ROAD NETWORK EXTRACTION BY A HIGHER-ORDER CRF MODEL BUILT ON CENTERLINE CLIQUES

O. Besbes

COSIM Lab., SUP'COM, Carthage Univ.,
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

The goal of this work is to recover road networks from aerial images. This problem is extremely challenging because roads not only exhibit a highly varying appearance but also are usually occluded by nearby objects. Most importantly, roads are complex structures as they form connected networks of segments with slowly changing width and curvature. As an effective tool for their extraction, we propose to resort to a Conditional Random Field (CRF) model. Our contribution consists in representing the prior on the complex structure of the roads by higher-order potentials defined over centerline cliques. Robust P^N-Potts potentials are defined over such relevant cliques as well as over background cliques to integrate long-range constraints within the objective model energy. The optimal solution is derived thanks to graph-cuts tools. We demonstrate promising results and make qualitative and quantitative comparisons to the state of the art methods on the Vaihingen database.

Index Terms— Road network, higher-order CRF, centerline cliques, graph-cuts, aerial images.

1. INTRODUCTION

Automatic network extraction has a wide application in remote sensing, medical imaging and computer vision. In particular, road network extraction in urban environments is notably a challenging problem because of not only the wide intra-class appearance variations, but also the strong occlusions due to nearby objects such as buildings, trees and their related shadows. Moreover, many background objects have road-like appearance as they are made of similar materials (e.g. concrete roofs). As a consequence, the detector may be prone to false positive misclassification and, it could risk to miss many network connections because they are broken by many gaps. Most existing methods [1–5] derive model road networks from ad-hoc rules and, most often, they rely on bottom-up processes to extract them. Such methods can perform well for rural and suburban areas where the background is relatively homogeneous with few shadows and occlusions. For more complex environments such as the urban ones, a large number of methods [6–8] recover a *complete* network structure only after detection through a post-processing. For instance, in [6, 8], the post-processing consists in firstly removing false segments, then a region linking is performed to eliminate the discontinuities between road segments. Similarly, in [7], a deep belief network is trained to fill small gaps, and improves the quality of the pixel classification. An alternative strategy consists in resorting to probabilistic models. In this respect, the challenge is to include the prior information about roads. Indeed, road segments are thin linear structures with smoothly changing curvature, and road segments connect each others at junctions and crossings. To the best of our knowledge, few reported works have accounted for such prior. In [9–11], powerful object-based probabilistic representations based on Marked Point Processes (MPPs) integrate priors on the connectivity and the intersection geometry of roads. Although MPPs are a powerful tool to impose high-level topological constraints, the inference has a high computational cost as it relies on Markov Chain Monte Carlo samplers or simulated annealing type methods. In [12], minimum cost paths are used to connect seeds with high foreground scores and recover the network structure of roads. Promising results are obtained on curvilinear structures. However, the imposed geometric constraints are still relatively local since they bear on consecutive edge pairs. Recently, in [13, 14] a higher-order Conditional Random Field (CRF) formulation was proposed for road network extraction. The structural prior is represented by long-range cliques with robust P^N-Potts potentials [15]. More precisely, in [13], higher-order cliques are either in form of elongated straight segments or T and Y junctions. Straight segments are obtained by connecting randomly two nodes whereas junctions are obtained by connecting randomly central nodes to three additional ones such that the sampled nodes have sufficiently high road likelihoods. However, too many irrelevant cliques are generated since a straight segment connecting randomly road nodes does not coincide mostly with a road segment. To overcome this drawback, minimum-cost paths

978-0-9928626-3-3/15/$31.00 ©2015 IEEE

are embedded in a higher-order CRF framework in order to construct an explicit prior about the shape of roads [14]. Nevertheless, important properties of the road network (e.g. T-junctions and crossings) that could help to obtain a complete road network [11] are still missing.

