Research Journal of Applied Sciences, Engineering and Technology 5(6): 2238-2244, 2013
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: September 03, 2012 Accepted: September 24, 2012 Published: February 21, 2013

Multiclass Image Segmentation Based on Pixel and Segment Level

Ling Mao and Mei Xie
Department of Electronic Engineering, University of Electronic Science and Technology of China, Chengdu, 611731, China

Abstract: Multi-class image segmentation (or pixel labeling) is one of the most important and challenging tasks in computer vision. Currently, many different methods for this task can be broadly categorized into two types according to their choice of the partitioning of the image space, i.e., pixels or segments. However, each choice of the two types of methods comes with its share of advantages and disadvantages. In this study, we construct a novel CRF model to integrate features extracted from pixel and segment levels. We exploit segments generated by Constrained Parametric Min Cuts (CPMC) algorithm in the proposed framework, instead of commonly used unsupervised segmentation method (e.g., mean-shift approach). Additionally, the recognition based on these segments is also integrated into the model, which possible corrects classification mistakes caused by the unary term based on information derived from pixel level. We experimentally demonstrate our model's quantitative and qualitative improvements over the baseline methods.

Keywords: Constrained parametric min cuts, CRF, higher order potential, non-linear support vector model

INTRODUCTION

As one of the most important and challenging tasks in computer vision, multi-class image segmentation (or pixel labeling) has received increasing attention in recent years (He *et al*., 2004; Shotton *et al*., 2006; Gould *et al*., 2008; Ladicky *et al*., 2009). The PASCAL Visual Object Classes Challenge 2007 added object class based image segmentation as the taster competition, which has been propelling this trend. Here multi-class image segmentation aims to assign each pixel in an image with a class label from a predetermined set, e.g., plane, car, people, sheep.

From the early 1990s, Markov Random Fields (MRFs) were exploited to address this problem of multi-class image segmentation (Bouman and Shapiro, 1994; Feng *et al*., 2002; Kumar and Hebert, 2003a), since these undirected graphical models allowed one to incorporate local contextual constraints in the labeling problems in a principled manner. However, the traditional MRF usually makes simplistic assumptions about the data, e.g., assuming the conditional independence of the observed data, which hinders capturing complex interactions in the observed data that might be required for classification purposes. Additionally MRF formulation often does not allow any use of data in label interactions.

Kumar and Hebert (2003b) firstly applied Conditional Random Fields (CRFs) to segment man-made structure from complex natural scenes. CRFs were proposed by Lafferty *et al*. (2001), which directly model the conditional distribution over labels given the observations and take observed data into account in label interactions. Therefore, the method presented in Kumar and Hebert (2003a) performed better than those using MRFs in Kumar and Hebert (2003b). He *et al*. (2004) and Shotton *et al.* (2006) used CRFs for semantic segmentation problems with more object classes other than two.

Turning to more recent times, many different methods have been proposed for multi-class pixel labeling, which can be broadly categorized into two types according to their choice of the partitioning of the image space. Some methods are formulated in terms of pixels (Shotton *et al*., 2006) and others used segments or groups of segments (Rabinovich *et al*., 2007; Pantofaru *et al*., 2008; Gould *et al*., 2009). Each choice of the two types of methods comes with its share of advantages and disadvantages. Those pixel-based methods assign each pixel a label using features extracted from a regularly shaped patch around it or at an offset from it Shotton *et al*. (2006). However, these small patches contain a limited amount of information. For example, they exclude useful shape-based cues or robust statistics about the appearance of larger regions. The former is very important in recognizing objects and the latter can help average out the random variations of individual pixels. Although the segment-based (or region-based) methods can avoid the problem of pixel-based methods, usually these segments do not capture

Corresponding Author: Ling Mao, Department of Electronic Engineering, University of Electronic Science and Technology of China, Chengdu, 611731, China

the boundaries between the objects in an image accurately (Rabinovich *et al*., 2007; Larlus and Jurie, 2008).

