Research Journal of Applied Sciences, Engineering and Technology 5(6): 2232-2237, 2013
DOI:10.19026/rjaset.5.4777
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: August 17, 2012 Accepted: September 08, 2012 Published: February 21, 2013

Research Article

Semantic Segmentation with Same Topic Constraints

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

Abstract: A popular approach to semantic segmentation problems is to construct a pair wise Conditional Markov Random Field (CRF) over image pixels where the pair wise term encodes a preference for smoothness within pixel neighborhoods. Recently, researchers have considered higher-order models that encode local region or soft non-local constraints (e.g., label consistency or co-occurrence statistics). These new models with higher-order terms have significantly pushed the state-of-the-art for semantic segmentation problems. In this study, we consider a novel non-local constraint that enforces consistent pixel labels among those image regions having the same topic. These topics are discovered by Probabilistic Latent Semantic Analysis model (PLSA). We encode this constraint as a robust Pn higher-order potential among all the image regions of the same topic in a unified CRF model. We experimentally demonstrate quantitative and qualitative improvements over a refined baseline unary and pair wise CRF models.

Keywords: CRF, higher-order potential, PLSA, topic

INTRODUCTION

Semantic segmentation aims to label each pixel in an image with a class label from a predetermined set, e.g. building, tree, face, body. For the purpose of semantic scene understanding, the task of image segmentation and labeling is a key challenge in computer vision that has received increasing attention in recent years (Feng *et al*., 2002; He *et al*., 2004; Shotton *et al*., 2006; Kohli *et al*., 2007; Ladicky *et al*., 2009). The PASCAL Visual Object Classes Challenge 2007 (VOC2007) added semantic segmentation as the taster competition, which has been propelling this trend.

In early studies, Markov Random Fields (MRFs) were commonly used, since these undirected graphical models allowed one to incorporate local contextual constraints in labeling problems in a principled manner. Bouman and Shapiro (1994) used Multiscale Random Field Models (MSRF) to segment image, where labels meant different texture types. Following the study of Bouman and Shapiro (1994) and Feng *et al*. (2002) considered the use of Tree-Structured Belief Networks (TSBNs) as prior models, successfully applied to images of outdoor scenes, with class labels such as sky, road, vegetation, etc. Similarly, Kumar and Hebert (2003) also adopted the MSRF as a prior model on the class labels (i.e., man-made structure or not) and modeled the distribution of the multiscale feature vector as mixture of Gaussians. Here the GMM capture the local dependencies in the observed data.

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. On the contrary, by using Conditional Random Fields (CRFs) proposed by Lafferty *et al*. (2001), one can directly estimate the conditional distribution over labels given the observations and thus avoid making simplistic assumptions about the data. Secondly, CRF models naturally consider observed data in label interactions. Therefore, Kumar and Hebert (2003) firstly incorporated CRFs to segment man-made structure from complex natural scenes. A generalized approach was proposed in (He *et al*., 2004), which encoded contextual information from different scales (local and global) and could be applied to complex dataset containing seven classes of objects. Shotton *et al*. (2006) exploited novel features based on texting and joint boost classifier in the CRF model, which performed best on the MSRC dataset at that time.

These CRF models are known as pair wise CRF models. In the pair wise CRF model, every pixel is associated with a random variable, both local features (encoded as unary potentials) and pair wise correlations between neighboring variables define the distribution over the joint assignment to all random variables. There has been a recent trend to improve results for semantic segmentation problems by incorporating higher-order

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

terms into the pair wise CRF models and obtain the state-of-the-art performance. These terms bias the energy-minimizing solution of the model towards one that has a more desirable label configuration, e.g., enforcing label consistency in image regions (Kohli *et al*., 2009) or biasing as few kinds of labels as possible (Ladicky *et al*., 2010).

Despite these CRF models' success, they still leave some problems to be solved. For example, multiple segmentations adopted in these studies cost too much computation and the higher-order terms often only consider adjacent regions consistent depending on local appearance matching.