In this work, we extend the baseline CRF models reported in [13, 14] by means of two main contributions. Firstly, centerline-driven road cliques are constructed in order to encompass different possible configurations of road segments, *i.e.* straight segments, blobs and junctions. Secondly, robust P^N-Potts potentials, defined over both road cliques and large background cliques, are computed using a boosted dense feature histogram-based classifier. Thereby, the label consistency is enforced by the integration of contextual priors and feature distributions. The figure 1 shows the block diagram of the proposed method. The paper is organized as follows. In Sec. 2, we describe the new higher-order CRF model built over these relevant cliques for road network extraction. In Sec. 3, we provide both qualitative and quantitative evaluations on the Vaihingen database of aerial images. Conclusions and future work are presented in Sec. 4.

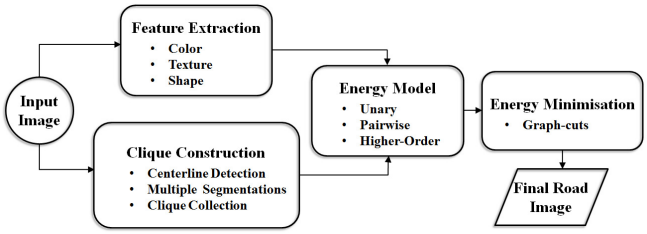

Fig. 1. The block diagram of the proposed method.

2. HIGHER-ORDER CRF MODEL FOR ROAD NETWORK EXTRACTION

We pose road network extraction problem as a pixel-wise labeling with two classes "road" and "background". We aim at defining a higher-order CRF model whose potentials incorporate appearance, shape, and context information as well as long-range contextual interactions to extract efficiently road networks. More precisely, this model reflects the prior assumptions about the roads, as they correspond to connected networks of smooth, elongated segments which meet at junctions and crossing. Likewise [15], the optimal labeling $\mathbf{x} = \{x_i\}$ of pixels i should minimize the Gibbs energy E, defined as a weighted sum of unary, pairwise and higher-order potentials:

$$E(\mathbf{x}) = \lambda_u \sum_{i \in \mathcal{V}} \psi_i(x_i) + \lambda_p \sum_{(i,j) \in \varepsilon} \psi_{ij}(x_i, x_j) + \lambda_h \sum_{c \in \mathcal{C}} \psi_c(\mathbf{x}_c) \quad (1)$$

where $\mathcal{V}$ corresponds to the set of all pixels, ε is the set of all edges connecting the neighboring pixels and, $\mathcal{C}$ refers to the set of cliques. The unary potentials ψ_i of the random field capture the local visual appearance of pixels, while the pairwise potentials ψ_{ij} encode a smoothness prior over neighboring variables. As regards the higher-order potentials ψ_c, they penalize inconsistent labeling by considering high-level dependencies between the clique variables. In particular, by means of long elongated centerline cliques, the higher-order potentials ensure the *continuity* of road networks and thus cut down gaps between road segments.

2.1. Features

It is well known that the raw RGB values of pixels alone are not very discriminant and, fail to produce an accurate object-class segmentation. To this end, several well-engineered features were defined for semantic segmentation of objects in images [16]. In order to ensure an efficient recognition, sophisticated potential functions are defined based on color, texture, location and shape features [14, 15, 17, 18]. In our work, we adopt the multi-feature extension [18] of TextonBoost [17]. A set of appearance features (Lab-color, textons [19], local binary patterns [20], multi-scale SIFT [21] and opponent SIFT [22]) is densely extracted. All features except Lab-color are quantized to 150 clusters using standard K-means clustering to construct the visual words.

2.2. Unary potentials

The unary potentials ψ_i measure how well the visual appearance of a pixel i matches a label x_i. Our model consists of respectively color and shape-texture unary potentials [17]:

$$\psi_i(x_i) = -\log(p_{\mathrm{C}}(x_i; y_i, \Theta_{\mathrm{C}})) - \log(p_{\mathrm{ST}}(x_i; \mathbf{y}, \Theta_{\mathrm{ST}})) \quad (2)$$

