CRF with Locality-Consistent Dictionary Learning for Semantic Segmentation

Yi Li*, Yanqing Guo, Jun Guo, Ming Li, Xiangwei Kong
School of Information and Communication Engineering, Dalian University of Technology
Dalian, China
{liyi, guojun}@mail.dlut.edu.cn; {guoyq, mli, kongxw}@dlut.edu.cn

Abstract

The use of top-down categorization information in bottom-up semantic segmentation can significantly improve its performance. The basic Conditional Random Field (CRF) model can capture the local contexture information, while the locality-consistent sparse representation can obtain the category-level priors and the relationship in feature space. In this paper, we propose a novel semantic segmentation method based on an innovative CRF with locality-consistent dictionary learning. The framework aims to model the local structure in both location and feature space as well as encourage the discrimination of dictionary. Moreover, an adapted algorithm for the proposed model is described. Extensive experimental results on Graz-02, PASCAL VOC 2010 and MSRC-21 databases demonstrate that our method is comparable to or outperforms state-of-the-art Bag-of-Features (BoF) based segmentation methods.

1. Introduction

As a fundamental problem in computer vision, semantic image segmentation [3, 5, 11, 22] has various applications ranging from pose estimation to object detection. Its task is to distribute a label from object categories to each pixel in the image. The challenges lie in that the objects may have significant variety of appearances, viewpoints as well as backgrounds, and that a single visual element (pixel, patch or super-pixel) is too small to maintain sufficient information for object categorization [20]. Therefore, a promising solution is to effectively leverage high-level object class priors as well as low-level contexture information among visual elements at the same time.

Recently, Conditional Random Field (CRF) models have been used in semantic image segmentation and achieved the state-of-the-art performance [11, 19, 20, 22]. Early works use the second-order CRFs based on unary and pairwise potentials for modeling. He *et al*. [5] propose a multi-scale framework to incorporate contextual features into CRF. TextonBoost, combining appearance, shape and context information, is established and incorporated in CRF for object recognition and segmentation [17]. These works can obtain efficient Maximum a Posteriori (MAP) solution via a graph cuts [8] or other approximate graph inference algorithms. However, the second-order CRF alone cannot deal with long-range interactions among objects or miscellaneous object deformations because of its poor expressive ability.

To address this issue, Kohli *et al*. shows the superior performance of higher-order potentials in the form of Robust P^n model to the traditional second-order CRF formulation [7]. Nevertheless, higher order brings higher computational complexity and is usually more time-consuming. Another sort of solutions focuses on shape priors, scene information or other clues of the objects. Based on Bag of Features (BoF) [1] model, [18] augments the second-order CRF by adding a global top-down categorization potential to the energy. And [19] further improves the model by the use of discriminative dictionary learning. While the mentioned methods consider discrimination of dictionary and the corresponding codes, the underlying local consistency of original input data is totally excluded. The loss of locality in original feature space impairs the use of top-down categorization information and may hinder the performance improvement. Inspired by [15], we exploit the locality consistency for segmentation.

In this paper, we propose a novel method of semantic segmentation based on an innovative CRF which can preferably preserve the locality of original data when utilizing category priors. Specifically, locality-consistent dictionary learning is integrated in CRF model to efficiently exploit local structure in both location and feature space. We further improve our model to make it adaptive, and its effectiveness is also investigated. Finally, extensive experimental results on Graz-02, PASCAL VOC 2010 and MSRC-21 show that our method is comparable to or outperforms state-of-the-art BoF based segmentation methods.

*The corresponding author of the paper is Yi Li. This work is supported by the National Natural Science Foundation of China (Grant No. 61402079).