In this study, we construct a novel CRF model based on the traditional pair wise CRF model to take full advantage of information derived from the two different types of partitioning of the image space, i.e., pixels or segments. Our contributions are two-fold: first, we incorporate the segments generated by Constrained Parametric Min Cuts (CPMC) algorithm (Carreira and Sminchisescu, 2012) into the CRF model, instead of commonly used unsupervised segmentation methods, e.g., mean-shift. Second, we introduce a new kind of higher-order term, which takes into account the probability of every segment to belong to each class.

METHODOLOGY

In the following subsections, we will first introduce the CPMC algorithm and the method of predicting the likelihood of the segments generated by the CPMC algorithm to belong to each class (Li *et al*., 2010). Then we will describe how to construct the novel CRF model based on the traditional pair wise CRF model, which integrates features extracted from pixels and segments here provided by the CPMC algorithm.

Segments generated by constrained parametric min cuts algorithm: A common method to unify pixels and segments is like that described in Kohli *et al.* (2009), which enforces the labels consistent in a segment. Usually multiple segmentations are needed to assure there is at least one segment aligning with the correct boundary of objects, as shown as Fig. 1. The best segmentation for car is (d), which almost captures the correct boundary of the car except the tire and gives rise to good pixel labeling (Fig. 4).

Figure 1 multiple segmentations using different methods or parameters. (a) is an image from VOC 2007 dataset. (b) is the result from kmeans segmentation

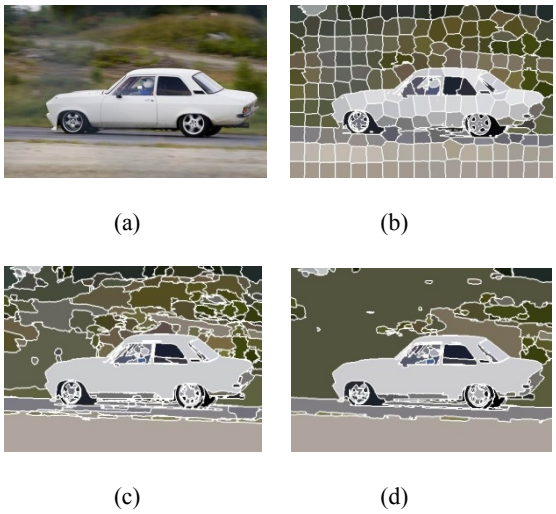

Fig. 1: Multiple segmentations using different methods or parameters (a) original image, (b) kmeans, (c) mean-shift 1, (d) mean-shift 2

Method. (c) and (d) are unsupervised image segmentation results generated by using different parameters values in the mean-shift segmentation algorithm.

However, the selection of unsupervised segmentation methods and decision of parameters values are not a trivial matter. Some methods are good for some objects, but may be bad for others, e.g., Fig. 1b is good for the person in the car, but is bad for the car. The CPMC algorithm proposed by Carreira and Sminchisescu (2012) avoids these problems to some extent.

For most images, the Constrained Parametric Min Cuts (CPMC) algorithm can create hundreds of figure-ground hypotheses and those segments covering full objects are usually ranked top 30~80 according to their prediction of putative overlap with ground truth. Figure 2 shows some examples from the 657 segments created by CPMC. There are good segments that cover the object of interest entirely, which are all ranked top

Fig. 2: Examples of segments generated by CPMC algorithm colored in green (best viewed in color)

50. The first segment in the first line overlaps the car perfectly even including the tire that results in better performance (Fig. 4). The segments shown in line 3 contain only background, which further discriminate the object of interest from the background. The segments depicted in line 2 probably cause some clutters, since they contain not only objects but also background. This problem will be resolved in section "our proposed CRF model".

When use CPMC algorithm, there are few parameters need to be adapted for different applications and segments capturing correct boundary of objects are often among top ranked ones. Therefore, we use the top ranked segments (top 50 in this study) generated by CPMC imposed on which the segment consistency constraint in our CRF model.

