44**)	5.11	7.02 (+1.91)
cow	29.24	37.21 (**+7.97**)	32.41	36.58 (+4.17)
diningtable	11.37	10.69 (-0.68)	9.45	11.22 (**+1.77**)
dog	17.61	23.15 (+5.44)	19.74	27.31 (**+7.57**)
horse	35.41	43.92 (**+8.51**)	27.66	30.58 (+2.92)
motorbike	27.90	32.59 (+4.69)	36.21	45.63 (**+9.42**)
person	44.00	49.65 (+5.65)	31.18	37.21 (**+6.03**)
pottedplant	17.07	17.60 (+0.43)	15.68	19.32 (**+3.64**)
sheep	26.68	37.38 (**+10.70**)	27.41	33.79 (+6.38)
sofa	9.72	9.49 (-0.23)	14.24	21.58 (**+7.34**)
train	20.34	23.55(+3.21)	33.21	38.24 (**+5.03**)
tvmonitor	43.51	47.50 (+3.99)	31.15	37.59 (**+6.44**)
average	28.17	32.21 (+4.04)	28.66	33.94 (**+5.28**)

Table 3. Accuracy comparison of our method on VOC2010 test set with other well performing methods [15]

	Other	Learning Based				Detection Based			
Method	Stanford	UC3M	Bonn SVR	CVC HCRF	Brooks	UOCTTI	Brox	Ours	Ranking
background	80.0	73.4	**84.2**	81.1	70.1	80.0	82.2	81.7	3
aeroplane	38.8	45.9	52.5	**58.3**	31.0	36.7	43.8	46.2	3
bicycle	21.5	12.3	**27.4**	23.1	18.8	23.9	23.7	21.9	5
bird	13.6	14.5	32.3	**39.0**	19.5	20.9	30.4	36.9	2
boat	9.2	22.3	34.5	**37.8**	23.9	18.8	22.2	30.3	3
bottle	31.1	9.3	47.4	36.4	31.3	41.0	45.7	**47.9**	**1**
bus	51.8	46.8	60.6	**63.2**	53.5	62.7	56.0	62.8	2
car	44.4	38.3	54.8	**62.4**	45.3	49.0	51.9	50.2	4
cat	25.7	41.7	**42.6**	31.9	24.4	21.5	30.4	34.0	3
chair	6.7	0.0	9.0	9.1	8.2	8.3	9.2	**10.4**	**1**
cow	26.0	35.9	32.9	36.8	31.0	21.1	27.7	**40.5**	**1**
diningtable	12.5	20.7	**25.2**	24.6	16.4	7.0	6.9	15.9	5
dog	12.8	**34.1**	27.1	29.4	15.8	16.4	29.6	32.9	2
horse	31.0	34.8	32.4	37.5	27.3	28.2	42.8	**43.7**	**1**
motorbike	41.9	33.5	47.1	**60.6**	48.1	42.5	37.0	48.9	2
person	44.4	24.6	38.3	44.9	31.1	40.5	**47.1**	41.5	4
pottedplant	5.7	4.7	**36.5**	30.1	31.0	19.6	15.1	29.3	4
sheep	37.5	25.6	**50.3**	36.8	27.5	33.6	35.1	40.2	2
sofa	13.0	13.0	21.9	19.4	19.8	13.3	23.0	**26.6**	**1**
train	33.2	26.8	35.2	44.1	34.8	34.1	37.7	**44.2**	**1**
tvmonitor	32.3	26.1	40.9	35.9	26.4	36.5	**48.5**	46.8	2
average	29.1	27.8	39.7	**40.1**	30.3	31.8	34.9	39.7	2

which requires extra manual annotation of the poselet, our method requires no hand-labeling and thus is more practical and robust.

