A higher-order CRF model for road network extraction

Jan D. Wegner, Javier A. Montoya-Zegarra, Konrad Schindler

Photogrammetry and Remote Sensing, ETH Zürich, Switzerland

Abstract

The aim of this work is to extract the road network from aerial images. What makes the problem challenging is the complex structure of the prior: roads form a connected network of smooth, thin segments which meet at junctions and crossings. This type of a-priori knowledge is more difficult to turn into a tractable model than standard smoothness or co-occurrence assumptions. We develop a novel CRF formulation for road labeling, in which the prior is represented by higher-order cliques that connect sets of superpixels along straight line segments. These long-range cliques have asymmetric P^N-potentials, which express a preference to assign all *rather than just* some *of their constituent superpixels to the road class. Thus, the road likelihood is amplified for thin chains of superpixels, while the CRF is still amenable to optimization with graph cuts. Since the number of such cliques of arbitrary length is huge, we furthermore propose a sampling scheme which concentrates on those cliques which are most relevant for the optimization. In experiments on two different databases the model significantly improves both the per-pixel accuracy and the topological correctness of the extracted roads, and outperforms both a simple smoothness prior and heuristic rule-based road completion.*

1. Introduction

The application problem behind this paper is the extraction of the road network from aerial or satellite images. This is a challenging vision problem with important applications in mapping and remote sensing. In spite of more than two decades of research [1, 7, 10, 18], the problem is largely unsolved—we are not aware of an operational system for automatic road extraction.

The most challenging part of the task is to get the network topology right: if existing connections are broken or inexistent ones hallucinated, then a road map is of little use for navigation, even if an overwhelming majority of pixels are correctly labeled as *road*, respectively *background*; on the other hand, once a correct and complete network of approximate centerlines has been recovered, the exact segmentation can be refined locally (*e.g.* with active contours [17]). We point out that this is an instance of a more general issue beyond road extraction. It exists in similar form for other image understanding tasks which involve objects with a "network" topology, *i.e.* they are made up of thin segments linked together by junctions and crossings (Fig. 1).

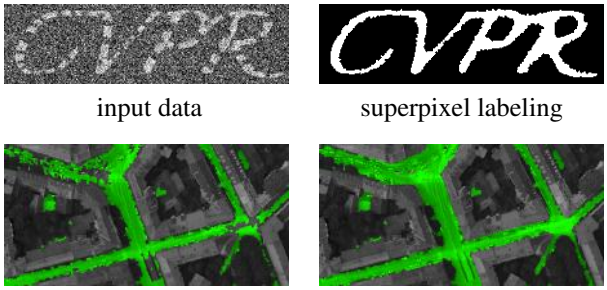


Figure 1. The proposed higher-order CRF favors networks of elongated segments *(top)*. We apply it to road extraction *(bottom)*.

A main difficulty of image understanding (*i.e.* semantic interpretation of the image content) is that the observation data is noisy, incomplete and ambiguous, such that prior knowledge about the layout of the observed scenes is necessary to obtain satisfactory results. As a consequence, a main focus of computer vision research over the past decade has been how to include such prior knowledge into the (usually probabilistic) models.

Maybe the simplest form of prior are expectations about an object's location, along the lines of "the sky is usually at the top". They are conditionally independent between different pixels and can directly be merged into the per-pixel likelihood, *e.g.* [21].

A lot more powerful models can be formulated by taking into account the relative location of objects, like for example "boats are usually found near water" [8] or "neighbouring regions tend to have the same semantic class" [2, 13]. Due to the constraints, the individual variables are no longer independent but form a conditional random field (CRF).[1]

[1] Or a Markov random field; for our purposes the distinction is irrelevant.

Inference in such models is more complicated, but often good solutions can still be found efficiently. Arguably, much progress in image understanding in the last decade is due to the fact that in CRFs with appropriately restricted clique potentials (approximate) MAP estimation is possible with variants of graph cuts [3] or message passing [6, 14].

However, for some object classes more complex priors are adequate, and these include our target class, the roads on the earth's surface. The characteristic feature of the roads is their network structure: road segments are thin linear structures with limited and smoothly changing curvature; and a road segment is usually connected to other road segments on both sides, sometimes connected only on one side, but almost never isolated. Note how these expectations differ qualitatively from standard smoothness assumptions: even a tiny minority of non-adjacent (super-)pixels with high road likelihood can be strong evidence for a road, if they are aligned along a straight line; and the evidence is even stronger if the first and last of the (super-)pixels lie on potential crossings.

In principle, it is of course possible to formalize all the desired constraints into a probabilistic model, and some research in that direction exists, *e.g*. [22, 15]. Unfortunately, the resulting likelihood functions tend not to be amenable to efficient inference algorithms. Solutions can only be found with Markov Chain Monte Carlo samplers or annealing-type methods, which are rather difficult to parameterise correctly and have high computational cost. In most of the literature, the network structure of roads is introduced only after detection, by filling gaps between detected road segments with heuristic rules (*c.f*. Sec. 2).