Categorization based on the segments: The shape-based cues or robust statistics derived from larger segments help to recognize pixels' class (Pantofaru *et al*., 2008; Gould *et al*., 2009). In this study, we exploit the approach proposed in Li *et al*. (2010). We will describe about how to incorporate the categorization results into the CRF model in section "our proposed CRF model".

Li *et al*. (2010) estimated the likelihood of each segment to belong to each class by computing the overlap between the segment and a ground truth object of that category. An image I is assumed with ground truth segments $\{G_q^I\}$. A group of segments $\{S_p^I\}$ for image I are generated by CPMC algorithm. There are also K object classes $\{c_1, c_2, \ldots, c_K\}$. K functions $f_1(S_p^I)$, …, $f_K(S_p^I)$ are learned by regression on an overlap measure Eq. (1) for segment:

$$O(S_p, G_q) = \frac{C\sqrt{|S_p|}}{\log|S_p|\sqrt{N_c^{fg}}}\frac{|S_p \cap G_q|}{|S_p \cup G_q|} + \frac{C\sqrt{|\overline{S_p}|}}{\log|\overline{S_p}|\sqrt{N^{bg}}}\frac{|\overline{S_p} \cap \overline{G_q}|}{|\overline{S_p} \cup \overline{G_q}|} \quad (1)$$

Here, N_c^{fg} and N^{bg} are the number of foreground and background pixels in the entire training set, with c the class of the ground truth segment and $\overline{S}$ is the image complement of a segment hypothesis. $C = 90$ is a normalization constant. For every putative segment S_p^I, we compute its overlap, given by (1). The target value v_{kp}^I for a segment S_p^I and a category c_k is the maximal overlap with ground truth segments that belong to c_k:

$$v_{kp}^I = \max_{G_q^I \in c_k} O(S_p^I, G_q^I) \quad (2)$$

where, $v_{kp}^I = 0$, for categories that do not appear in an image.

Finally, a non-linear Support Vector Model (SVR) is used to regress on v_{kp}^I against y_p^I, the multiple types of features from segments S_p^I. The SVR optimization problem can be derived as:

$$\begin{aligned} &\min_{w,\xi,\eta} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i + C\sum_{i=1}^{n}\eta_i \\ &\text{s.t.} \quad \xi_i \geq 0; \eta_i \geq 0, \forall i \\ &\qquad \langle w, \varphi(y_i)\rangle \geq O(v_i, v)\text{-}\varepsilon\text{-}\eta_i \\ &\qquad \langle w, \varphi(y_i)\rangle \leq O(v_i, v)+\varepsilon+\xi_i \end{aligned} \quad (3)$$

where, $\varphi(y_i)$ is a nonlinear feature transform of the input y_i, defined implicitly by the kernel $K(y_i, y_j) = \langle \varphi(y_i), \varphi(y_j) \rangle$; ε is a small constant, usually 0.05 or 0.1. It is notable that the input y_i means seven types of features, including four bags of words of gray-level SIFT and color SIFT and three pyramid HOGs (pHOG). Readers can see Li *et al*. (2010) for details.

The pair wise CRF model based on pixels: For multi-class image segmentation, CRFs are usually the basis of the most successful approaches, since these models based on CRFs unify local appearance information (such as color and texture) and a smoothness prior that enforce the labels of neighboring pixels to be the same.

The traditional pair wise CRF model is formulated as the energy function (4):

$$E(\mathbf{x}) = \sum_i E_i(x_i) + \lambda \sum_{ij} E_{ij}(x_i, x_j) \quad (4)$$

Here, $\mathbf{x}$ means the joint labeling over all pixels of a given image and all the labels are from a predefined set, e.g., person, car, sheep. The random variable x_i denotes the label assigned to pixel i (Shotton *et al*., 2006), or segment i (Gould *et al*., 2009). In this study, we adopt the former. E_i is the unary potential encoding local appearance information and E_{ij} is the smoothness term that penalizes adjacent pixels i and j for taking different labels. The non-negative constant λ trades-off the strength of the smoothness prior against the unary potential. It's notable that we omit the input features y in (4).

