Object Segmentation Using Structural Relationship between Super-pixels

Yonghui Gao[1, a], Lei Zhou[2] and Xiaoxiao Li[3,b]

[1,2] School of Medical Instrument and Food Engineering,

University of Shanghai for Science and Technology, Shanghai, China

[3] College of Sciences, Shanghai Institute of Technology, Shanghai, China

[a]email: gaoyonghui1978@163.com, [b]email:xiaoxiao_li@sina.com

Keywords: Segmentation; Super-pixels; Structural Relationship; CRF; Patches.

Abstract. We address the problem of describing and integrating long range information efficiently, such as the information demonstrated by super-pixels (patches), into conditional random field (CRF) model for object segmentation. For those purpose, a novel structural relationship between patches are defined for evaluating super-pixels' similarity. The structural relationship between super-pixels will focus on whether two patches can display similar information of objects' global appearance. Furthermore, a regression model is learned for super-pixels classification based on analyzing their structural relationship between super pixels and initial object hypothesis. Finally, a pixel-level CRF model that integrates information of color, texture and super-pixels is constructed to obtain segmentation results. Compared with traditional super-pixels or solely pixels based model, our method can combine the complementary information provided by pixels and super-pixels and generate better performance.

Introduction

Image segmentation is an important problem in computer vision, which focuses on partitioning image pixels into several distinct regions. There are mainly two classes of segmentation algorithms. One class includes automatic segmentation methods, such as active contour [1], normalized cut [2] and clustering [3]. The other class relates to interactive methods with user guidance. In interactive segmentation, users can label pixels as foreground or background with interaction approaches. These user guidance may be helpful to reduce the complexity of pattern modeling as well as its ambiguity for segmentation. In the past few years, various interactive segmentation methods have been proposed [4, 5, 6].

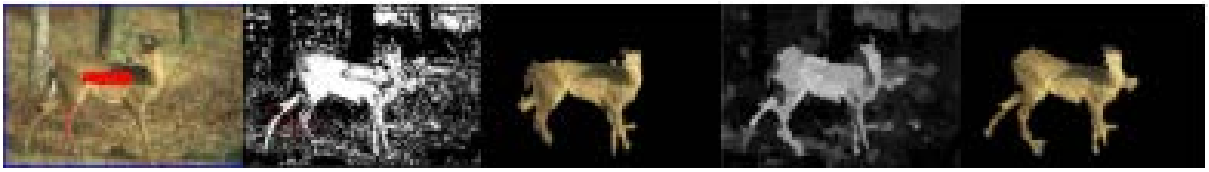
As to integrate super-pixels information into segmentation, many methods have been proposed. In [7], a tree-structured conditional random field is constructed to integrate prior in patches. In [8], image is divided into several patches via over segmentation. Then patches are labeled via specially trained models. In [9], the image is represented by a hierarchical segmentation tree then a "pylon" model is built to combine the segments come from different layers of the tree. In [10], a graph-based image segmentation method (patch-cuts) that incorporates features and spatial relations obtained from image patches is presented. In [11], the robust PN model is proposed to ensure label consistency of pixels in the same super-pixels. Unlike those models, we take the prior information provided by the super-pixels as a global feature which is integrated into CRF model with optimal weighting.

Our intuition is when the pattern exhibited by pixels is sparse, method such as graph cut may fail due to it cannot find a path. As shown in Fig.1 (a), the pixels displayed in red-boxes cannot display discriminative cues for classification. By introducing long-range cues (super-pixels' information) as the global features, pixels will display more powerful intermediate representation of an object. Our main contributions are the following:

1). We construct a novel method for measuring patches' similarity, considering whether they can display similarity global appearance.

2). A patch classification method is presented and the results are projected on pixel level.

3). A pixel-level CRF model is constructed for segmentation, based on optimal feature fusion, followed by patches classification.



(a) (b) (c) (d) (e)

Fig. 1. From left to right: (a) Initialized seeds. (b) The hypothesis generated by gaussian mixture model learned from seeds. (c) Result of GC. (d) Hypothesis of combined features. (e) Result of proposed method.