Comparison to State-of-the-Art Approaches. In order to compare with other state-of-the-art object segmentation methods, we run the full model on the test set of VOC2010 dataset. Table 3 shows our result with the top-ranking methods in the challenge. Our approach achieves a performance of 39.7%, that is comparable with the state-of-the-art approaches. Among the 20 object classes, our approach shows the best results on 6 categories (tied with the best performing algorithms) and has an average ranking of 2 - 3 for other categories. We also compare our approach with other detection based approach like UOCTTI and the Brox's poselet approach mentioned above. From Table 3, it is observed that our approach significantly outperforms these two algorithms.

Figure 3 shows some exemplar segmentation results based on our algorithm from the VOC2010 test set. The first 4 rows are mainly results with only one category while the last row contains results with multiple categories. From these results, it is observed that our method can handle background clutters, objects with low contrast with the background and multiple objects in the same image. Both the *cat* and *boat* in the middle column have particularly low contrast, while the *cat* is partially occluded by the tree. The *bird* in the right column contains a very cluttered background with low contrast against the foreground. In the last row, the two *horses* in the middle image have quite different poses, while the *cars* are partially visible behind the fence in the right image.

However, there still exist some failure cases, due to various different types of reasons. First comes from the inability to correctly select masks, the detector might predict a wrong label for the bounding box area. Also in some cases, the algorithm does not successfully handle multiple interacting objects, especially

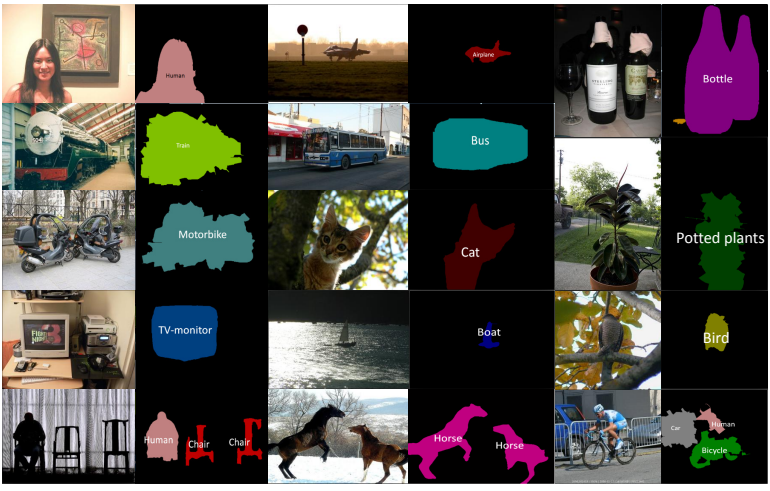

Fig. 3. Some exemplar segmentation results (Better viewed in color)

when the neighboring objects have low mutual contrast and occlude each other, such as *man* on the *bicycle*. From the last image in Figure 3, although the *human* and *bicycle* are successfully segmented, the interacting part like the human legs are still not well recovered. The third types of failure comes from the confusion of similar category pairs, like *cow* and *horse*, *dog* and *cat*, etc.. Otherwise, if the objects are correctly detected and labeled, which can be achieved through choosing detection bounding boxes with relatively higher confidence scores, such problems could be avoided. The last failure comes from the fact that sometimes the detectors cannot determine the accurate spatial bounding box of objects, especially for objects with rare poses or only partially visible. In some extreme cases, such partially visible objects are neglected by detectors and the later mask refinement procedure shall not be activated.

7 Conclusions and Future Work

In this paper, we presented a novel approach for object segmentation based on object detection by coupled global and local sparse representations. Unlike previous methods, we frame segmentation as a mask refinement problem from the coarse masks predicted from object detectors. Through global sparse reconstruction that could generally select the most similar training masks and local reconstructions that could handle locally spatial deformation, the proposed algorithm could achieve competitive results with the state-of-the-art algorithms on VOC2007 and VOC2010 benchmarks and outperforms other detection based object segmentation algorithms.

Current performance of object segmentation on VOC2010 benchmark is around 40%, which remains a great potential for improvement. One key property of our algorithm is that it heavily relies on object detection algorithms, therefore, with better object detectors in future, such as the one that could well handle partial objects and occlusions, significant improvement can be expected for object segmentation performance.