Probabilistic Latent Semantic Analysis model (PLSA) is originally developed in the statistical text literature (Hofmann, 2001). Sivic *et al*. (2005) applied the PLSA model to discover both the object categories and their approximate spatial layout in unlabelled images. The result is good. In further experiments, we find that the regions with the same topic discovered by the PLSA model are likely to belong to the same objects. This observation inspires our idea imposing a constraint on these regions of the same topic. In this study, we incorporate, into a unified CRF model, novel higher-order terms that encourage consistent labeling among all the image regions with the same topic. The new higher-order terms actually encode non-local constraints and perform well on the test dataset.

METHODOLOGY

Our CRF model extends the standard pair wise CRF model (He *et al*., 2004; Shotton *et al*., 2006) for semantic segmentation by adding novel higher-order energy terms. We will first introduce the basic pair wise CRF model and then describe the unified CRF model with new higher-order terms, followed by the brief introduction of the PLSA model and details of obtaining these new higher-order energy terms.

Pair wise CRF model: Conditional random field models are originally introduced by Lafferty *et al*. (2001), of which the common type used in semantic segmentation problem is formulated as formula (1) (Kumar and Hebert, 2003):

$$P(\mathbf{x}\mid\mathbf{y}) \propto \exp\Big(\sum_i E_i(x_i,\mathbf{y}) + \lambda\sum_{ij} E_{ij}(x_i,x_j,\mathbf{y})\Big) \quad (1)$$

Here every pixel in an image of size W×H is assigned a label from a discrete label set L. The joint labeling over all pixels is denoted by $x\in L^{W\times H}$ and y represents all the features extracted from the image. E_i is the unary potential for assigning label xi to pixel E_{ij} and i is a contrast-dependent smoothing prior 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 and is chosen by cross-validation on the training set.

The final joint labeling x is decided by the Maximum A Posteriori (MAP) solution of (1), so this formula can be transformed to the equivalent energy function (2). It is notable that the observation y is omitted here. Now the values of x maximizing the energy function are the desired labeling:

$$E(\mathbf{x}) = \sum_i E_i(x_i) + \lambda\sum_{ij} E_{ij}(x_i,x_j) \quad (2)$$

In the proposed model, the pixel-based unary term E_i is identical to that used in Ladicky *et al*. (2009) and is derived from Texton 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. Triplets of feature type, feature cluster and rectangular region define shape filters 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 (Torralba *et al*., 2004). To enforce local consistency between neighboring pixels we use the standard contrast sensitive Potts model (Boykov and Jolly, 2001) as the pair wise potential E_{ij} on the pixel level.

Higher-order CRF model: We append to the pair wise CRF Eq. (2) one higher-order potential for all the regions of the same topic to give:

$$E(\mathbf{x}) = \underbrace{\sum_i E_i(x_i)}_{\text{unary term}} + \lambda \underbrace{\sum_{ij} E_{ij}(x_i,x_j)}_{\text{smoothness term}} + \mu \underbrace{\sum_t E_t(\mathbf{x}_t)}_{\text{higher-order term}} \quad (3)$$

Here $\mathbf{x}_t$ means a segment (or a super pixel) in an image, defined on which E_t is the higher-order potential, which enforces label consistency in image regions. This idea is reasonable, because these segments obtained through unsupervised segmentation method are likely to belong to the same object, i.e., the labels of those pixels in each segment are the same. Unfortunately, one segmentation does not make sure that every segment contains only one object. Hence, current works usually use multiple segmentations of an image to obtain $\mathbf{x}_t$ in the hope that there is always at least one correct segmentation (Kohli *et al*., 2009; Ladicky *et al*., 2010). However, multiple segmentations take an expensive computation time and the higher-order terms defined on those segments only consider local information in each image region. In this study, we define $\mathbf{x}_t$ as all the regions of the same topic discovered by Probabilistic Latent Semantic Analysis model (PLSA) and encode the constraint that all the regions with the same topic in the image agree on their label. Only one segmentation is used in our method, but it can perform as well as those using multiple segmentations. Additionally the method outperforms the baseline unary and pair wise CRF models described in detail in the experiments and results section.

PLSA model: Probabilistic Latent Semantic Analysis (PLSA) is originally developed in the statistical text literature and used to discover topics in a corpus using the bag-of-words document representation (Hofmann, 2001). Sivic *et al*. (2005) used this model to discover both the object categories (treated as topics) and their approximate spatial layout without supervision, i.e., find the topics present in a group of unlabelled images. The proposed unified CRF model is inspired from the observation that these regions of the same topic discovered by PLSA model probably belong to the same class of objects, as shown in Fig. 5.