Formulation

Image segmentation can be modeled with a conditional random field (CRF). Consider a random field F defined over a set of variables $\{F_1, F_2, ..., F_N\}$. The domain of each variable is a set of labels $L = \{\ell_1, \ell_2, ..., \ell_k\}$. Denote C the semantic label. Let $I = \{I_1, .., I_N\}$ be the observed data corresponding to image information. I_i is the feature vector at pixel i. F_i represents the label assigned to pixel i. A CRF model is described by a Gibbs distribution:

$$P(F, C \mid I, w) = \frac{1}{Z} e^{-\sum_{c \in C_G} E(F_c \mid I_c, w, \psi)} \tag{1}$$

where $G = (v, \varepsilon)$ is a graph on F and c is a clique belonging to a set of cliques C_G. The weigh parameters w is a $N \times 2$ matrix $w_i = \{w_1, ..., w_{|V|}\}$. w_i is a N-dimensional vector $[w_{i1}, w_{i2}, ..., w_{iN}]^T$ and N is the feature dimension. Z is the normalizing coefficient. In our approach, an energy function E is designed as the linear function of parameters and the prior information

$$E(F, C \mid I, w) = \sum_{i \in V} w_i^T . E^{(1)}(F_i \mid I_i) + E^{(2)}_{pair}(F_i, F_j \mid I). \tag{2}$$

The energy function $E^{(1)}_i(F \mid I)$ and $E^{(2)}_{i,j}(F_i, F_j \mid I)$ is expressed in terms of single-site and pair-site clique potentials, which means node feature function and feature vector for edge (i, j) with respective node labeling F_i and F_j. Three features, global feature P_{gb}, color model P_{gmm} and texture P_{tex} are used to define the single-site model. $E^{(2)}_{pair}$ is a pairwise term between neighboring pixels.

Information Provided by Super-pixels

Global Appearance Generated by Patches.

The basic assumption is that the same global appearance can be learned from similar patches. Mainly there is the difference between GMM parameters, we want to measure it by hypothesis. It is measured by the information generated by pixels in related patches. The spatial distribution of figure-ground hypothesis generated via pattern learned from a patch. To train GMM for patch i and the related figure-ground is described using H_i. The basic assumption is similar patches may display similar hypothesis. The classifier is written as:

$$H_q(i) = \begin{cases} \{1\}, P_{gmm+tex}(F_i = 1) > P_{gmm+tex}(F_i = 0) \\ \{0\}, P_{gmm+tex}(F_i = 1) \le P_{gmm+tex}(F_i = 0) \end{cases} \tag{3}$$

where i is pixels in region q and $P_{gmm+tex}$ is combined feature of color and texture and averaging strategy is applied. We take image of antelope shown in Fig.1 for example. From Fig.2, even if in the complex scene, confusing part, such as hindquarter, have higher overlap soccer compared with grass region.

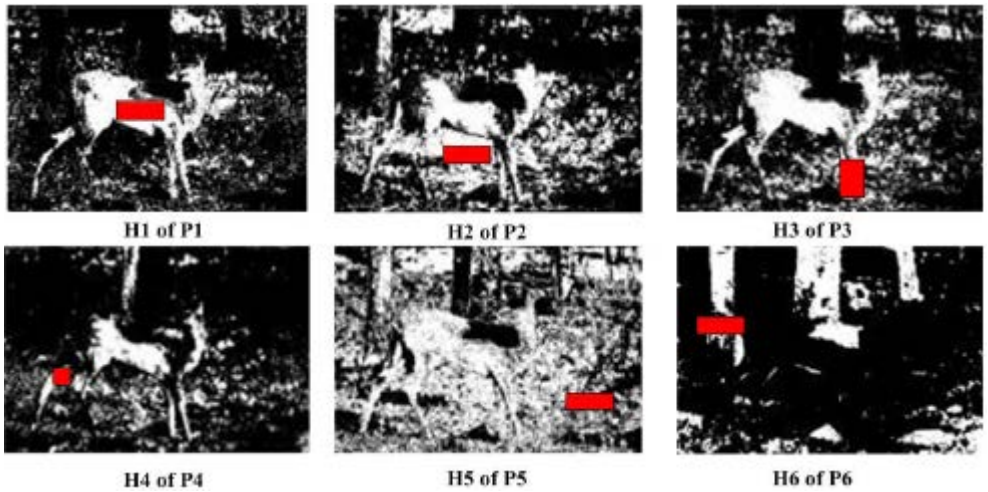


Fig. 2. From Hypothesis generated via statistics of pixels in related patch labeled as red.