978-1-4799-6100-9/15/$31.00 ©2015 IEEE

2. Conditional Random Fields

CRF was proposed by Lafferty *et al*. [9] and firstly used in the field of natural language processing. It directly builds the posterior distribution of the label field conditioned on the observation field. Let $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_n] \in \mathbb{R}^{m \times n}$ be an image with m patches, and $\mathbf{y} = [y_1, \cdots, y_n] \in \mathbb{R}^n$ be the corresponding labels where $y_i \in \{-1, +1\}$ with -1 denoting background and $+1$ foreground. Each patch $\mathbf{x}_i$ is a node and two nodes are connected by an edge if they are adjacent. Then a graph $G = \{V, E\}$ is built on the image where V refers to the nodes and E the edges. The CRF models the posterior distribution with a Gibbs distribution in the form of

$$P(\mathbf{y}\,|\mathbf{X}, \mathbf{w}\,) = \frac{1}{Z} \exp(-E(\mathbf{y}, \mathbf{X}, \mathbf{w})) \tag{1}$$

where $\mathbf{w}$ is the set of model parameters and Z is the normalization term. In a second-order CRF, the energy can be written as

$$E(\mathbf{y}, \mathbf{X}, \mathbf{w}) = \sum_{i \in V} \Phi_1(y_i, \mathbf{X}, \mathbf{w}) + \sum_{(i,j) \in E} \Phi_2(y_i, y_j, \mathbf{X}, \mathbf{w}). \tag{2}$$

The unary potential Φ_1 models the cost of assigning label y_i to node $\mathbf{x}_i$, while the pariwise potential Φ_2 models the cost of assigning a pair of labels (y_i, y_j) to a pair of adjacent nodes $(\mathbf{x}_i, \mathbf{x}_j)$ [19]. There are multiple CRF potentials and the adopted functions are elaborated later.

The objective is to seek an optimal labeling that maximizes the condition probability, *i.e*., achieves MAP, which is also equivalent to minimizing the energy E as is shown in

$$\mathbf{y}^* = \arg\max_{\mathbf{y}} P(\mathbf{y}|\mathbf{X}, \mathbf{w}) = \arg\min_{\mathbf{y}} E(\mathbf{y}, \mathbf{X}, \mathbf{w}). \tag{3}$$

There is an intuitive meaning of CRF which can be explained as image regions which are alike or close to each other tend to share the same category label. Conversely, those regions that are disparate or far from each other tend to have different labels.

3. Proposed method

3.1. Motivation

Locality consistency could be defined as the similar inputs (*i.e*., neighbors in feature space) have similar codes [15]. As is mentioned above, locality consistency of original input data is a common scenario in realistic applications while is often ignored. When integrating sparse coding into a CRF model, to achieve a higher segmentation quality, we wish to efficiently leverage global top-down categorization priors and preserve the locality relationship of original data synchronously. As thus, the similarities of image regions in both location and feature space are embodied in our method. Figure 1 is an illustration of our proposed model.

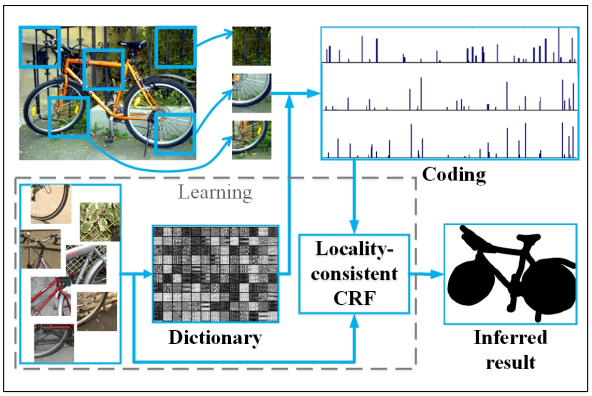

Figure 1. System overview. Given a test image, we first encode all the patches on the trained dictionary and then use locality-consistent CRF to infer the segmentation. Locality consistency is well preserved by obtaining similar codes for visually similar patches. The dictionary and CRF parameters in dashed box are jointly learned.