where p_{C} (resp. p_{ST}) is the probability that label x_i is assigned to pixel i given its color feature y_i (resp. the extracted dense multiple features $\mathbf{y}$). Theses probabilities are outputted by the learned classifiers of parameters Θ_{C} and Θ_{ST} respectively. The unary color potentials capture the color distribution of object classes by means of Gaussian Mixture Models in the Lab color space. The unary shape-texture potentials reflect the shape, texture and appearance context of each class. They are estimated by boosting weak classifiers based on a set of shape filter responses. These dense feature filters are defined by triplets [feature type f, feature cluster t, rectangular region r] and their feature response $v^f_{[t,r]}(i)$ for a given pixel i. The latter measures the number of features of type f belonging to cluster t in the region r associated to i. The pool of weak classifiers contains a comparison of responses of dense-feature shape filters against a set of thresholds. Further details about this procedure could be found in [17]. At this level, it is worth pointing out that we have considered independent distributions for color and shape-texture features in order to achieve robust model in challenging cases where the road segments and background have a similar appearance.

In this way, if one of the cues (color, shape-texture) is not enough discriminant, we rely on the other one to prohibit one distribution to leak into the other one.

2.3. Pairwise potentials

The pairwise potentials ψ_{ij} encode a smoothness prior which promotes neighboring pixels to share the same label. They have the form of a contrast sensitive Potts model [15, 17]:

$$\psi_{ij}(x_i, x_j) = \begin{cases} 0 & \text{if } x_i = x_j, \\ g(x_i, x_j) & \text{otherwise,} \end{cases} \quad (3)$$

where $g(x_i, x_j) = \exp\left(-\beta \|y_i - y_j\|^2\right)$ is an edge feature based on the difference in colors of neighboring pixels. The parameter $\beta = \left(2\langle \|y_i - y_j\|^2 \rangle\right)^{-1}$ is an image-dependent contrast term, where $\langle . \rangle$ denotes an average over the image.

2.4. Higher-order potentials

The higher-order potentials ψ_c promote all pixels of a clique to have the same label. In our work, we have focused on a robust P^N-Potts model [15] defined by:

$$\psi(\mathbf{x}_c) = \min_{l \in \{0,1\}} \left(\gamma_c^{\max}, \gamma_c^l + \sum_{i \in c} k_c^l \left[x_i \neq l \right] \right) \quad (4)$$

where $[x_i \neq l] = 1$ if true and 0 otherwise. For a clique $c \in \mathcal{C}$, the potential has a cost of $\gamma_c^l \leq \gamma_c^{\max}$ if all the pixels in the clique take the label l. However, each pixel whose label is different of l is penalized by an additional cost of k_c^l. The maximum cost of the potential is truncated to $\gamma_c^{\max}$. For instance, if in one hand, there is enough road evidence within a clique, then its member pixels are assigned to the road class. Therefore, gaps caused by false negatives within a clique are corrected. On the other hand, if there is not enough total road evidence in a clique, then the related pixels are assigned to the background. As a result, false positive are removed. In [13, 14], the costs $\{\gamma_c^{\max}, \gamma_c^l, k_c^l\}$ are assumed constants for labels $l \in \{0, 1\}$ and for all the cliques $c \in \mathcal{C}$. In this case, the higher-order potentials encourage all the variables in a clique c to take the same *dominant* label, while omitting the clique appearance. Nevertheless, it is well known that the distributions of pixel-wise feature responses are very useful for discriminating strongly object-classes. Thereby, motivated by [18], we learn a binary boosting classifier over the normalized histograms of multiple clustered pixel-wise features $\{\mathcal{H}_t^f\}$, to capture the clique appearances for both road and background classes. The negative log-likelihood of the classifier is incorporated into the energy as:

$$\gamma_c^l = -\log\left(p\left(\mathbf{x}_c = l; \{\mathcal{H}_t^f(c)\}\right)\right)|c|, \quad (5)$$

$$\gamma_c^{\max} = |c|\alpha^h, k_c^l = \frac{\gamma_c^{\max} - \gamma_c^l}{0.1|c|}, \quad (6)$$