Structural Relationship between Random Patches.

The intuition is that if the two patches can reveal similar statistics, the related object hypotheses learned by them may be similarity. The patch set is defined as $\{P_1,...,P_n\}$. The similarity matrix $M_{n\times n}$ between all patches is defined, $M(i,j), i,j \in 1,..,n$ means the similarity probability with patch i and j.

$$M_p(i,j)=\frac{|H_i \cap H_j|}{|H_i|}, M_n(i,j)=\frac{|\overline{H_i \bigcup H_j}|}{|\overline{H_i}|}, \tag{4}$$

where H_i represents the segmented foreground area of initial hypothesis generated by statistic of patch i and its complement is $\overline{H_i}$. $M(i,j)=exp((M_p(i,j)+M_n(i,j))/2-1)$. Relationship of six patches in Fig.2 with S is listed in Table 1.

Table 1. Relationship with Seed Patch.

	P1	P2	P3	P4	P5	P6
$M(s,*)$	0.57,	0.51,	0.54,	0.57,	0.48,	0.41
$M(*,s)$	0.65,	0.55,	0.61,	0.59,	0.52,	0.39

Table 2. Relationship between Patches.

	P1	P2	P3	P4	P5	P6
P1	1.00	0.64	0.69	0.66	0.62	0.56
P2	0.65	1.00	0.66	0.68	0.63	0.57
P3	0.70	0.66	1.00	0.65	0.63	0.54
P4	0.64	0.65	0.63	1.00	0.61	0.59
P5	0.64	0.67	0.69	0.61	1.00	0.54
P6	0.58	0.59	0.58	0.59	0.59	1.00

Clearly, patches P1, P3, P4 demonstrate higher similarity with initialized hypothesis compared with other patches. Pairwise Relationship between six patches is listed in Table 2. P1, P4 and P3 belong to object, they demonstrate higher similarity rate with each other. For example, P3 and p4 is more close to P2 compared with other patches. Encouraged by those observation, the structural relationship is coded into CRF model.

Global Classification of Super-pixels.

When pixels cannot exhibit continuous pattern, method such as graph cut may fail due to it cannot find path. To modify the probabilities given by the local pixel-level classifiers, a global level classification is presented considering super-pixels' structural relationship between initial segmentation. For segmentation, the initial object hypothesis s can be obtained by users' input or objects' prior model. The patch i's overlap rate with initial object hypothesis $P_{ov}(i,s)=\frac{H_i \bigcap H_s}{H_i \bigcup H_s}$ is computed firstly. The histogram P_{hist} of color and texture information in a region is used for classification. Moreover, the similarity between initial segmentation and hypothesis generated by each patch $P_{hs}(i)$ is taken as a feature on image level. Then the sparse feature set $f_{sp}=\{P_{ov},P_{hist},P_{hs}\}$ is used for training a regression model and the region i's association probability with object is $P_{reg}(i)=\mu^T f_{sp}$, where μ is the parameters for regression model. As shown in Fig.1, the combined feature is more discriminative combined with the region probability .

Energy Function

Local Features.