Acknowledgement. This research is for CSIDM Project No. CSIDM-200803, which was supported in part by a grant from the National Research Foundation (NRF) administered by the Media Development Authority (MDA) of Singapore.

References

1. Gonfaus, J., Bosch, X., Weijer, J., Bagdanov, A., Gual, J.: Harmony potentials for joint classification and segmentation. In: CVPR (2010)
2. Li, F., Carreira, J., Sminchisescu, C.: Object recognition as ranking holistic figure-ground hypotheses. In: CVPR (2010)
3. Brox, T., Bourdev, L., Maji, S., Malik, J.: Object segmentation by alignment of poselet activations to image contours. In: CVPR (2011)
4. Yang, Y., Hallman, S., Ramanan, D., Fowlkes, C.: Layered object detection for multi-class segmentation. In: CVPR (2010)

5. Felzenszwalb, P., Girshick, R., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. TPAMI 32 (2010)
6. Vedaldi, A., Gulshan, V., Varma, M., Zisserman, A.: Multiple kernels for object detection. In: ICCV (2009)
7. Kumar, M., Torr, P., Zisserman, A.: OBJ CUT. In: CVPR (2005)
8. Ladicky, L., Sturgess, P., Alahari, K., Russell, C., Torr, P.: What, Where and How Many? Combining Object Detectors and CRFs. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part IV. LNCS, vol. 6314, pp. 424–437. Springer, Heidelberg (2010)
9. Ladicky, L., Russell, C., Kohli, P., Torr, P.: Associative hierarchical CRFs for object class image segmentation. In: ICCV (2009)
10. Ladicky, L., Russell, C., Kohli, P., Torr, P.: Graph Cut Based Inference with Co-occurrence Statistics. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part V. LNCS, vol. 6315, pp. 239–253. Springer, Heidelberg (2010)
11. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge, VOC 2007 (2007) (results)
12. Yuan, X., Yan, S.: Visual classification with multi-task joint sparse representation. In: CVPR (2010)
13. Mori, G., Ren, X., Efros, A., Malik, J.: Recovering human body configurations: Combining segmentation and recognition. In: CVPR (2004)
14. Martin, D., Fowlkes, C., Malik, J.: Learning to detect natural image boundaries using brightness and texture. In: NIPS (2002)
15. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge, VOC 2010 (2010) (results)
16. Malisiewic, T., Gupta, A., Efros, A.: Ensemble of exemplar-svms for object detection and beyond. In: ICCV (2011)
17. Wright, J., Yang, A., Ganesh, A., Sastry, S., Ma, Y.: Robust face recognition via sparse representation. TPAMI 31 (2009)
18. Yang, J., Yu, K., Gong, Y., Huang, T.: Linear spatial pyramid matching using sparse coding for image classification. In: CVPR (2009)
19. Liu, X., Feng, J., Yan, S., Jin, H.: Image segmentation with patch-pair density priors. In: ACM Multimedia (2010)
20. Rao, S., Tron, R., Vidal, R., Ma, Y.: Motion segmentation via robust subspace separation in the presence of outlying, incomplete, or corrupted trajectories. In: CVPR (2008)
21. Elhamifar, E., Vidal, R.: Sparse subspace clustering. In: CVPR (2009)
22. Zhu, L., Chen, Y., Yuille, A., Freeman, W.: Latent hierarchical structural learning for object detection. In: CVPR (2010)
23. Chen, Y., Zhu, L., Yuille, A.: Active Mask Hierarchies for Object Detection. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part V. LNCS, vol. 6315, pp. 43–56. Springer, Heidelberg (2010)
24. Emmanuel Candes, J.R.: L1-magic: Recovery of sparse signals via convex programming (2005)
25. Nesterov, Y.: Smooth minimization of non-smooth functions. Math. Program. (2005)
26. Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., Gong, Y.: Locality-constrained linear coding for image classification. In: CVPR (2010)
27. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: CVPR (2006)
28. Lowe, D.: Distinctive image features from scale-invariant keypoints. IJCV 60 (2004)