The PLSA model is an unsupervised latent space model based on the Bag of Words model (BoW) (Csurka *et al*., 2004). Suppose we have a collection of images $D = d_1,..., d_N$ $D = d_1,..., d_N$ with visual words from a visual vocabulary $W = w_1,..., w_M$. The corpus of images may be summarized in an M×N co-occurrence table of counts $N_{ij} = n(w_i, d_j)$, where $n(w_i, d_j)$ stores the number of occurrences of a visual word wi in image d_j. This is the bag of words model. The approach of obtaining these visual words will be described in the following sections. In addition, there is also a hidden (latent) topic variable z ∈ $Z = z_1,..., z_k$ associated with each observation, that is the key point in the PLSA model.

A joint probability $P(w_i, d_j, z_k)$ over the M×N co-occurrence table is modeled as the form of the graphical model depicted in Fig. 1a. Marginalizing over topics z_k determines the conditional probability $P(w_i|d_j)$:

$$P(w_i \mid d_j) = \sum_{k=1}^{K} P(z_k \mid d_j) P(w_i \mid z_k) \tag{4}$$

where, P (z_k | d_j) is the probability of topic z_k occurring in image d_j and P (wi| z_k) is the probability of word wi occurring in a particular topic z_k (Fig. 3). This formula amounts to a matrix decomposition as shown in Fig. 1b with the constraint that both the vectors and mixture coefficients are normalized to make them probability distributions. For more informative descriptions of the model, refer to Sivic *et al*. (2005) and Hofmann (2001).

In learning phase, the probability P(wi|z_k) for every visual word and topic is obtained by maximum likelihood estimation and meanwhile the probability P(z_k | d_j) could be computed. The objective function is:

$$L = \prod_{i=1}^{M} \prod_{j=1}^{N} P(w_i \mid d_j)^{n(w_i, d_j)} \tag{5}$$

where, (4) give $P(w_i|d_j)$. The Expectation Maximization (EM) algorithm as described in (Hofmann, 2001) is usually adopted to get the solution. Given P(wi| z_k) and P(z_k | d_j), we can easily compute the posterior probability P(z_k|wi, d_j) (6) which decides the topic type of each segment:

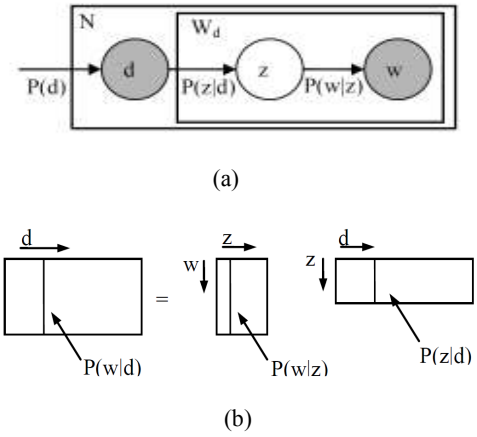


Fig. 1: PLSA model (a) PLSA graphical model, (b) $P(w_i| d_j)$ presentation

$$P(z_k \mid w_i, d_j) = \frac{P(w_i \mid z_k) P(z_k \mid d_j)}{\sum_{l=1}^{K} P(z_l \mid d_j) P(w_i \mid z_l)} \tag{6}$$

At last, we seek a vocabulary of visual words, which will be insensitive to changes in viewpoint and illumination. Here we use the same interest point detectors and descriptors as Sivic *et al*. (2005). In our experiment, all the SIFT descriptors are vector quantized into the visual words for the vocabulary. The vector quantization is carried out here by k-means clustering. Six hundred clusters are used for each of the Shape Adapted and Maximally Stable regions and the resulting total vocabulary has 1,200 words (Fig. 3).