Our CRF model is defined on pixel level and three features, including local and global features, are integrated into the model. We choose color and texture as the local features. The gaussian mixture model (GMM) is used for describing color model. Let $\mathrm{P}_r(I_q;F_q=l)$ be the likelihood probability density function (PDF) of the color on a pixel p associated with label l. The color models are represented as GMM $\{\alpha_c,\mu_c,\Sigma_c\}_{c=1}^{C}$ in the color space, where $\{\alpha_c,\mu_c,\Sigma_c\}$ represent the weight, the mean color and the covariance matrix of the c th component. For pixels labeled as BG or FG, two sets of GMM parameters are learned. The variance of pixel x associated with pixels labeled as l is defined as:

$$V(l\mid I_x)=\sum\nolimits_c \alpha_{cl}N(I_x\mid\mu_{cl},\Sigma_{cl}),l\in\{BG,FG\}. \tag{5}$$

where $\{\alpha_{cl},\mu_{cl},\Sigma_{cl}\}$ represent the weight, the mean color and the covariance matrix of the c th component learned from pixels labeled as class l. Then, the likelihood probability density function (PDF) of the color on a pixel p associated with label l is represented as

$$\mathrm{P}_{gmm}(I_q;F_q=l)=\frac{V(l\mid I_x)}{\sum_{m=1}^{|L|}V(m\mid I_x)}. \tag{6}$$

The likelihood probability density function (PDF) of the texture on a pixel p associated with label l. $P_{tex}(I_q;F_q=l)$ is the similarly defined as P_{gmm}.

Global Features.

Regions form a powerful intermediate representation, but our end goal requires us to make pixel level decisions. The term will express the dependency between the local pixels and other pixels whose labels are inferred jointly. The patch scores should be projected onto pixels. For this purpose, we present a method to estimate probability associated with certain class of pixels rather than of regions. The estimated likelihood $P_{gb}(l\mid x_i)$ of pixel x_i from related region likelihood P_{reg} is defined as the weighted average of its corresponding region likelihood.

Let $P_{gb}(l\mid x_i)=\sum_{k=1}^{N_y}P_{ik}P_{reg}(l\mid R_k)$, $P_{ik}=w_{ik}/\sum_m w_{im}$ measure pixel i's association rate with region k. Then the related potential is written as:

$$E_1^{App}(F \mid I) = \{V_1^{App}(F_1), ..., V_N^{App}(F_N)\}^T,$$
$$V_1^{App}(F_q = l) = -\log(\mathrm{P}_{gmm/tex/gb}(I_q; F_q = l)). \quad (7)$$

Pairwise Term.

The first part is global pairwise reflected by PATCH. The distances between color and texture histograms are computed. Regions of the same material will often have similar texture histograms, regardless of differences in shading. When regions have both similar color and texture, they are likely to be same illumination pairs. The distance between two regions is taken into account. Features' difference between two regions/pixels is measured using the χ^2 distance of likelihood probability related, and it is defined as [12],

$$\chi^2(p,q) = \sum_i \frac{(P_r(p_i) - P_r(q_i))^2}{P_r(p_i) + P_r(q_i)}. \quad (8)$$

where p and q denote two selected regions (pixels). Then, the distance is transformed into a log likelihood ratio. The distance between two regions i and j is measured using color and texture histogram, χ^2_{gmm} and χ^2_{tex}, described as $R_{dis}(i,j) = exp(-|\chi^2_{gmm}(i,j) + \chi^2_{tex}(i,j)|/\sigma)$. Then the pairwise term between pixel i and j is composed of two parts $E^{(2)}_{pair}(i,j) = E^{(2)}_{pixel}(i,j) + E^{(2)}_{patch}(i,j)$:

$$E^{(2)}_{pixel}(i,j) = exp(-|x_i - x_j|/\sigma), i, j \in NEB$$
$$E^{(2)}_{patch}(i,j) = \lambda M(P_i, P_j) + \beta R_{dis}(P_i, P_j), P_i, P_j \in S - NEB, \quad (9)$$

where NEB is set of pixels in neighborhood, P_i and P_j denote the patches the pixels i and j belong to. $S - NEB$ is the set of super-pixels. The structural relationship and region distance is weighted. The results of pixels based (Graph cut), patch merging (MSRM), global based model (Patch formulation) and ours are listed. The goat in last row is typical failure case.

