JOINT ROAD NETWORK EXTRACTION FROM A SET OF HIGH RESOLUTION SATELLITE IMAGES

O. Besbes

COSIM Lab., ISITCOM, Sousse Univ.,
G.P.1 Hammam Sousse, 4011, Tunisia
olfa.besbes@supcom.rnu.tn

A. Benazza-Benyahia

COSIM Lab., SUP'COM, Carthage Univ.,
Cité Technologique, 2080, Tunisia
benazza.amel@supcom.rnu.tn

ABSTRACT

In this paper, we develop a novel Conditional Random Field (CRF) formulation to jointly extract road networks from a set of high resolution satellite images. Our fully unsupervised method relies on a pairwise CRF model defined over a set of test images, which encodes prior assumptions about the roads such as thinness, elongation. Four competitive energy terms related to color, shape, symmetry and contrast-sensitive potentials are suitably defined to tackle with the challenging problem of road network extraction. The resulting objective energy is minimized by resorting to graph-cuts tools. Promising results are obtained for developed suburban scenes in remotely sensed images. The proposed model improve significantly the segmentation quality, compared against the independent CRF and two state-of-the-art methods.

Index Terms— Road network, joint segmentation, CRF.

1. INTRODUCTION

Automated road network extraction from a satellite image is a challenging task with important applications in mapping and remote sensing. The well-known computer vision techniques alone are not relevant to automatically extract the road network from a given image due to several constraints. Indeed, the road networks observed in a high resolution satellite image may exhibit a wide variability in visual appearance (e.g. spectral response, shape, contrast) that it is complex to model accurately. Roads can also be occluded by other nearby objects like buildings and trees and, even vehicles especially for high resolution images.

Most methods have attempted to solve this problem using a semi-supervised or semi-automatic system as they rely respectively on a training stage or a manual identification of seed points. They can be grouped as knowledge-based methods [1], mathematical morphology [2], snakes [3], marked point process [4] and classification [5–9]. In [1], the geometric and radiometric properties of road are expressed and then a top-down process is applied to check their validity on image regions. This knowledge-based method fails to cope with large intra-class variability. In [2], mathematical morphology operations are applied to preserve the elongated road areas, filter non-road objects and bridge gaps due to shadows, overhanging trees etc. However, gaps may remain if roads are completely broken and there is no useful information assisting the linkage. Subsequently, the road segmentation is refined with a pair of coupled active contours [3]. In [4], an object-based probabilistic representation is derived with marked point processes to integrate priors on the connectivity and intersection geometry of roads. The drawbacks are due to their difficult parameterization and high computational cost of inference. Alternative methods formulate the road network extraction as a binary classification problem. For instance, in [5], the image is pre-processed via a series of wavelet based filter banks, a fuzzy inference algorithm is then applied to detect the roads. In [6], a variety of network structures with different iteration times are used to determine the best network structure. In [8], a deep belief network is trained to detect image patches containing roads based on massive amounts of training data. A recent work [9] proposes a higher-order CRF model to capture long-range structures such as roads: the prior is represented by higher-order cliques that connect superpixels along straight line segments. It is a supervised method as it relies on some training databases.

In this paper, we are interested in detecting roads from a collection of images. To this end, we adopt a novel point of view: under a fully unsupervised framework, we pose such joint road extraction as a joint binary labelling problem on a multi-image graph of superpixels. The first contribution concerns the lack of supervision as a pairwise CRF model is defined on a set of test images to encode prior assumptions about the roads (e.g. thinness, elongation). The second novelty is related to the considered objective function consisting of data-driven competitive energy terms suitably defined to tackle with the challenging problem of road network extraction. To the best of our knowledge, there is no reported work on co-segmentation of roads that simultaneously exploits the geometric structures (elongation and symmetry) and, the rich modelling possibilities of CRF potentials.