In our proposed model (5), the pixel-based unary term E_i is identical to that used in Ladicky *et al*. (2009) and is derived from Text on Boost (Shotton *et al*., 2006). It estimates the probability of a pixel taking a certain label by boosting weak classifiers based on a set of shape filter responses. Shape filters are defined by triplets of feature type, feature cluster and rectangular region and their response for a given pixel is the number of features belonging to the given cluster in the region placed relative to the given pixel. The most discriminative filters are found using the Joint Boosting algorithm. To enforce local consistency between neighboring pixels we use the standard contrast sensitive Potts model as the pair wise potential E_{ij} on the pixel level.

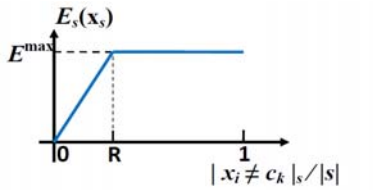

Fig. 3: Behaviour of the new higher order potential (7)

Our proposed CRF model: To integrate features from pixel and segment levels, we append higher order terms to the pair wise CRF (4):

$$E(\mathbf{x}) = \underbrace{\sum_i E_i(x_i)}_{\text{unary term}} + \lambda \underbrace{\sum_{ij} E_{ij}(x_i, x_j)}_{\text{smoothness term}} + \mu \underbrace{\sum_s E_s(\mathbf{x}_s)}_{\text{higher-order term}} \quad (5)$$

where,

$\mathbf{x}_s$: A segment from a set of image segments generated by CPMC

E_s: The higher-order potential, which enforces label consistency in $\mathbf{x}_s$. E_s could be formulated as (6) like Potts model

$|s|$: The cardinality of the segment s, which in our case is the number of pixels constituting segment s

while θ is the parameter controlling strength of the term. Formula (6) means that if the pixels in segment s are not assigned the same class label c_k, the cost $|s|^\theta$ will be added into the energy of this labeling Eq. (5). In this way, the labels of pixels in a segment tend to be the same to obtain lower energy:

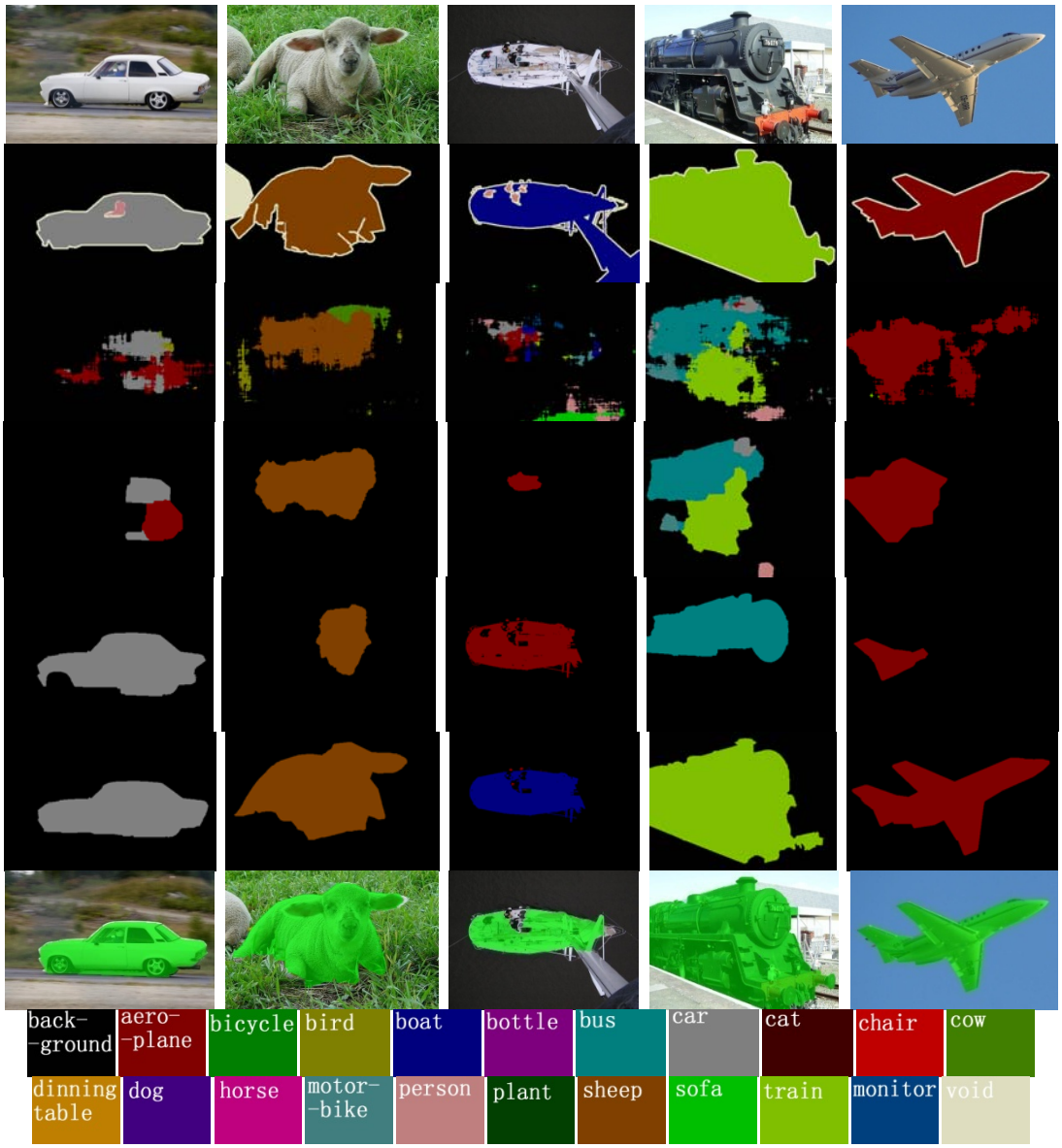

Fig. 4: Qualitative comparison of results obtained through different approaches

$$E_s(\mathbf{x}_s) = \begin{cases} 0 & \text{if } x_i = c_k, \forall i \in s \\ |s|^{\theta} & \text{otherwise} \end{cases} \quad (6)$$

The higher order potential in terms of (6) can perform very well on segments, which contain only objects of interest or background, e.g., the samples shown in line 1 and 3 of Fig. 2, but this type of potential will cause wrong labeling while encountering cluttered segments as shown in line 2 of Fig. 2. To resolve this problem, we redefine the higher order term as:

$$E_s(\mathbf{x}_s) = \min_k \left\{ \frac{|x_i \neq c_k|_s}{R|s|} E^{\max}, E^{\max} \right\} \quad (7)$$

where,

$|x_i \neq c_k|_s$: The number of pixels whose class label is different from c_k in segment s

$E^{\max}$: The assumed value of the maximum cost caused by each segment

R : The truncation parameter, which controls the ratio of pixels different from the dominant label in a segment

Unlike the old higher order potential (6), our newly defined potential (7) gives rise to a cost that is a linear truncated function of the ratio of number of inconsistent variables as shown in Fig. 3, which allows some variables x_i to take different labels from the dominant label. Therefore, our model can work well over the mixed segments. It is shown in line 2nd line of Fig. 2 and this segment can also be shown in Fig. 4.