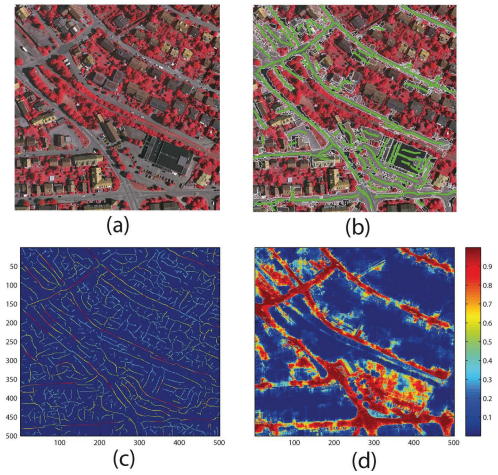

Fig. 2. (a) an aerial image, (b) dectection centerlines and samples of higher-order cliques, (c) the detector's symmetry reponse, and (d) the learned classifier's likelihood.

where α^h is a truncation parameter and, $|c|$ denotes the cardinality of the clique c.

2.5. Clique construction

The set $\mathcal{C}$ of higher-order cliques should contain *relevant* cliques for higher-order potentials. On the one hand, they should be shaped, for the road class, like long elongated sets of pixels, putted usually together in junctions. On the other hand, they should form large regions, as far as possible, for the background. To this end, we detect for an image the centerlines, and we perform multiple segmentation of it. The road and background cliques are then derived by accounting for both the centerlines and the segmentations.

• **Centerline detection**: For an aerial image, we extract firstly centerlines using the symmetry axes detector [23]. More precisely, this detector focuses on ribbon-like structures, *i.e.* contours marking local and approximate reflection symmetry. In order to remove *non-road* ones, we apply a non-maximum suppression on the detector's symmetry response, which is weighted by the learned classifier's likelihood (Fig .2(c) − (d)). Then, we perform a hysteresis thresholding to extract the road centerlines. D-centerlines refer to these detected centerlines (Fig .2(b)).

Generally, many gaps among them appear because of the occlusions and missing detections. However, the main characteristic of the roads is their network structure: a road segment is usually connected to other road segments on both sides, sometimes connected only on one side, but almost never isolated. Accordingly, we resort to the Constrained Delaunay Triangulation (CDT) algorithm to *complete* D-centerlines across gaps, after their piece-wise linear approximation. At this stage, we obtain completion centerlines (called C-centerlines) filled in by the CDT. Finally, we dis-

card C-centerlines whose average road likelihood is below a threshold.

• **Generating multiple segmentations**: Multiple segmentations allow to obtain accurate segmentations of objects with thin structures, possibly the road segments [15]. Therefore, we produce for an image three segmentations in an unsupervised manner by using the mean-shift algorithm [24] by varying both spatial and range bandwidth parameters.

• **Collecting relevant cliques**: For each segmentation, we construct firstly road cliques such that the estimated D/C-centerlines are their medial-axis. More precisely, we perform morphological dilation on these centerlines with a disk. The disk radius was adjusted to 5 so as the narrowest roads in the tested images can be extracted. Then, we collect all superpixels of the segmented image, that overlap with the resulting ribbons. These superpixels constitute the road cliques at fine resolution. In order to correct the first stage's miss-detection, we add afterward all superpixels with sufficiently high road likelihoods. Concerning background cliques, we construct a binary mask such that a pixel is assigned 0 if it belongs to a *road* superpixel and 1 otherwise. Thus, the background cliques are the set of connected components in this binary mask. Thereby, we obtain a new segmentation which consists of both road superpixels and background cliques. Then, we derive a coarser segmentation by merging all superpixels explained by the same centerline segment, corresponding also to a CDT's edge. Finally, we merge centerline segments, constituting a junction in the binary C/D-centerlines mask. Figure 2(b) shows samples of constructed higher-order cliques. It is worth noting that due to such data-driven clique construction different road configurations are considered such as straight segments, blobs and various-type junctions.