The paper is organized as follows. Sec. 2 and 3 present respectively the image set representation in terms of a multi-

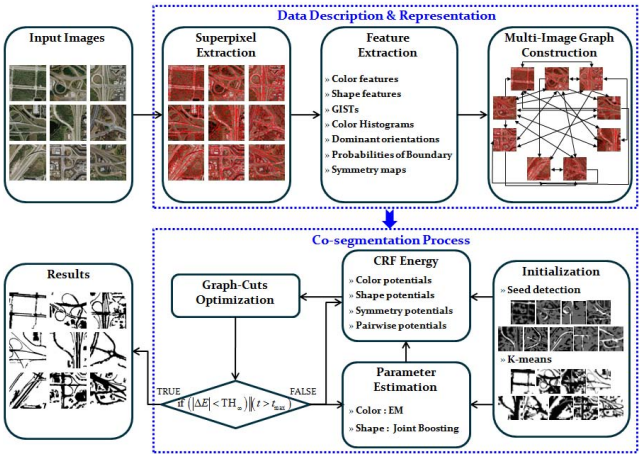

Fig. 1. The flowchart of our iterative method which relies on two steps: (1) Data description and representation, (2) co-segmentation process. In the first step, features are extracted and the multi-image graph is built. Each arrow indicates a dependency of an image on its neighboring image. In the second step, an iterative co-segmentation process is applied to extract jointly road networks.

image graph and the joint CRF model built over it (see the flowchart in Fig. 1). Sec. 4 gives experimental results on high resolution satellite images. Conclusions and future work are presented in Sec. 5.

2. IMAGE SET REPRESENTATION

Let $\mathcal{I} = \{I^1, \ldots, I^M\}$ denote a set of high resolution, RGB satellite images containing road networks. Rather than working with individual pixels, each image is over-segmented into small, regular superpixels by using the entropy rate algorithm which is known to preserve image structures [10]. We use superpixels for practical reasons. On the one hand, they yield more meaningful representation than do pixels because of their larger support, on the other hand, they are expected to speed up inference processing. The goal is to assign each superpixel either to the road class or to the background one. To this end, we first extract local and global features from images. Then, we construct a multi-image graph connecting superpixels based on their appearance similarity and spatial relationships. This graph represents *intra-image* as well as *inter-image* dependencies between superpixels.

2.1. Local/global features

Concerning local features, we compute a set of 3 color and 4 shape features to describe respectively the visual appearance and the geometric structure of each superpixel. More precisely, color is represented by the means of the Lab-color at each superpixel. Four shape features are also considered namely the extent e_x, the aspect ratio a_r, the circularity c_i and the convexity c_o scores defined as follows:

$$e_x = \frac{\mathcal{A}}{\mathcal{A}_B}, \quad a_r = \frac{\ell}{L}, \quad c_i = \frac{4\pi\mathcal{A}}{\mathcal{P}^2}, \quad c_o = \frac{\mathcal{A}}{\mathcal{A}_c} \qquad (1)$$

where $\mathcal{A}$, $\mathcal{A}_B$, ℓ, L, $\mathcal{P}$ and $\mathcal{A}_c$ denote respectively the superpixel area, the area of its bounding box, the lengths of minor and major axes of the ellipse that has the same normalized second central moments as the superpixel, its perimeter and its convex hull area. These simple shape features serve as measures of how circular or elongated the superpixels are. Structures of road superpixels should be elongated whilst other non-road superpixels are more circular. Stacking all the features yields a 7-dimensional descriptor $\mathcal{D}_i^m$ for each superpixel x_i^m in each image I^m. Then, we evaluate the similarity between any couple of superpixels $(x_i^m, x_j^{m'})$ by the squared Mahalanobis distance to account for the correlations between local features:

$$D\left(x_i^m, x_j^{m'}\right)^t = \left(\mathcal{D}_i^m - \mathcal{D}_j^{m'}\right)^t \Sigma^{-1} \left(\mathcal{D}_i^m - \mathcal{D}_j^{m'}\right) \qquad (2)$$

where Σ is the covariance matrix estimated on the image set. As global features, for every I^m in $\mathcal{I}$, we extract the GIST descriptor G^m [11] and, the 3D color histogram $\mathcal{H}_c^m$. The GIST descriptor represents the dominant spatial structure of a scene and has recently received increasing attention in the context of scene recognition. The similarity between images I^m and $I^{m'}$ is assessed by the sum of the squared Euclidean and chi-squared distances, the latter being widely employed to measure the similarity between normalized histograms:

$$K\left(I^m, I^{m'}\right) = \|G^m - G^{m'}\|^2 + \chi^2\left(\mathcal{H}_c^m, \mathcal{H}_c^{m'}\right). \qquad (3)$$