3.2. Model

To take full advantage of high-level categorization priors and low-level contexture information among patches, we establish our model by integrating locality-consistent dictionary learning and CRF. Assume $\mathbf{D} \in \mathbb{R}^{m \times K}$ is a dictionary to be learnt, where K denotes the dictionary size. Each m-dimensional patch descriptor $\mathbf{x}_i$ is encoded over $\mathbf{D}$ via sparse representation with locality consistency, by optimizing the following problem:

$$\mathbf{a}_i = \arg\min_{\mathbf{a}} \left\{ (1-\gamma)\|\mathbf{x}_i - \mathbf{D}\mathbf{a}\|^2 + \gamma \frac{1}{k} \sum_{\mathbf{z}_i \in \mathcal{N}(\mathbf{x}_i)} \|\mathbf{z}_i - \mathbf{D}\mathbf{a}\|^2 + 2\lambda \|\mathbf{a}\|_1 \right\} \tag{4}$$

where $0 \leq \gamma < 1$ weights the contribution of nearby points denoted by $\mathbf{z}_i$, and λ controls the sparse penalty. If $\gamma = 0$, we have a LASSO model [4] to produce a sparse representation. $\mathcal{N}(\mathbf{x}_i)$ is obtained by searching for the k nearest neighbors in image $\mathbf{X}$. Hence, the medial term is the locality constrain in the formula which enforces that the reconstruction of $\mathbf{x}_i$ still lies into its local neighborhood. It models the property mentioned above that similar original inputs lead to similar codes, *i.e*., locality consistency in feature space is preserved.

Note that $\mathbf{a}_i$ represents an optimization problem instead of a variable. Accordingly, $\mathbf{A} = [\mathbf{a}_1, \cdots, \mathbf{a}_n] \in \mathbb{R}^{K \times n}$ is incorporated in Eq. (1) as a latent variable, and the energy function in Eq. (2) becomes

$$E(\mathbf{y}, \mathbf{A}, \mathbf{w}) = \sum_{i \in V} \Phi_1(y_i, \mathbf{A}, \mathbf{w}) + \sum_{(i,j) \in E} \Phi_2(y_i, y_j, \mathbf{A}, \mathbf{w}) \tag{5}$$

in which $\mathbf{A}$ depends nonlinearly on $\mathbf{D}$. The unary and pairwise potentials can be further written as

$$\Phi_1(y_i, \mathbf{A}, \mathbf{w}) = \langle \mathbf{w_1}, -y_i \mathbf{a_i} \rangle \tag{6}$$

and

$$\Phi_2(y_i, y_j, \mathbf{A}, \mathbf{w}) = w_2 I(y_i \neq y_j) \tag{7}$$

where $\langle \cdot, \cdot \rangle$ represents the inner products and $I(\cdot)$ equals 1 if the input is true. Apparently, we have $\mathbf{w} = [\mathbf{w}_1; w_2] \in \mathbb{R}^{K+1}$.

3.3. Inference

Given the optimal CRF parameters $\mathbf{w}^*$ and the normalized dictionary $\mathbf{D}^*$, the proposed model is settled and can be utilized to infer the segmentation of a test image. Note that our energy function in Eq. (5) does not employ any complex calculation of the latent variables and it is plausible to decompose the inference into two steps. For a test image $\mathbf{X}$, we first evaluate the sparse codes $\mathbf{A}$ over $\mathbf{D}^*$ by optimizing Eq. (4). Then, the problem is converted to minimizing the energy E, which can be efficiently solved by α expansion, $\alpha - \beta$ swap, or other approximate inference algorithms.

3.4. Learning

Suppose that $\{\mathbf{X}^i\}_{i=1}^N$ is the training image set and $\{\mathbf{y}^i\}_{i=1}^N$ is the corresponding segmentation labels. Our model is linear with CRF parameters $\mathbf{w}$. However, it is nonlinear with the dictionary $\mathbf{D}$ and the dependency is implicit, which makes the learning procedure really challenging. Despite all this, if $\mathbf{D}$ is known, the proposed model turns out to be similar to most CRF models and the large-margin framework can be used for learning, which is to solve the following optimization

$$\begin{aligned} &\min_{\mathbf{w}, \{\xi_i\}} \ \frac{\beta}{2}||\mathbf{w}||^2 + \sum_{i=1}^{N} \xi_i \\ s.t. \quad &E(\mathbf{y}, \mathbf{A}^i, \mathbf{w}) - E(\mathbf{y}^i, \mathbf{A}^i, \mathbf{w}) \geq \Delta(\mathbf{y}, \mathbf{y}^i) - \xi_i, \\ &\forall i \in \{1, \cdots, N\}, \end{aligned} \tag{8}$$