Although the segment consistency constraint encoded by the higher order potential (7) improves the performance of original pair wise CRF model, it's almost impossible to recover from any errors caused by the basic unary potential $E_i(x_i)$. For example, in the fifth line of Fig. 4, the classification results for boat and train are wrong because of the wrong recognition based on pixel level. As known to all, shape-based cues derived from larger regions help to recognize the class of objects correctly. We incorporate the recognition results based on these shape-based cues into the term (7), in the hope that these cues can complement the features on pixel level exploited by the unary term and improve further the performance:

$$E_s(\mathbf{x}_s) = \min_k \left\{ \frac{|x_i \neq c_k|_s}{R|s|} E_s^k, E^{\max} \right\} \quad (8)$$

In formula (8), $E_s^k = -\log(f_k(s))$, where, $f_k(s) \in [0, 1]$ computed through this approach described in section "categorization based on the segments". It is easy to find that if $f_k(s)$ takes larger value, the cost E_s is smaller. In other words, variable x_i tends to take the class label c_k which is the most probable class for the segment s. Then, correct labels could be decided in the soft competition among the different potentials, which fully integrate the information from pixels and segments (see experiments).

Now we have constructed the whole CRF model which allows integration of features obtained at different levels of image partitioning, i.e., pixels and segments. The final joint labeling $\mathbf{x}$ can be determined by maximizing the objective function (5) using graph cuts (Kohli *et al*., 2009).

EXPERIMENTS AND RESULTS

Evaluation dataset: We evaluate our model on PASCAL VOC 2007 dataset. VOC 2007 is one of the most challenging datasets, which consists of 209 training, 213 validation and 210 test images for semantic segmentation task. There are 20 object classes and 1 background class. Some sample images are shown in the first line of Fig. 4. We decide the model parameters, e.g., λ, μ, $E^{\max}$, over the validation images and train the CRF model over training and validation images. Readers can refer to Ladicky *et al*. (2009) for details.

Results: Quantitative results are shown in Table 1 and some qualitative results are shown in Fig. 4.

In the experiments, the baseline models are basic unary CRF model, pair wise CRF model and associative CRF (Ladicky *et al*., 2009). The pair wise CRF model is given by formula (4), from which the smoothness term is removed gives rise to the unary CRF. These two basic CRF models consider sole information from pixel level and thus perform not so well. It is shown in line 3rd and 4th of Fig. 4. Ladicky *et al.* (2009) adds segment consistency constraint

Table 1: VOC 2007 multiclass image segmentation results on the test set obtained from pair wise CRF model (4), associative CRF model and our CRF model separately, bold numbers is denote the best performance for each class

	Average	Background	Aero plane	Bicycle	Bird	Boat	Bottle	Bus	Car	Cat	Chair
Pair wise CRF	18.2	**83**	12	**28**	24	**2**	2	25	8	3	1
Associative CRF	19.2	78	14	27	**26**	0	0	29	10	**3**	0
Cur CRF model	**20.1**	75	**18**	25	25	**3**	1	**33**	**12**	2	**2**

	Cow	Dining table	Dog	Horse	Motor bike	Person	Potted plant	Sheep	Sofa	Train	TV/ monitor
Air wise CRF	1	**23**	22	17	20	40	2	15	2	30	22
Associative CRF	1	22	**24**	**19**	22	55	2	17	0	33	22
Cur CRF model	**3**	20	22	**16**	**22**	**56**	**3**	**18**	**3**	**38**	**25**

into the basic CRF and thus discoveries the nearly correct areas of objects, e.g., the results for car and boat in the 5th line of Fig. 4. However, that depends strongly on the initial segmentations (e.g., mean-shift segmentation), since some bad initial segmentations probably cause bad results as shown in the 5th line of Fig. 4 for sheep, train and plane (please see the analysis in section "segments generated by constrained parametric min cuts algorithm"). Additionally, the associative model possibly causes wrong labeling as shown in the results for boat and train. In contrast, our model could achieve better performance as depicted in the 6th line of Fig. 4. The segments generated by CPMC can often well overlap the objects of interest (see the 7th line) and thus our model can discover the correct areas of objects. On the other hands, integration of the recognition based on segment (formula 8) obtains better semantic segmentation, e.g., the boat and train can be categorized correctly.

Quantitatively, our model provides a small increase in accuracy: 2% than the pair wise model and 1% than the associative model (Table 1).