2.2. Multi-image graph

Considering the spatial consistency and appearance similarity between superpixels, we construct a multi-image graph $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$. The set of vertices $\mathcal{V} = \{x_i^m\}_{i\in[1,N_m],m\in[1,M]}$ consists of $\sum_{m=1}^{M} N_m$ superpixels obtained from all images where N_m denotes the number of superpixels in I^m. The set $\mathcal{E}$ of connections between superpixels capture their appearance similarity and spatial relationships. More precisely, each superpixel x_i^m is connected to its adjacent neighbors $\mathcal{N}_i^m$ to encode spatial dependencies, as for a single-image segmentation [12, 13]. In order to incorporate long-range dependencies and exploit inter-image information, we should ideally connect all the superpixels between all the images, but this would produce an overly complex model. For the sake of simplicity, we only keep similar superpixels connections. More precisely, we first find for each image I^m its $q \simeq M/2$ most similar images according to the image similarity metric K. Noting that we can set q equal to a given small constant for a large image set. Furthermore, we retain for each superpixel its $p = 10$ most similar superpixels in one image to keep a balanced contribution of the q images. Finally, each superpixel

$\{x_i^m\}$ is connected to only $k = p \times q/2$ most similar superpixels in these q images, according to the superpixel similarity metric D. It is worth noting that this graphical model construction is inspired from [14] where a multi-image model was first used to jointly estimate the labels of superpixels over the weakly supervised training set as well as the parameters of appearance models for the semantic classes. Then, this advanced model was applied to label superpixels of *one* new test image. In contrast, in our case, we address the problem of recovering the labels of superpixels in *multiple* test images without any training images. Compared to [15], we avoid dense correspondences between pixels by measuring similarities between larger region supports as superpixels.

3. JOINT ROAD NETWORK EXTRACTION FORMULATION

We propose a new probabilistic framework for the joint extraction of roads, *i.e.* symmetric elongated objects surrounded by a dominant background. We define a suitable pairwise CRF model over the resulting multi-image graph $\mathcal{G}$. Our goal is to compute the binary masks $\mathcal{B} = \{\mathbf{b}^1, \ldots, \mathbf{b}^M\}$ where $b_i^m = 1$ indicates road, and $b_i^m = 0$ indicates background at superpixel x_i^m in I^m. Likewise [14, 15], the optimal labelling corresponds to the minimization of the following global energy function E which defines this CRF model over $\mathcal{I}$:

$$\begin{aligned} E(\mathcal{B}) &= \sum_{m=1}^{M} \sum_{x_i^m \in I^m} [\Phi^m(b_i^m) + \lambda_{\text{int}} \sum_{x_j^m \in \mathcal{N}_i^m} \Psi_{\text{int}}^m(b_i^m, b_j^m) \\ &+ \lambda_{\text{ext}} \sum_{x_j^{m'} \in \mathcal{K}_i^m} \Psi_{\text{ext}}^{mm'}\left(b_i^m, b_j^{m'}\right)], \end{aligned} \tag{4}$$

where the Φ^m, Ψ_{int}^m and $\Psi_{\text{ext}}^{mm'}$ are potentials, $\mathcal{N}_i^m$ are the spatial neighbors of superpixel x_i^m in I^m and, $\mathcal{K}_i^m$ are its k nearest neighbors in other images. The unary potential Φ^m reflects the visual properties of the superpixel x_i^m and represents its likelihood to be road or background. Since visual appearance is often ambiguous at the superpixel level, we propose to regularize the objective function E by two additional pairwise potentials Ψ_{int}^m and $\Psi_{\text{ext}}^{mm'}$. Note that pairwise potentials have also been used in other tasks such as the segmentation of textured images [16]. In our case, Ψ_{int}^m aims at strengthening a consistent labelling between adjacent superpixels in the same image I^m whereas $\Psi_{\text{ext}}^{mm'}$ makes the labelling to be consistent between images I^m and $I^{m'}$ by encouraging connected superpixels to take the same label.