where $\{\xi_i\}$ are slack variables of the constrains, $\Delta(\cdot, \cdot)$ is a loss function, $\mathbf{y}^i$ is the ground truth label and $\mathbf{y}$ denotes the precision. In general, $\Delta(\mathbf{y}^i, \mathbf{y}^i) = 0$, but $\Delta(\mathbf{y}, \mathbf{y}^i) > 0$ for any $\mathbf{y} \neq \mathbf{y}^i$. In this paper, we define $\Delta(\mathbf{y}, \mathbf{y}^i) = \sum_{j=1}^{n} I(y_j \neq y_j^i)$. The intuitive meaning of Eq. (8) is to find the $\mathbf{w}$ with small norm to make the energy of ground truth label $E(\mathbf{y}^i, \mathbf{A}^i, \mathbf{w})$ is smaller than any other incorrect labels $E(\mathbf{y}, \mathbf{A}^i, \mathbf{w})$ by at least a margin $\Delta(\mathbf{y}, \mathbf{y}^i)$ [11]. Similar with [19, 21], we first seek for the most violated constraint by solving

$$\tilde{\mathbf{y}}^i = \arg\min_{\mathbf{y}} E(\mathbf{y}, \mathbf{A}^i, \mathbf{w}) - \Delta(\mathbf{y}, \mathbf{y}^i). \tag{9}$$

Then, we rewrite the objective optimization in Eq. (8) as

$$f(\mathbf{w}, \mathbf{D}) = \frac{\beta}{2}||\mathbf{w}||^2 + \sum_{i=1}^{N} E(\mathbf{y}^i, \mathbf{A}^i, \mathbf{w}) - E(\tilde{\mathbf{y}}^i, \mathbf{A}^i, \mathbf{w}). \tag{10}$$

Inspired by [19, 21], we employ stochastic gradient descent to solve the optimization. Given the initiation $\mathbf{D}_0$ and $\mathbf{w}_0$, at each iteration, we randomly select an instance $(\mathbf{X}, \mathbf{y})$ (*i.e.*, $N = 1$) from training set to update $\mathbf{D}$ and $\mathbf{w}$ by their gradients respectively. Let $\Phi(\mathbf{y}, \mathbf{A}) \triangleq \left[\sum_{i \in V} -y_i \mathbf{a}_i; \sum_{(i,j) \in E} I(y_i \neq y_j)\right]$, $\mathbf{D}_t$ and $\mathbf{w}_t$ be the results of the t-th iteration, then the gradient of f with respect to $\mathbf{w}$ of the selected instance can be easily calculated by

$$\frac{\partial f}{\partial \mathbf{w}} = \beta \mathbf{w}_t + \Phi(\mathbf{y}, \mathbf{A}) - \Phi(\tilde{\mathbf{y}}, \mathbf{A}). \tag{11}$$

As for the gradient of f with respect to $\mathbf{D}$, since the energy depends implicitly on the dictionary, we utilize the chain rule of differentiation in the form of

$$\frac{\partial f}{\partial \mathbf{D}} = \sum_{i \in V} \left(\frac{\partial f}{\partial \mathbf{a}_i}\right)^T \frac{\partial \mathbf{a}_i}{\partial \mathbf{D}}. \tag{12}$$

In the following, we omit the subscripts i and t for simplicity. Now we calculate $\frac{\partial \mathbf{a}_i}{\partial \mathbf{D}}$. Under certain conditions, the sparse representation $\mathbf{a}$ in Eq. (4) must satisfy [13]

$$\mathbf{D}^T \left\{ \mathbf{D}\mathbf{a} - \left((1-\gamma)\mathbf{x} + \gamma \frac{1}{k} \sum_{\mathbf{z}_i \in \mathcal{N}(\mathbf{x})} \mathbf{z}_i \right) \right\} = -\lambda sign(\mathbf{a}). \tag{13}$$