The higher-order terms: For obtaining the regions of the same topic, we first segment the image into dozens of hundreds of regions. Provided the probability P(z_k|wi, d_j), we can determine the topic types of all visual words in each region by choosing the z_k making the maximum posterior. Then count the number of the occurrences of every topic in each region and sort these topics in descending order. If the ratio between the first two topics is bigger than some threshold, e.g., 0.6, this region is assumed to belong to the first topic, otherwise this region is ignored. Finally, we can set all the regions of the same topic as $\mathbf{x}_t$, on which a robust Pn higher-order potential is defined (Kohli *et al*., 2009)). Now we have constructed the completely unified CRF model with a novel higher-order term.

The proposed algorithm procedure: The proposed algorithm contains two main parts: training phase and test phase, detailed in the following:

Training phase:

- Use the ground truth images (see evaluation dataset section) to learn unary and pair wise potentials, i.e., E_i and E_{ij} in Eq. (3). The learning method

(a) Original frame (b) The ground truth

Fig. 2: Some sample images from the datasets
The first column (a) shows the original images and column (b) shows the ground truth images from MSRC; There are three classes: face, body and background, assigned different colors; It is notable that we just set two topics in the experiment; (best viewed in color)

is identical to the one described in (Ladicky *et al*., 2009).

- Use the original unlabelled images to learn the PLSA model (see PLSA model section).

Test phase:

- Segment the test image into some appearance consistent regions. In this study, we employ mean-shift method to find these unsupervised segments (Comaniciu and Meer, 2002).
- Obtain the distribution of visual words in each segment according to the learnt PLSA model and decide to which topic each region belongs. Now we get the higher-order term E_t that represents all the regions belonging to the same topic (see higher-order terms section).
- Determine the final joint labeling x by maximizing the objective function (3) using graph cuts (Kohli *et al*., 2009).

EXPERIMENTS AND RESULTS

We performed experiments on the multiclass pixel-labeling task and compared results on CRF models with and without our higher-order potentials.

Evaluation dataset: The evaluation dataset consists of the images of the face class from the Caltech 101 datasets (Fergus *et al*., 2003). There are 435 images in this dataset. Some examples are shown in Fig. 2a. For learning the unary and pair wise potentials in the unified CRF model, we also use the 30 ground truth images from MSRC dataset (Shotton *et al*., 2006). These 30-ground truth images correspond to the face images in Caltech, depicted in Fig. 2b. There are three classes in this data set: face, body and background. Each class is assigned a unique color. It is notable that few interest points are extracted from the body regions, so we define only two topics in the experiment, in the hope of finding face areas and background areas.

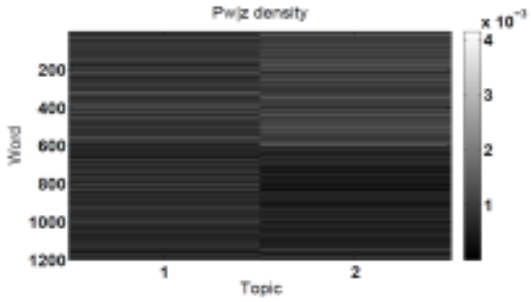
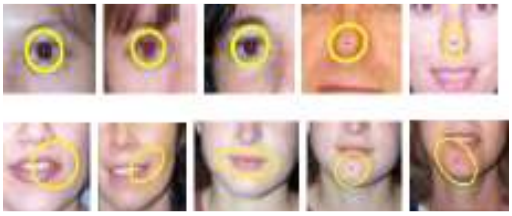


Fig. 3: The plot of the learnt P(w|z)



(a) Faces

(b) Background

Fig. 4: The most likely words for the two learnt topics. (a) Shows the visual words for the face topic, which capture the typical features of face, e.g., eye, tip of the nose, corner of the mouth, chin, (b) shows the visual words for the background topic

All the face images are used to learn the PLSA model, except the 30 images corresponding to the ground truth images in MSRC. All the values of P(wi| z_k) for every word and topic are obtained in the learning phase. We randomly split the 30 images into equal halves, 15 images for training the unary and pair wise potentials in the unified CRF model and the rest for test.

The learnt PLSA model and topic discovery: The P(w|z) learnt from the 405 original images is plotted in Fig. 3, in which the gray values mean the probability. It is also interesting to see the visual words, which are most probable for a topic by selecting those with high topic specific probability $P(w_i|z_k)$. These are shown in Fig. 4. It is easy to find that these visual words do capture typical features of the learnt two topics.