3.1. Unary potentials

The unary potential measures how well the local apparence of a superpixel x_i^m matches the label b_i^m. We define it as a linear combination of three terms:

$$\begin{aligned} \Phi^m(b_i^m) = \quad & -\lambda_C \log p\left(b_i^m; \mathcal{C}_i^m, \Theta^C\right) \\ & -\lambda_S \log p\left(b_i^m; \mathcal{S}_i^m, \Theta^S\right) + \lambda_{S_y} \Phi_{S_y}^m(b_i^m) \end{aligned}$$

where $p\left(b_i^m; \mathcal{C}_i^m, \Theta^C\right)$ is the posterior probability of x_i^m belonging to the class b_i^m given its color descriptor $\mathcal{C}_i^m$, according to the common color appearance model of this class. This model is shared among all images to account for the wide intra-class appearance variability. More precisely, we model each class road/background using a full-covariance Gaussian Mixture Model (GMM) of $N_c = 3$ components in the color feature space as the GMM has been successfully applied to model color appearance of objects [17]. Hence, the **first term** can be easily derived. During the superpixel assignment, each color appearance model is used to assess how fits a superpixel to one class. During the class modelling, each color appearance model is updated by learning from the superpixels allocated to the class (see Subsec. 3.3).

The **second term** measures how well the shape appearance $\mathcal{S}_i^m$ of x_i^m matches b_i^m, according to the joint boosting classifier [18] parameterized by Θ^S. We take its outputting posterior probabilities $p\left(b_i^m; \mathcal{S}_i^m, \Theta^S\right)$ to define the second unary term. Note that the joint boosting classifier is trained online over shape descriptors of the road segments and background estimation of $\mathcal{I}$. Joint boosting explicitly learns to share features (*i.e.* weak classifiers) across classes. It is efficient for detecting object classes in cluttered scenes [18]. Indeed, many fewer features are needed to achieve a desired level of performance than if weak classifiers were learned independently. Besides, the features selected by independently trained classifiers are often specific to the object class whereas the features selected by the jointly trained classifiers are more generic features. Furthermore, constructing independent distributions for color and shape makes the model robust against challenging cases where the road segments and background have a similar appearance. When one of the cues (color, shape) is not enough discriminative, we rely on the other to prohibit one distribution to leak into the other.

Finally, the **third term** in (5) is introduced to reflect the symmetric aspect of roads. Given the probability of symmetry map $\mathcal{S}^m$ [19] for each image I^m, we can directly define:

$$\Phi_{S_y}^m(b_i^m) = \begin{cases} -\log \mathcal{S}^m(x_i^m) & b_i^m = 1 \\ \beta & b_i^m = 0 \end{cases} \tag{5}$$

where β is a constant parameter for adjusting the likelihood of background superpixels. Decreasing β makes every superpixel more likely to belong to the background, thus resulting a more accurate estimation of the road class. $\mathcal{S}^m$ is the outputting posterior probability of the symmetry axes detector, introduced in [19]. This detector focuses on ribbon-like structures, *i.e.* contours marking local and approximate reflection symmetry. As the symmetry cue is useful in road network extraction, as shown in Fig. 2, we exploit it in three ways:

$\mathcal{S}^m$ is first used as a location prior of road segments into the unary potential, detected center lines are considered as seeds at the initialization; and pairwise costs for the CRF model are finally adjusted according to $\mathcal{S}^m$ as it will be explained in the next subsections.