3. EXPERIMENTAL RESULTS

Experiments were conducted on the urban aerial Vaihingen database [25] composed of 14 color infrared images of size 500×500 pixels and a ground resolution of 0.5 m: 4 images were used for training and 10 for testing. The Vaihingen road networks have irregular and complex structures. There are many short and narrow roads, with cast shadows and overhanging trees, making road extraction more challenging. The proposed approach is compared to 4 methods: the baseline pairwise CRF (PCRF), the baseline Higher-Order CRF (HOCRF) [15] defined on superpixels collected from three segmentations (generated by the second stage) and the two recent and competitive methods reported in [13, 14]. As performance metrics, we employ the commonly used metrics for evaluating road detection methods, namely *correctness*, *completeness* and *quality* scores [26]. Ground truth centerline pixels are considered true positives TP if they lie within a buffer of width B around the estimated centerline, and false negatives FN otherwise, so as a variant of recall $completeness = \frac{TP}{TP+FN}$. On the other hand, estimated centerline pixels are TP if they lie within B pixels of the ground truth centerline, and false positives (FP) otherwise, thus as a variant of precision $correctness = \frac{TP}{TP+FP}$. Finally, both criteria are combined into a quality metric according to $quality = \frac{TP}{TP+FP+FN}$. The buffer width is set to $B = 5$ pixels for all the considered methods, corresponding to the narrowest roads we wish to extract. Qualitative and quantitative results are reported respectively in Figure 3 and Table 1.

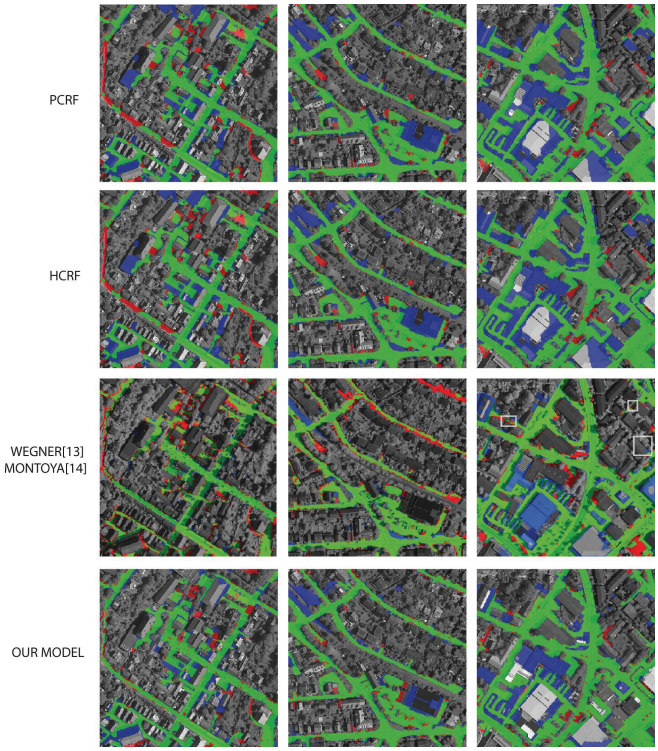

Fig. 3. Road networks extracted in 3 different aerial images. True positives are displayed green, false positives blue, and false negatives red.

Table 1. Detection performance of road extraction methods. All numbers are percentages.

	Completeness	*Correctness*	*Quality*
PCRF	68.5	76	56.1
HOCRF	68.8	76.4	56.4
WEGNER [13]	69.4	75.0	55.6
MONTOYA [14]	88.4	81.1	73.3
OUR MODEL	86.7	74.3	67.8

It can be noted that both the PCRF and HOCRF methods, despite they do not incorporate any prior about roads, have a higher quality compared to [13] thanks to the unary classfier's efficiency. The performance of HOCRF is little better than PCRF, due to the consistency prior over superpixels obtained from multiple segmentation. Our model outperforms PCRF, HCRF and [13] since we take into account the specific char-

acteristics of roads and incorporate them as priors by means of higher-order potentials. On the other hand, the recent method [14] outperforms all the four methods by dint of a more powerful unary classifier. In fact as reported in [14], the proposed multi-scale, contextual unary classifier provides: $completeness = 89.6$, $correctness = 79.8$ and $quality = 73$.

4. CONCLUSION

In this paper, we have proposed an efficient higher-order CRF model for road network extraction based on the robust P^N-Potts model. As future work, we plan to enhance the CRF's unary classifer by means of additional contextual features. Moreover, we plan to extend our model to tackle a multi-label classification problem with class-specific priors for other objects like buildings.