The key insight of the proposed algorithm is the discovery of those regions belonging to the same topic and constraint on the semantic segmentation by higher-

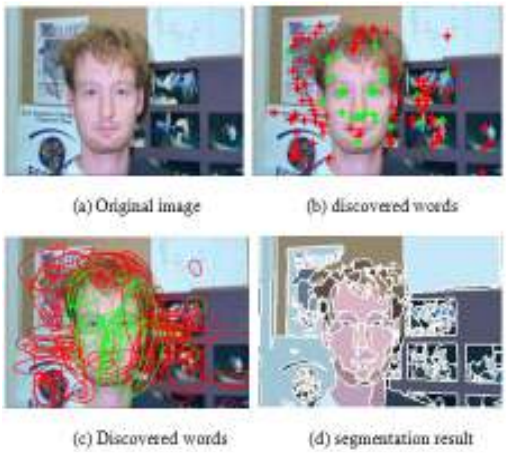
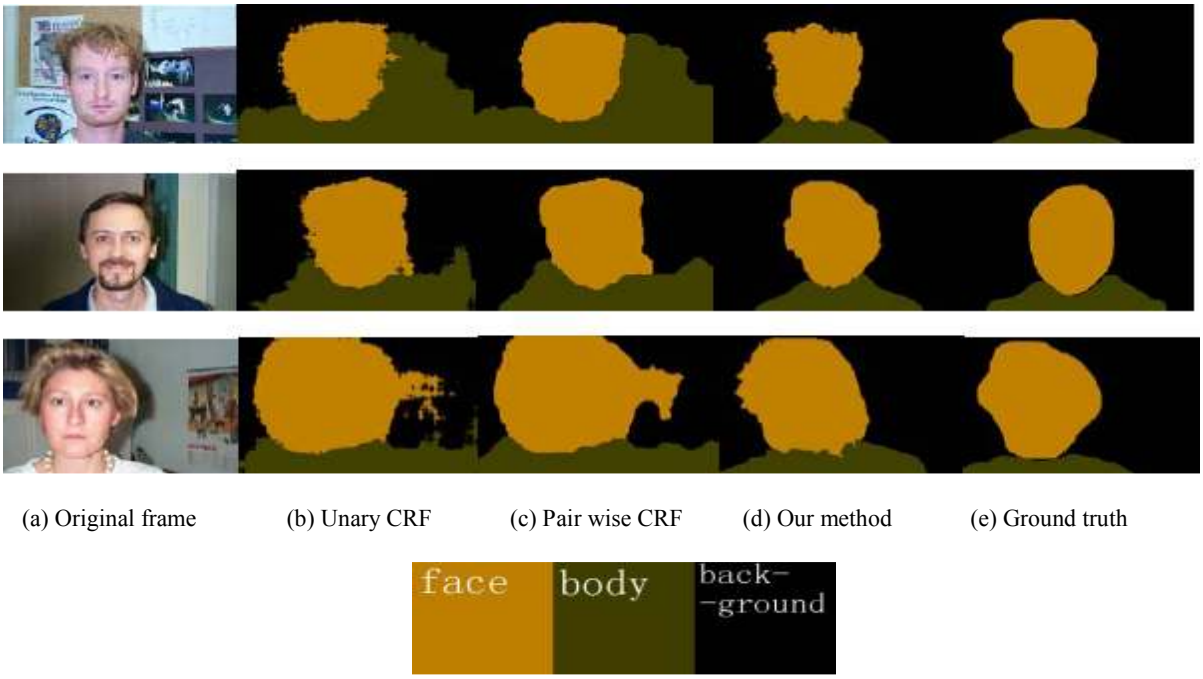


Fig. 5: Image as a mixture of visual topics using two learnt topics. (a) Is the original frame. (b) and (c) depict the image as a mixture of a face topic (green) and background topic (red) separately using cross markers or ellipses. (d) Shows the segmentation result by mean-shift (Best viewed in color)

order terms. Figure 5 gives some intuitive sense about this insight. Figure 5b-c show the discovered two visual topics face (green) and background (red). Here only visual words with P(z|w, d) greater than 0.8 are plotted. The cross markers represent the locations of detected interest points and the ellipses represent the supporting domains of each interest point. There is an impressive alignment of the words with the corresponding object areas of the image. Therefore, it is easy to decide to which topic the segments by mean-shift (shown in (d)) belong (see the higher-order terms section). Then the same topic constraint on the unified CRF model improves the semantic segmentation accuracy as shown in Fig. 6. For example, our method can recognize the body and background regions correctly which tend to be wrongly labeled in the baseline models.