3.2. Pairwise potentials

The masks $\{\mathbf{b}^m\}$ should be spatially consistent within each image structure. Indeed, along homogenous road, feature vectors are very similar and close to the mean. Thus, we use a Gaussian kernel $\exp -D(x_i^m, x_j^m)$ defined over similarity measures D to satisfy this smoothness constraint. Road segments have also linear structure with limited and slowly varying curvature. We use here again a Gaussian kernel $\exp -|\varphi_i^m - \varphi_j^m|^2/\sigma^2$ to enforce curvature consistency of road segments. $\Delta\varphi_{ij}^m = |\varphi_i^m - \varphi_j^m|^2$ is the squared dominant orientation deviation between adjacent superpixels. Furthermore, we integrate the average of probabilities of boundary Pb_{ij}^m along the boundary between adjacent superpixels to encourage segmentation aligned with the images gradients. As aforementioned, we enforce pairwise costs for superpixels belonging at symmetric axes to have the same label. By combining all the considered pairwise costs, an *intra-image* pairwise potential $\Psi_{\text{int}}^m\big(b_i^m, b_j^m\big) = [b_i^m \neq b_j^m] g\left(x_i^m, x_j^m\right)$, defined between adjacent superpixels x_i^m, x_j^m in I^m is obtained:

$$g\left(x_i^m, x_j^m\right) = \begin{cases} \lambda_r & \text{if center line} \\ \overline{\mathrm{Pb}}_{ij}^m \exp \frac{-D\left(x_i^m, x_j^m\right)|\varphi_i^m - \varphi_j^m|^2}{\sigma^2} & \text{otherwise.} \end{cases} \tag{6}$$

where $[b_i^m \neq b_j^m] = 1$ if $b_i^m \neq b_j^m$ otherwise 0, $\overline{\mathrm{Pb}}_{ij}^m = 1 - \mathrm{Pb}_{ij}^m$, and λ_r is a constant parameter for adjusting the contribution of center line cliques. Hence, we reduce the over-smoothing of thin structures as road segments.

The *inter-image* smoothness term is defined to promote *connected* superpixels $\{x_i^m, x_j^{m'}\}$ to take the same label if their appearance similarity is high:

$$\Psi_{\text{ext}}^{mm'}\left(b_i^m, b_j^{m'}\right) = [b_i^m \neq b_j^{m'}] \exp -D(x_i^m, x_j^{m'}). \tag{7}$$

This *inter-image* pairwise potential allows to strengthen the consistency of the labelling between images.

3.3. Optimization

Minimizing the whole energy E in (4) reduces to maximize the joint posterior probability of superpixel labels and appearance model parameters Θ, given the observed features. As in [15], we resort to an iterative algorithm that alternates between model parameter estimation and binary mask $\mathcal{B}$ estimation until convergence. Note that for given parameters Θ, E is submodular and so can be efficiently minimized by the graph-cuts algorithm [20]. In contrast, when the labelling $\mathcal{B}$ is fixed, the GMM parameters can be estimated using a standard EM algorithm. The later alternates between performing an expectation step, which creates a function for the expectation of the log-likelihood evaluated using the current estimate for the parameters, and a maximization step, which computes parameters maximizing the expected log-likelihood found on the expectation step. We initialize Θ by exploiting the detected center lines from the symmetry map $\mathcal{S}^m$ to extract automatically road/background seeds (Fig. 2). With these initial seeds, initial labelling is obtained by the k-means algorithm.

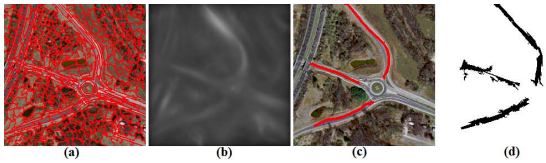

Fig. 2. (a) Superpixels. (b) Symmetry map. (c) Detected center lines. (d) Detected road seeds.

4. EXPERIMENTAL RESULTS

We validate our approach on high resolution (1m/pixel) satellite images of developed suburban scenes. Figure 3 illustrates road extraction results obtained by both independent (Ind-CRF for $M = 1$) and Joint CRF (JCRF for $M > 1$) models. Both models perceptually produce higher quality segmentations than the baseline k-means since they exploit contextual interactions and more flexible representation of prior knowledge. It is also observed that the joint model outperforms significantly the independent model due to the sharing of information between images. It is important to note that our intermediate models achieves good performance without any post-processing scheme, e.g. filling gaps between detected road segments and removing small non-road regions. In most of the literature, the network structure of roads is introduced only after detection with a heuristic post-processing scheme. Indeed, we compare the proposed method with two state-of-the-art techniques, those of Tuncer [5] and Mokhtarzade et al. [6]. We have discarded the method described in [7] because it includes a heuristic post-processing scheme. It can be observed that the results of our method, given in Fig. 4 are better than those approaches and are quite close to the ground truth. Overall, the joint model clearly extracts the road network most faithfully.