Experiment

To learn global weights for each feature in our model, we use the structural output support vector machine framework [13]. Unlike the schemes in [14] and [15], we mainly focus on performance of classification on unary term of pixels without considering the pairwise relationship. To evaluate the discrimination and to measure the related importance of feature, the feature set that we formed for parameter learning is $F = \{P_{gmm}(1) - P_{gmm}(0), P_{tex}(1) - P_{tex}(0), P_{gb}(1) - P_{gb}(0)\}$. To enforce valid constraint on w, the energy of ground truth $\overline{y}$ should be less than any segmentation mask y. That is $w^T E_1(x,y) - w^T E_1(x,\overline{y}) > 0$. We introduce formulation of margin rescale described in [13] to learn optimal parameters**.**

Interactive Figure/Ground Segmentation.

We test our method on 100 images from Berkeley segmentation database [14] and the Grab-Cut database [16]. The initial super-pixels is generated using Mean-shift [3]. Our method is compared with graph cut modeled on pixel-level [16], MSRM [17] using region merging, and the patch based formulation of our method (Take a super-pixel as a node). In the implementation of patch based method, color and texture histogram is taken as feature for patch and pairwise term is calculated as distance between histogram.

As shown in Fig.3, graph cut and MSRM will fail in the case when the pattern of pixels is disconnected, while our combining strategy generates a better performance. The numeral results are displayed in Table 3, pixel-level model combined with patch information is more flexible. It achieves highest TPF and TNF.

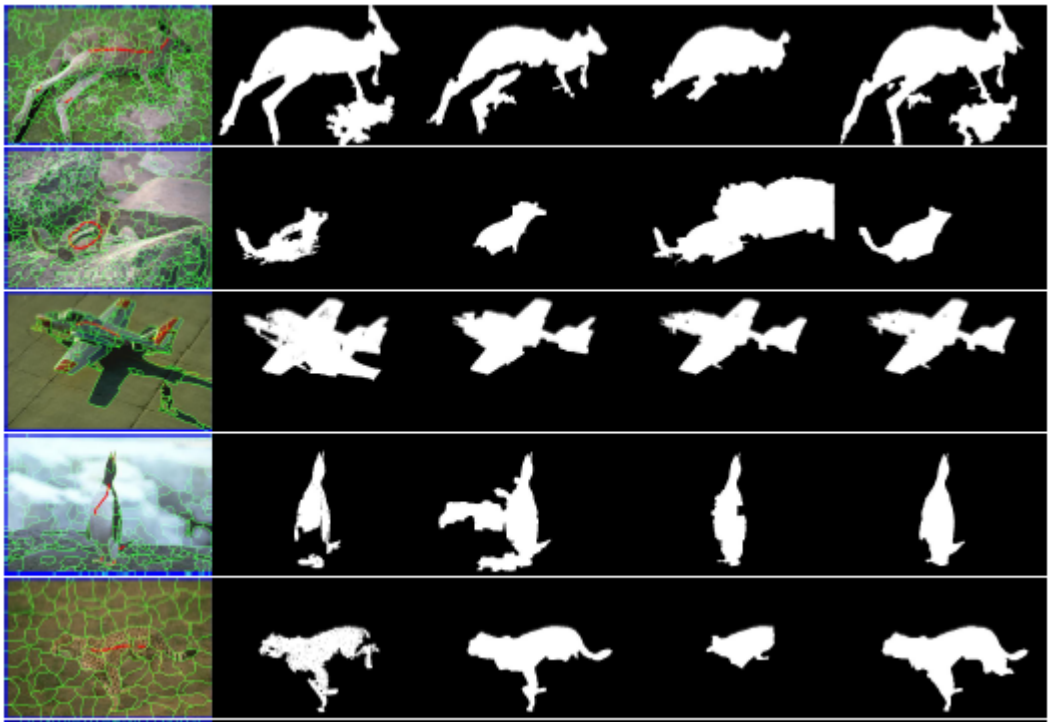


Fig. 3. Interaction segmentation result of typical hard examples. The initialized Green-super-pixels, red-seeds for object, blue-bk seeds. The results of Pixels based (Graph cut), Patch merging (MSRM), Global based Model (Patch formulation) and ours are listed.