In Fig. 4, the first line contains the original images from the VOC 2007 database, the second gives the human ground truth annotations of objects and the third, fourth, fifth and sixth show the multiclass image segmentation results obtained through the baseline unary and pair wise CRF models, associative CRF model and our proposed model separately. The 7th line shows the best segment among the top ranked 50 segments generated by CPMC. In this figure, different colors mean different object classes as shown in the last two lines. (Best viewed in color)

CONCLUSION

Many current works on multi-class image segmentation problems focus on the choice of the partitioning of the image space, i.e., pixels or segments. In this study, we have explored how to integrate information derived from both the two levels into a unified CRF model. We introduce CPMC algorithm and recognition based on it in our framework. The experiments demonstrate that our algorithm is efficient and performs better.

ACKNOWLEDGMENT

This study is supported by Sichuan Provincial Department of Science and Technology. The Grant Number is: 2010GZ0153.

REFERENCES

Bouman, C.A. and M. Shapiro, 1994. A multiscale random field model for Bayesian image segmentation. IEEE Trans. Image Process, 3(2): 162-177.

Carreira, J. and C. Sminchisescu, 2012. CPMC: Automatic object segmentation using constrained parametric min-cuts. IEEE Trans. Patt. Anal., 34(7): 1312-1328.

Feng, X.J., C.K.I. Williams and S.N. Felderhof, 2002. Combining belief networks and neural networks for scene segmentation. IEEE Trans. Patt. Anal., 24(4): 467-483.

Gould, S., T. Gao and D. Koller, 2009. Region-based segmentation and object detection. Proceedings of the 23rd Annual Conference on Neural Information Processing Systems, Vancouver, BC, pp: 655-663.

Gould, S., J. Rodgers, D.S Cohen, G. Elidan and D. Koller, 2008. Multi-class segmentation with relative location prior. Int. J. Comput. Vision., 80(3): 300-316.

He, X.M., R.S. Zeme and M.A. Carreira-Perpinan, 2004. Multiscale conditional random fields for image labeling. Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Toronto University, pp: 695-702.

Kohli, P., L. Ladicky and P.H.S. Torr, 2009. Robust higher order potentials for enforcing label consistency. Int. J. Comput. Vision, 82(3): 302-324.

Kumar, S. and M. Hebert, 2003a. Man-made structure detection in natural images using a causal multiscale random field. Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Madison, WI, United States, pp: 119-126.

Kumar, S. and M. Hebert, 2003b. A discriminative framework for contextual interaction in classification. Proceedings of the IEEE 9th International Conference on Computer Vision Nice, pp: 1150-1157.

Ladicky, L., C. Russell, P. Kohli and P.H.S. Torr, 2009. Associative hierarchical CRFs for object class image segmentation. Proceedings of the IEEE 12th International Conference on Computer Vision, pp: 739-746.

Lafferty, J., M. Andrew and C.N.P. Fernando, 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. Proceedings of the 18th International Conference on Machine Learning, Williams College, pp: 282-289.

Larlus, D. and F. Jurie, 2008. Combining appearance models and markov random fields for category level object segmentation. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Anchorage, pp: 1-7.

Li, F.X., J. Carreira and C. Sminchisescu, 2010. Object recognition as ranking holistic figure-ground hypotheses. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, C.A. San Francisco, pp: 1712-1719.

Pantofaru, C., C. Schmid and M. Hebert, 2008. Object recognition by integrating multiple image segmentations. Proceedings of the 10th European Conference on Computer Vision, Marseille, pp: 481-494.

Rabinovich, A., A. Vedaldi, C. Galleguillos, E. Wiewiora and S. Belongie, 2007. Objects in Context. Proceedings of the IEEE 11th International Conference on Computer Vision. Univ. of California, San Diego, pp: 1-8.

Shotton, J., J. Winn, C. Rother and A. Criminis, 2006. Text on Boost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. Proceedings of the 9th European Conference on Computer Vision, Graz, Austria, pp: 1-15.