5. CONCLUSION

In this paper, we have formulated a powerful CRF model for jointly extracting road networks from a set of high resolution satellite images. The proposed method performs promising segmentations due to suitable unary and pairwise potentials. In future work, we plan to incorporate additional higher-order potentials in order to capture the long-range structure of roads and thus to fill gaps between road segments. We also plan to extend the data description by other discriminative cues to

deal with urban scenes, where roads are affected by occlusions, shadows, cars, etc.

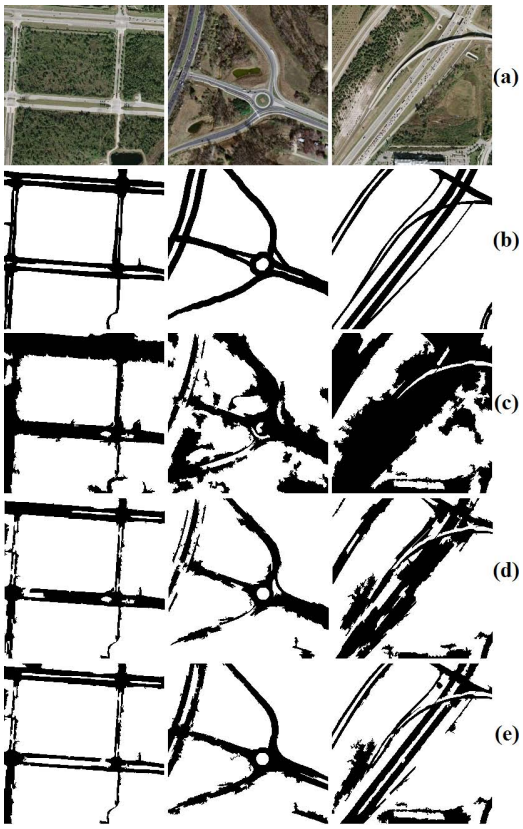

Fig. 3. (a) Three satellite images of developed suburban scenes. (b) Hand-drawn road maps. (c), (d), (e) Results of respectively k-means, IndCRF and JCRF ($M = 9$) methods.

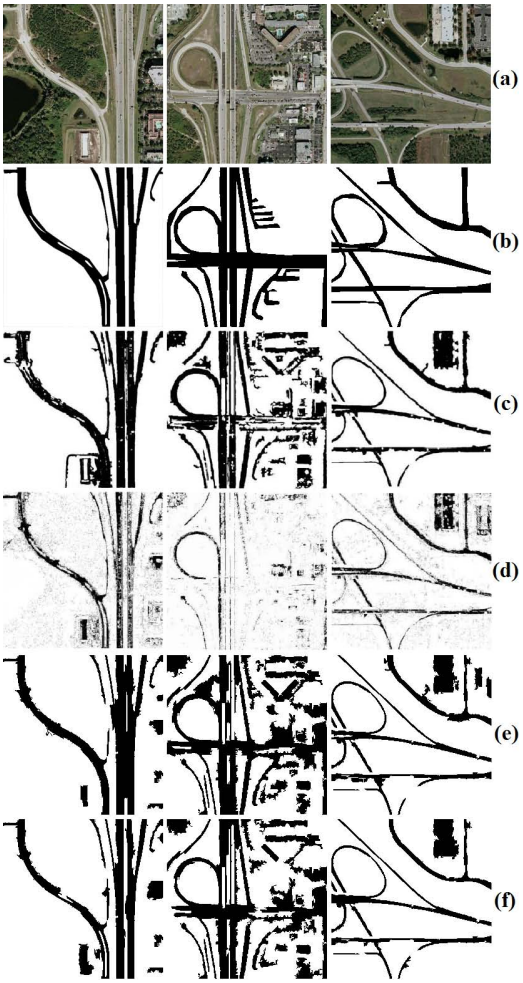

Fig. 4. (a) Three satellite images of developed suburban scenes. (b) Hand-drawn road maps. (c), (d), (e), (f) Results of respectively [5], [6], IndCRF and JCRF ($M = 9$) methods.