In the present paper, we explore the possibility to construct an intermediate model, which captures important properties of the road network while still being amenable to efficient inference techniques. The main contributions are *(i)* a formulation of the constraints as non-local higher-order cliques with asymmetric P^N-Potts potentials [13], such that they can be solved within the graph-cuts framework; and *(ii)* a data-driven sampling strategy to find the relevant cliques and make inference tractable. To our knowledge this is the first work which exploits the rich modeling possibilities of the P^N-Potts model for network modeling in general and for road extraction in particular.

2. Related work

There is an extensive body of work on road extraction, and we can only review a representative selection here. For a more complete overview please refer to the evaluation papers [10] (up to 1997) and *et al*. [18] (1998–2006).

Road detection in images goes back to at least [1], where road pixels are identified with a sequence of local image processing operations. Only shortly afterwards [7] was probably the first work to explicitly incorporate topology, by searching long 1-dimensional structures. A local road score is computed at each pixel with a line detector and roads are found iteratively as minimum cost paths with an A^*-type algorithm. In [17] road extraction is based on multi-scale line detection. A heuristic completion scheme is employed to bridge gaps due to shadows, overhanging trees *etc*. Subsequently the road segmentation is refined with a pair of coupled active contours ("twin-snakes"). Detecting oriented road segments also forms the basis of [12]. The most road-like of these segments are then designated as seeds and the network is iteratively grown from there. In a final step, the network is pruned with a shape-based classifier to remove false positives. In [22, 15] marked point processes (MPP) are introduced as representation for short road segments. MPPs provide a powerful framework to construct an object-based probabilistic representation and allows them to include elaborate priors on the connectivity and intersection geometry of roads. On the downside, inference with MPPs is only possible with all-purpose methods like simulated annealing or reversible jump Markov Chain Monte Carlo (RJMCMC), which are computationally very demanding and also brittle to set up, such that in practice they do not always achieve satisfactory results. In [19, 20] a deep belief network is trained to detect image patches containing roads. A second network is trained to take the output of the first one as input and fill small gaps. Using massive amounts of training data—extracted automatically with the help of existing road databases—they achieve promising results, on images with largely unoccluded roads.

The works mentioned so far have focused on rural and suburban areas, where the road network is relatively sparse and regular, and less affected by occlusions, shadows, cars *etc*. One of the few works on urban roads is [11]. Given high-resolution images and a height map, road segments are detected using multiple cues (dark homogeneous areas, valley lines of the height map, lane markings, vehicles). The segments are then connected by iteratively inserting potentially missing connections and verifying that they have sufficiently homogeneous brightness. Overall, little research exists on road extraction in dense urban scenes.

Road extraction has also been attempted from other data sources, *e.g*. [31] extract road center lines from range images generated with airborne laser scanning, and [25, 24] extract roads from synthetic aperture radar (SAR) imagery. Both approaches are surprisingly similar: detect oriented lines, link them to straight road segments, hypothesize additional segments to "fill the gaps" in the network with simple geometric rules, and select which of the hypotheses to keep by inference in a pairwise MRF over the segments.

3. CRF Model of the road network

We pose road extraction as a binary labeling problem on superpixels, linked together in a CRF which encodes the

prior assumptions about the roads. In the following we describe each component of the model.

3.1. Image representation and unaries

Rather than working with individual pixels, the raw image is over-segmented into small, regular superpixels, which are the atomic units for all further processing. We use the graphcut-based segmentation algorithm of [26].

While our method can in principle be extended to individual pixels, we prefer to use superpixels for practical reasons. On the one hand, they yield more stable unaries because of their larger support, on the other hand, they greatly speed up processing, both during clique generation (Sec. 3.4) and during inference. Their main disadvantage is that in certain cases they will lead to jagged and incorrect road boundaries. We are mainly interested in improved extraction of the network topology, and believe segmentation boundaries are best cleaned up in a subsequent step with a stronger shape prior (*e.g.* [17]).

The next step is to estimate, for each superpixel, the likelihood of being *road* respectively *background*. To that end we train a Random Forest classifier [4] with 20 trees. As features we extract standard color and texture features. In detail, we convert the image to opponent Gaussian color space [5] and convolve it with the 17-dimensional filter bank proposed by Winn *et al.* [29]. The filter bank consists of Gaussian kernels at three scales, first-order Gaussian derivatives in x and y at two scales, and LoG responses at four scales. The Gaussian kernels are evaluated for all three channel, whereas the derivative filters are evaluated only in the intensity image. The 34-dimensional feature vector for a superpixel is made up of the means and standard deviations of the individual filter channels.