The Jaccard index is used to measure the overall performance and it is the ratio of the areas of the intersection between the segmented object and the ground truth, and of the union. $JI = \frac{TPF}{TPF + FNF + FPF}$. We can see that single pixel or patch based method may not be able to generate satisfied results. GC shows the highest FNF, it may suffer from over-segmentation easily. Patch-based method shows the highest FPF, it may suffer from short-cutting easily. Pixel-based miss-classify and patch-based will miss many details.

Table 3. The TNF, TPF, FNF, FPF and Jaccard index results of different methods are reported.

	TPF(%)	FNF(%)	TNF(%)	FPF(%)	JI(%)
GC	85.40	7.66	92.34	14.60	79.31
MSRM	84.36	5.13	94.87	15.64	80.13
Patch	80.42	4.99	95.01	19.58	76.54
Proposed	89.21	3.58	96.42	10.79	86.22

Automatic Figure/Ground Segmentation.

We also explore the way for extending our framework to automatic segmentation. Since the most salient region is often associated with the most salient object, we select it as the initial region for the object segmentation. The bottom-up visual saliency model (GBVS) is used to locate and generate the most salient region of the image. The saliency map is generated by combining multi-scale image features including color, intensity and orientation into a single topographical saliency map [18]. A threshold operation with $P_{gvbs} > 0.8$ is used to extract the salient object as initialization.

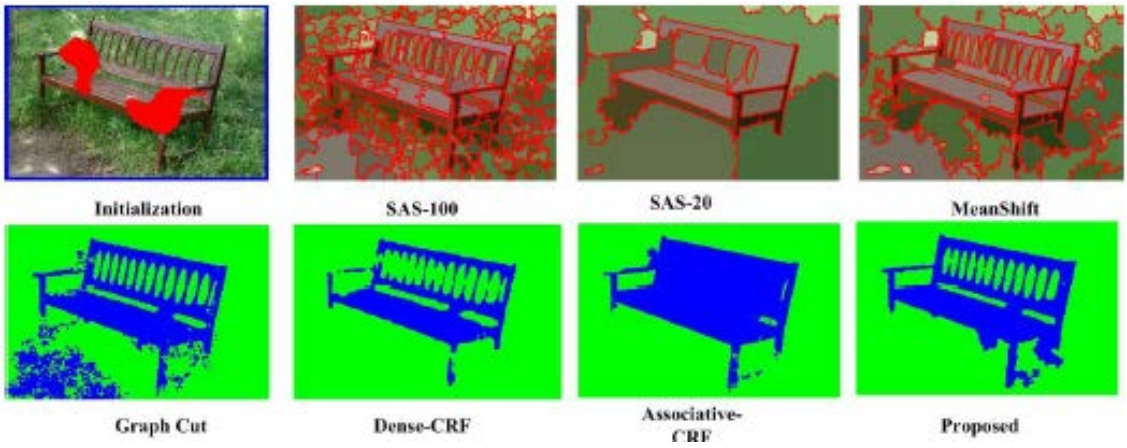


Fig. 4. Typical results in automatic framework.

Our method is compared with Mean-shift [3], SAS [19] (separated into 20/100 regions) which is designed to aggregate multi-layer super-pixels, Graph-cuts with initialization [16], dense CRF (Fully connected CRF) [20] and associative CRF [21] (in which PN model is used to ensure label consistency). As shown in Fig.4, our model generates the best result using the global features.

Conclusion

In this paper, we propose a novel method for combining information provided by super-pixels and pixels in a CRF model. Different from other methods, the patch information is evaluated using the global appearance exhibited by each patch and the information is projected onto pixels. Structural SVM is applied to learn the parameters in CRF model. Results show that our hybrid model can improve the accuracy.

References

[1] V. Caselles, R. Kimmel, and G. Sapiro, "Geodesic active contours," International journal of computer vision [J], vol. 22, no. 1, pp. 61–79, 1997.

[2] J. Shi and J. Malik, "Normalized cuts and image segmentation," Pattern Analysis and Machine Intelligence, IEEE Transactions on [J], vol. 22, no. 8, pp. 888–905, 2000.