REFERENCES

[1] J. Trinder and Y. Wang, "Knowledge-based road interpretation in aerial images," *Int. Arch. Photogramm. Remote Sens.*, vol. 32, no. 4, pp. 635–640, 1998.

[2] C. Zhu, W. Shi, M. Pesaresi, L. Liu, X. Chen, and B. King, "The recognition of road network from high-resolution satellite remotely sensed data using image morphological characteristics," *IJRS*, vol. 26, no. 24, pp. 5493–5508, 2005.

[3] I. Laptev, H. Mayer, T. Lindeberg, W. Eckstein, C. Steger, and A. Baumgartner, "Automatic extraction of roads from aerial images based on scale space and snakes," *MVA*, vol. 12, no. 1, pp. 23–31, 2000.

[4] C. Lacoste, X. Descombes, and J. Zerubia, "Point processes for unsupervised line network extraction in remote sensing," *IEEE PAMI*, vol. 27, no. 10, pp. 1568–1579, 2005.

[5] O. Tuncer, "Fully automatic road network extraction from satellite images," in *3rd RAST*, 2007, pp. 708–714.

[6] M. Mokhtarzade, M. Javad, and V. Zoej, "Road detection from high-resolution satellite images using artificial neural networks," *Int. J. Applied Earth Observation and Geoinformation*, vol. 9, no. 1, pp. 32–40, 2007.

[7] S. Das, T. T. Mirnalinee, and K. Koshy Varghese, "Use of salient features for the design of a multistage framework to extract roads from high-resolution multispectral satellite images," *IEEE TGRS*, vol. 49, no. 10, pp. 3906–3931, 2011.

[8] V. Mnih and G. E. E. Hinton, "Learning to detect roads in high-resolution aerial images," in *ECCV*, 2010, pp. 210–223.

[9] J. D. Wegner, J. A. Montoya-Zegarra, and K. Schindler, "A higher-order crf model for road network extraction," in *CVPR*, 2013, pp. 1698–1705.

[10] M. Y. Liu, O. Tuzel, S. Ramalingam, and R. Chellappa, "Entropy rate superpixel segmentation," in *CVPR*, 2011, pp. 2097–2104.

[11] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *IJCV*, vol. 42, no. 3, pp. 145–175, 2001.

[12] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context," *IJCV*, vol. 81, no. 1, pp. 2–23, 2009.

[13] A. Bogdan, T. Deselaers, and V. Ferrari, "Classcut for unsupervised class segmentation," in *ECCV*. 2010, pp. 380–393, Springer.

[14] A. Vezhnevets, V. Ferrari, and J. M. Buhmann, "Weakly supervised structured output learning for semantic segmentation," in *CVPR*, 2012, pp. 845–852.

[15] M. Rubinstein, A. Joulin, J. Kopf, and C. Liu, "Unsupervised joint object discovery and segmentation in internet images," in *CVPR*, 2013, pp. 1939–1946.

[16] W. Pieczynski and Tebbache A. N., "Pairwise markov random fileds and segmentation of textured images," *MGV*, vol. 9, no. 3, pp. 705–718, 2000.

[17] S. Vicente, V. Kolmogorov, and C. Rother, "Cosegmentation revisited: Models and optimization," in *ECCV (2)*, 2010, pp. 465–479.

[18] A. Torralba, K.P Murphy, and W.T. Freeman, "Sharing visual features for multiclass and multiview object detection," *IEEE PAMI*, vol. 29, no. 5, pp. 854–869, 2007.

[19] S. Tsogkas and I. Kokkinos, "Learning-based symmetry detection in natural images," in *ECCV*, 2012, pp. 41–54.

[20] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision," *IEEE PAMI*, vol. 26, no. 9, pp. 1124–1137, 2004.