Suppose the active set in $\mathbf{a}$ denoted by $\mathbf{a}_\Lambda$ does not change when $\mathbf{D}$ has a small perturbation. $\mathbf{D}_\Lambda$ is composed of the columns in $\mathbf{D}$ which are corresponding to $\mathbf{a}_\Lambda$ and let $\mathbf{M} \triangleq (\mathbf{D}_\Lambda^T \mathbf{D}_\Lambda)^{-1}$, $\mathbf{X}_\mathbf{z} \triangleq (1-\gamma)\mathbf{x} + \frac{\gamma}{k} \sum_{\mathbf{z}_i \in \mathcal{N}(\mathbf{x})} \mathbf{z}_i$. We then calculate the derivative of $\mathbf{D}$ of Eq. (13) and have

$$\frac{\partial \mathbf{a}_{(k)}}{\partial \mathbf{D}} = (\mathbf{X}_\mathbf{z} - \mathbf{D}\mathbf{a})\, \mathbf{M}_{[k]} - \left(\mathbf{D}\mathbf{M}^T\right)_{\{k\}} \mathbf{a}^T, \qquad \forall k \in \Lambda \tag{14}$$

where (k) indicates the k-th entry of $\mathbf{a}$, $[k]$ the k-th row of $\mathbf{M}$ and $\{k\}$ the k-th column of $\mathbf{D}\mathbf{M}^T$. Hence, for the selected training instance $(\mathbf{X}, \mathbf{y})$, the gradient of f with respect to $\mathbf{D}$ can be computed by

$$\begin{aligned} \frac{\partial f}{\partial \mathbf{D}} &= \sum_{i \in V} \left(\frac{\partial f}{\partial \mathbf{a}_i}\right)^T \frac{\partial \mathbf{a}_i}{\partial \mathbf{D}} = \sum_{i \in V} \sum_{k \in \Lambda_i} \frac{\partial f}{\partial \mathbf{a}_{i(k)}} \frac{\partial \mathbf{a}_{i(k)}}{\partial \mathbf{D}} \\ &= \sum_{i \in V} \sum_{k \in \Lambda_i} \mathbf{g}_{i(k)} \left(-\left(\mathbf{D}\mathbf{M}^T\right)_{\{k\}} \mathbf{a}_i^T + (\mathbf{X}_\mathbf{z} - \mathbf{D}\mathbf{a}_i)\, \mathbf{M}_{[k]} \right) \\ &= \sum_{i \in V} (\mathbf{X}_\mathbf{z} - \mathbf{D}\mathbf{a}_i)\, (\mathbf{M}\mathbf{g}_i)^T - \mathbf{D}\mathbf{M}^T \mathbf{g}_i \mathbf{a}_i^T \end{aligned} \tag{15}$$

where $\mathbf{g}_i = \frac{\partial f}{\partial \mathbf{a}_i} = (\tilde{y}_i - y_i)\mathbf{w}_1$. The parameter settings of learning algorithm are described in the next section.

4. Experimental studies

We demonstrate the effectiveness of our method by conducting experiments on the Graz-02 database, the PASCAL VOC 2010 database and MSRC-21 dataset. Two typical segmentation performance metrics are used in our experiments: accuracy and average intersection-union metric over all classes (*i.e.*, VOC measure), both of which are on pixel level. Specifically, accuracy includes the mean accuracy (the mean of per-class accuracy) denoted by "Acc-mean" and the global accuracy (the percentage of all correctly classified pixels from all classes) denoted by "Acc-global".

4.1. Parameter settings

We extracted the SIFT descriptors [12] on grid patches to represent the image and the patch size is 64×64 pixels with a shifting space of 16 pixels. As for the corresponding labels, we propagate the original pixel-level ground truth to patch-level in the following scheme: a label is set to be 1 if more than a quarter of the patch pixels are foreground, otherwise the label is -1. Note that joint $\mathbf{D}$ and $\mathbf{w}$ learning is non-convex, a good initialization is vital to the algorithm. We employ the dictionary learning algorithm in [10] to initialize the dictionary $\mathbf{D}$ and evaluate the sparse coding of training data on it. A linear SVM is trained to obtain the initialization of $\mathbf{w}$. Based on the results of [21], the dictionary size is set to 512 and the sparse penalty $\lambda = 0.15$. The weight γ in Eq. (4) is set to 0.1, $k = 3$ and the weight penalty $\beta = 1e - 5$ in Eq. (10). The inferred scores share the same scale of its patch grid. We upsample the scores to the size of the original image and obtain the final score matrix for segmentation.