RESULTS

Quantitative results are shown in Table 1 and some qualitative results are shown in Fig. 6. We compare



(a) Original frame (b) Unary CRF (c) Pair wise CRF (d) Our method (e) Ground truth

Fig. 6: Example results from our multiclass pixel labeling experiments on the Caltech 101 dataset
Each row shows a different instance; The test image is shown in column (a); Semantic class predictions for the unary, pair wise model and one with higher-order potentials constrained by same topic are shown in columns (b), (c) and (d), respectively (best viewed in color)

Table 1: Pixel wise semantic labeling accuracy for the face images from the Caltech 101 dataset

	Method			
	Baseline			
Class	Unary	Pair wise	Our algorithm	Hierarchical CRF model
Face	90.06	92.13	93.21	94.08
Body	73.29	74.89	75.98	77.04
Background	89.92	92.08	94.62	95.60

baseline unary and pair wise CRF model against our proposed model with the same topic constraint. For the three classes, the proposed higher-order potential with the same topic provides a small increase in accuracy: at least 1.1%. We note that our result is below the state-of-the-art result by Ladicky *et al*. (2009). It is not surprised, since Ladicky segmented every image multiple times with different and adapted parameters, but we just do only one segmentation.

CONCLUSION

Much recent study on semantic segmentation problems has focused on the addition of higher-order energy terms to encode preferences for particular label configurations. We have explored one such term that encodes a novel preference for consistent label assignments among those regions of the same topic. Instead of multiple segmentations, only one segmentation is used to discover these topics by an unsupervised approach. The experiment demonstrates that our algorithm is efficient and performs well.

ACKNOWLEDGMENT

Sichuan Provincial Department of Science and Technology support this study. The Grant Number is M1101020120l0GZ0153.

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.

Boykov, Y.Y. and M.P. Jolly, 2001. Interactive graph cuts for optimal boundary & region segmentation of objects in N-D images. Proceedings of the 8th IEEE International Conference on Computer Vision, Vancouver, BC, July 9-12, pp: 105-112.

Comaniciu, D. and P. Meer, 2002. Mean shift: A robust approach toward feature space analysis. IEEE Trans. Pattern Anal., 24(5): 603-619.

Csurka, G., C.R. Dance, L. Fan, J. Willamowski and C. Bray, 2004. Visual Categorization with Bags of Keypoints. Workshop on Statistical Learning in Computer Vision, Czech Technical University in Prague, May 11-14, pp: 1-22.

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

Fergus, R., P. Perona and A. Zisserman, 2003. Object class recognition by unsupervised scale-invariant learning. Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Madison, WI, United States, June 18-20, pp: 264-271.

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

Hofmann, T., 2001. Unsupervised learning by probabilistic latent semantic analysis. Mach. Learn. (Netherlands), 42(1-2): 177-196.

Kohli, P., M.P. Kumar and P.H.S. Torr, 2007. P3 & beyond: Solving energies with higher order cliques. Proceedings of the 2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, United States, pp: 1-8.

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

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, Kyoto, Sept., pp: 739-746.

Ladicky, L., C. Russell, P. Kohli and P.H.S. Torr, 2010. Graph cut based inference with co-occurrence statistics. Proceedings of the 11th European Conference on Computer Vision, Heraklion, Crete, Greece, Sep. 5-11, pp: 239-253.

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.

Shotton, J., J. Winn, C. Rother and A. Criminisi, 2006. Texton 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.

Sivic, J., B.C. Russell, A.A. Efros, A. Zisserman and W.T. Freeman, 2005. Discovering objects and their location in images. Proceedings of the 10th IEEE International Conference on Computer Vision, Beijing, China, Oct. 17-21, pp: 370-377.

Torralba, A., K.P. Murphy and W.T. Freeman, 2004. Sharing features: Efficient boosting procedures for multiclass object detection. Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, United States, Washington, DC, pp: 762-769.