REFERENCES

[1] P. Doucette, P. Agouris, and A. Stefanidis, "Automated road extraction from high resolution multispectral imagery," *ASPRS PE&RS*, vol. 70, no. 12, pp. 1405–1416, 2004.

[2] J. B. Mena and J. A. Malpica, "An automatic method for road extraction in rural and semi-urban areas starting from high resolution satellite imagery," *PRL*, vol. 26, no. 9, pp. 1201–1220, 2005.

[3] C. Poullis and S. You, "Delineation and geometric modeling of road networks," *ISPRS P&RS*, vol. 65, no. 2, pp. 165–181, 2010.

[4] C. Ünsalan and B. Sirmaçek, "Road network detection using probabilistic and graph theoretical methods," *IEEE Trans. GRS*, vol. 50, no. 11, pp. 4441–4453, 2012.

[5] Z. Miao, W. Shi, H. Zhang, and X. Wang, "Road centerline extraction from high-resolution imagery based on shape features and multivariate adaptive regression splines," *IEEE GRSL*, vol. 10, no. 3, pp. 583–587, 2013.

[6] S. Hinz and A. Baumgartner, "Automatic extraction of urban road networks from multi-view aerial imagery," *ISPRS P&RS*, vol. 58, no. 1-2, pp. 83–98, 2003.

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

[8] 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 Trans. GRS*, vol. 49, no. 10, pp. 3906–3931, 2011.

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

[10] F. Florent Lafarge, Gimel'farb G. L., and X. Descombes, "Geometric feature extraction by a multimarked point process," *IEEE Trans. PAMI*, vol. 32, no. 9, pp. 1597–1609, 2010.

[11] D. Chai, W. Förstner, and F. Lafarge, "Recovering line-networks in images by junction-point processes," in *IEEE CVPR*, Portland, Oregon, 2013, pp. 1894–1901.

[12] E. Türetken, F. Benmansour, B. Andres, H. Pfister, and P. Fua, "Reconstructing loopy curvilinear structures using integer programming," in *IEEE CVPR*, Portland, Oregon, 2013, pp. 1822–1829.

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

[14] J.A. Montoya-Zegarra, J. D. Wegner, L. Ladick, and K. Schindler, "Mind the gap: Modeling local and global context in (road) networks," in *GCPR*, Münster, Germany, 2014, pp. 212–223.

[15] P. Kohli, L. Ladický, and P.H. Torr, "Robust higher order potentials for enforcing label consistency," *IJCV*, vol. 82, no. 3, pp. 302–324, 2009.

[16] I. Guyon, *Feature Extraction: Foundations and Applications*, Studies in Fuzziness and Soft Computing. Springer, 2006.

[17] J. Shotton, J. M. 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.

[18] L. Ladicky, C. Russell, P. Kohli, and P. H. S. Torr, "Associative hierarchical crfs for object class image segmentation," in *IEEE ICCV*, Kyoto, Japan, 2009, pp. 739–746.

[19] J. Malik, S. Belongie, T. Leung, and J. Shi, "Contour and texture analysis for image segmentation," *IJCV*, vol. 43, no. 1, pp. 7–27, 2001.

[20] T. Ojala, M. Pietikainen, and D. Harwood, "Performance evaluation of texture measures with classification based on kullback discrimination of distributions," in *ICPR*, Jerusalem, Israel, 1994, vol. 1, pp. 582–585.

[21] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, vol. 60, no. 2, pp. 91–110, 2004.

[22] K. E. A. Van de Sande, T. Gevers, and C. G. M. Snoek, "Evaluation of color descriptors for object and scene recognition," in *IEEE CVPR*, Alaska, USA, 2008, pp. 1–8.

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

[24] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Trans. PAMI*, vol. 24, no. 5, pp. 603–619, 2002.

[25] ISPRS benchmark, "The Vaihingen dataset," `http://www.itc.nl/ISPRS_WGIII4/tests_datasets.html`.

[26] C. Wiedemann, C. Heipke, and H. Mayer, "Empirical evaluation of automatically extracted road axes," in *CVPR Workshop EEMCV*, Santa Barbara, USA, 1998, pp. 172–187.