4.2. Results

The Graz-02 database [14] is composed of four categories—bike, cars, person and background, each of which contains 300 images of size 480×640. Following [21], we use the 150 odd-numbered images in each category for training and the others for testing. The challenges of the dataset are the various object poses, scales and background. We compare our results with other models by pixel accuracy and average VOC measure in Table 1. Ours-wgh refers to the weighted scheme mentioned before, which means replacing $\frac{1}{k}$ with $\exp\{-(\frac{dist(\mathbf{x},\mathbf{z}_i)}{t})^2\}$ where $dist(\cdot,\cdot)$ represents the Euclidean distance and $t = 1$. The best result is achieved by our model with the Gaussian kernel weighted scheme. Moreover, there are apparent improvement of both our methods over other algorithms. With the consideration of locality in both location and feature space, our model is robust to viewpoints variation and complex background. Some visualized segmentation results are presented in Figure 2.

The PASCAL VOC 2010 database [2] consists of 1928 images from 20 categories and background class. Since official ground truth of the test data is not available, we use the training set for learning and the validation set for testing. The average VOC measure over all classes is shown in Table 2. Our results are slightly worsen than state-of-the-art BoF based methods. It reveals that the performance of our model is susceptible to the structure information on the patch size level. The reason lies in that similar patch appearance from different categories, such as bus and car, may mislead the learning and inferring results.

The MSCR-21 dataset [16] is a multi-class benchmark containing 591 images from 21 categories and the standard train-validation-test split is used. We compare the accuracy and VOC measure with two previous methods in Table 3. The results of our method is marginally better than those of the compared methods on the whole. Moreover, [22] uses detection scores and [19] needs to over-segment images into superpixels beforehand, while ours only uses grid patches, so our method successfully achieves better results with even simpler model.

Table 1. Accuracy and VOC measure on Graz-02.

Method	[6]	[19]	Ours	Ours-wgh
Acc-mean	79.3	75.4	81.2	**83.3**
Acc-global	–	87.6	90.2	**91.6**
VOC measure	53.1	56.1	56.5	**56.9**

Table 2. Results on PASCAL VOC 2010.

Method	[22]	[19]	Ours	Ours-wgh
VOC measure	**31.2**	30.3	30.6	31.1

Table 3. Results on MSRC-21.

Method	[22]	[19]	Ours	Ours-wgh
Acc-mean	79.3	78.4	79.5	**79.7**
Acc-global	86.2	84.5	**87.2**	87.1

5. Conclusion

In this paper, we propose a novel semantic segmentation method based on a new CRF with locality-consistent dictionary learning. The framework effectively leverages high-level object class priors and low-level contexture information at the same time. In addition to considering the discrimination of dictionary, our method also models the local structure in both location and feature space. Moreover, we propose a scheme to jointly learn the dictionary and the model parameters. Experimental results show that our method is comparable to or outperforms state-of-the-art BoF based segmentation methods, revealing the locality considered in our method helps ameliorate the performance.