[3] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," Pattern Analysis and Machine Intelligence, IEEE Transactions on [J], vol. 24, no. 5, pp. 603–619, 2002.

[4] Y.Y. Boykov and M.P. Jolly, "Interactive graph cuts for optimal boundary & region segmentation of objects in nd images," in Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on [C]. IEEE, 2001, vol. 1, pp. 105–112.

[5] L. Grady, "Random walks for image segmentation," Pattern Analysis and Machine Intelligence, IEEE Transactions on [J], vol. 28, no. 11, pp. 1768–1783, 2006.

[6] X. Bai and G. Sapiro, "Geodesic matting: A framework for fast interactive image and video segmentation and matting," International journal of computer vision [J], vol. 82, no. 2, pp. 113–132, 2009.

[7] J. Reynolds and K. Murphy, "Figure-ground segmentation using a hierarchical conditional random field," in Computer and Robot Vision, 2007. CRV'07. Fourth Canadian Conference on [C]. IEEE, 2007, pp. 175–182.

[8] N. Plath, M. Toussaint, and S. Nakajima, “Multi-class image segmentation using conditional random fields and global classification,” in Proceedings of the 26th Annual International Conference on Machine Learning [C]. ACM, 2009, pp. 817–824.

[9] V. Lempitsky, A. Vedaldi, and A. Zisserman, “A pylon model for semantic segmentation,” in Proceedings of Advances in Neural Information Processing Systems (NIPS) [C], 2011.

[10] G. Brunner, D.R. Chittajallu, U. Kurkure, and I.A. Kakadiaris, “Patch-cuts: A graph-based image segmentation method using patch features and spatial relations,” in Biologically Motivated Computer Vision (BMCV), IEEE Conference on [C], 2010.

[11] P. Kohli, L. Ladick` y, and P.H.S. Torr, “Robust higher order potentials for enforcing label consistency,” International Journal of Computer Vision [J], vol. 82, no. 3, pp. 302–324, 2009.

[12] X. Ren and J. Malik, “Learning a classification model for segmentation,” in Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on [C]. IEEE, 2003, pp. 10–17.

[13] B.T.C.G.D. Roller, “Max-margin markov networks,” in Advances in Neural Information Processing Systems 16: Proceedings of the 2003 Conference [C]. MIT Press, 2004, vol. 16, p. 25.

[14] M. Szummer, P. Kohli, and D. Hoiem, “Learning crfs using graph cuts,” Computer Vision–ECCV 2008 [C], pp. 582–595, 2008.

[15] Z. Kuang, D. Schnieders, H. Zhou, K.Y.K. Wong, Y. Yu, and B. Peng, “Learning image-specific parameters for interactive segmentation,” in Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on [C]. IEEE, 2012, pp. 590–597.

[16] V. Gulshan, C. Rother, A. Criminisi, A. Blake, and A. Zisserman, “Geodesic star convexity for interactive image segmentation,” in Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on [C]. IEEE, 2010, pp. 3129–3136.

[17] J. Ning, L. Zhang, D. Zhang, and C. Wu, “Interactive image segmentation by maximal similarity based region merging,” Pattern Recognition [J], vol. 43, no. 2, pp. 445–456, 2010.

[18] Harel J, Koch C, and Perona P, “Graph-based visual saliency,” in Neural Information Processing Systems (NIPS) [C], 2006.

[19] Z. Li, X.M. Wu, and S.F. Chang, “Segmentation using superpixels: A bipartite graph partitioning approach,” in Computer Vision and Pattern Recognition(CVPR), 2012 IEEE Conference on [C]. IEEE, 2012, pp. 789–796.

[20] P. Krähenbühl and V. Koltun, “Efficient inference in fully connected crfs with gaussian edge potentials,” arXiv preprint [J], arXiv:1210.5644, 2012.

[21] L. Ladicky, C. Russell, P. Kohli, and P.H.S. Torr, “Associative hierarchical crfs for object class image segmentation,” in Computer Vision, 2009 IEEE 12th International Conference on [C]. IEEE, 2009, pp. 739–746.