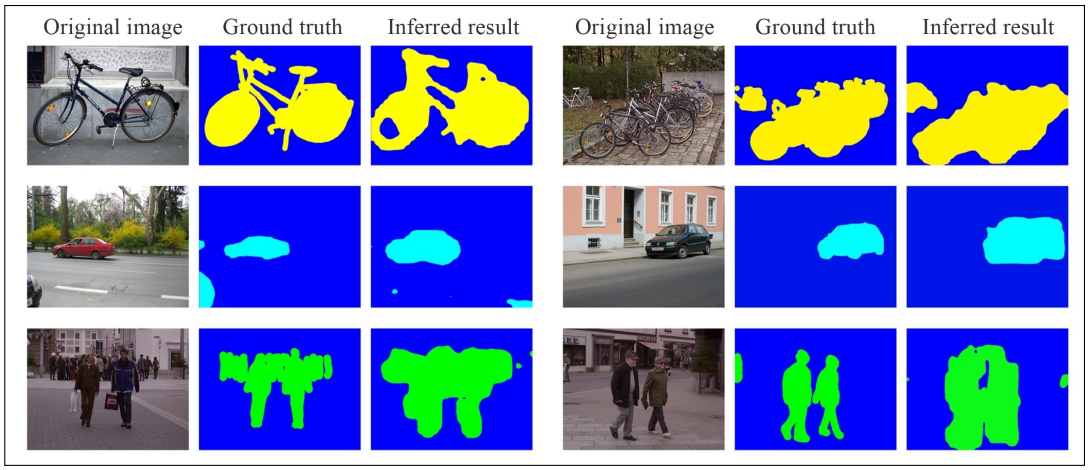

Figure 2. Inferred segmentation results on Graz-02.

References

[1] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. *Workshop on Statistical Learning in Computer Vision, ECCV*, 1(1-22):1–2, 2004. 1

[2] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010. 4

[3] S. Gould, J. Rodgers, D. Cohen, G. Elidan, and D. Koller. Multi-class segmentation with relative location prior. *International Journal of Computer Vision*, 80(3):300–316, 2008. 1

[4] T. Hastie, R. Tibshirani, and J. Friedman. Linear methods for regression. *The Elements of Statistical Learning*, pages 43–99, 2009. 2

[5] X. He, R. S. Zemel, and M. Á. Carreira-Perpiñán. Multiscale conditional random fields for image labeling. *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2:II–695, 2004. 1

[6] A. Jain, L. Zappella, P. McClure, and R. Vidal. Visual dictionary learning for joint object categorization and segmentation. *Computer Vision–ECCV 2012*, pages 718–731, 2012. 4

[7] P. Kohli, P. H. Torr, et al. Robust higher order potentials for enforcing label consistency. *International Journal of Computer Vision*, 82(3):302–324, 2009. 1

[8] V. Kolmogorov and R. Zabin. What energy functions can be minimized via graph cuts? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(2):147–159, 2004. 1

[9] J. Lafferty, A. McCallum, and F. C. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001. 2

[10] H. Lee, A. Battle, R. Raina, and A. Y. Ng. Efficient sparse coding algorithms. *Advances in Neural Information Processing Systems*, pages 801–808, 2006. 4

[11] F. Liu, G. Lin, and C. Shen. Crf learning with cnn features for image segmentation. *Pattern Recognition*, 2015. 1, 3

[12] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004. 4

[13] J. Mairal, F. Bach, and J. Ponce. Task-driven dictionary learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(4):791–804, 2012. 3

[14] A. Opelt and A. Pinz. Tu graz-02 database, 2004. 4

[15] X. Peng, L. Zhang, Z. Yi, and K. K. Tan. Learning locality-constrained collaborative representation for robust face recognition. *Pattern Recognition*, 47(9):2794–2806, 2014. 1, 2

[16] J. Shotton, M. Johnson, and R. Cipolla. Semantic texton forests for image categorization and segmentation. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2008. 4

[17] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. *Computer Vision–ECCV 2006*, pages 1–15, 2006. 1

[18] D. Singaraju and R. Vidal. Using global bag of features models in random fields for joint categorization and segmentation of objects. *2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2313–2319, 2011. 1

[19] L. Tao, F. Porikli, and R. Vidal. Sparse dictionaries for semantic segmentation. *Computer Vision–ECCV 2014*, pages 549–564, 2014. 1, 2, 3, 4

[20] J. Yang, Y.-H. Tsai, and M.-H. Yang. Exemplar cut. *2013 IEEE International Conference on Computer Vision (ICCV)*, pages 857–864, 2013. 1

[21] J. Yang and M.-H. Yang. Top-down visual saliency via joint crf and dictionary learning. *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2296–2303, 2012. 3, 4

[22] J. Yao, S. Fidler, and R. Urtasun. Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation. *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 702–709, 2012. 1, 4