3.2. Higher-order CRF model

CRFs have become a standard tool of computer vision to represent image priors. Initially researchers concentrated on first-order CRFs (*e.g.*, [23, 21, 8, 9]), mainly because efficient inference methods existed only for these. More recently it has turned out that the crucial property for a CRF to allow efficient inference is not so much the size of the cliques, but rather how many different states the clique potentials can take on: higher-order CRFs can still be solved efficiently, if the higher-order potentials are restricted appropriately [13].

Several recent works exploit the rich modeling possibilities of higher-order cliques for semantic pixel labeling, for example by introducing global co-occurrence statistics [16], or by simultaneously inferring the scene type of an image as well as the spatial extent, location, and class of objects [30]). Here, we adapt higher-order potentials to network extraction. Recall that our aim is to model the posterior distribution $P(\mathbf{y}|\mathbf{x})$ of labels $\mathbf{y}$ given $\mathbf{x}$. With a slight abuse of notation we denote both the raw data and the features derived from it $\mathbf{x}$ (respectively x_i for a particular superpixel). In a CRF the posterior is a Gibbs distribution, $P(\mathbf{y}|\mathbf{x}) \propto \exp(-E(\mathbf{x},\mathbf{y}))$. In our case the variables to be labeled are the set S of superpixels, and the label set is $y_i \in \{0,1\}$, where 1 denotes *road* and 0 *background*. Instead of only allowing unary and pairwise potentials, the Gibbs energy for a higher-order CRF is given by

$$E(\mathbf{x},\mathbf{y}) = \sum_{i\in S} \psi_i(x_i,y_i) + \sum_{c\in H} \psi_c(\mathbf{x}_c,\mathbf{y}_c) \ , \qquad (1)$$

where H denotes the set of cliques (note, for convenience of notation we also include possible pairwise cliques in H), ψ_i are the unaries, and ψ_c are the clique potentials that encode dependencies between the variables of a clique. MAP inference consists in maximizing $P(\mathbf{y}|\mathbf{x})$, which is the same as minimizing the energy $E(\mathbf{x},\mathbf{y})$.

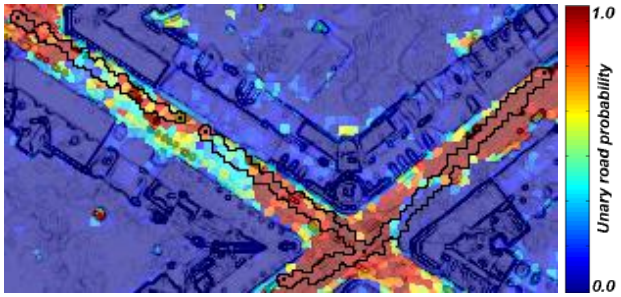


Figure 2. Cliques connect superpixels on straight line segments or 3-junctions. They are sampled by connecting superpixels with high *road* likelihood. (marked with black centroids).

The aim of our work is to extract roads, *i.e.* thin elongated objects surrounded by a dominant background. Obviously, a CRF with standard pairwise potentials will not be able to encode these long-range structures, but rather tend to smooth away thin structures such as roads, a well-documented phenomenon in image segmentation (*e.g.* [27], see also Sec. 4). Instead, we require a higher-order potential over long elongated sets of superpixels (Fig. 2), which encourages them to take on the *road* label *if the cumulative evidence over the entire clique is strong enough.*

Still many such cliques will also contain some *background* superpixels, thus the penalty for non-*road* labels in the clique should increase gracefully rather than abruptly with the first deviating superpixel. Furthermore, the prior is not symmetric: if the dominant label along an elongated clique is *background*, then one can not in general deduce that all superpixels should be labeled *background*, since in an urban or suburban environment any sufficiently long straight segment will intersect several roads, and contain several superpixels with a strong preference for that class.

The higher order potential $\psi_c(\mathbf{x}_c,\mathbf{y})$ that we propose has the following form:

$$\psi_c(\mathbf{x}_c, \mathbf{y}_c) = \begin{cases} \min\left(\alpha, P_b \cdot \frac{\alpha-\beta}{\gamma} + \beta\right) & \text{if} \quad P_b < P_r \\ 0 & \text{else} \end{cases} \quad (2)$$

with $P_r = \sum(w_c^i \cdot y_i)$ the weighted sum of *road* superpixels in the clique and $P_b = \sum(w_c^i \cdot (1 - y_i))$ the weighted sum of *background* superpixels. The w_c^i are weights that determine the influence of individual superpixels on the clique potential (*c.f*. Sec. 3.3). α is an upper bound on the potentials, and β, γ are the remaining parameters of a truncated linear cost function. Using a truncated linear function ensures the desired graceful increase of the penalty, while penalizing only *background* pixels in *road*-dominated cliques introduces the desired asymmetry.

The potential (2) is designed in such a way that it is a special case of the *robust P^N-Potts model*, a class of higher-order CRFs introduced in [13] whose energies can be minimized in low polynomial time with graph cuts.

3.3. Contrast-sensitive node weighting

In order to avoid over-smoothing, we use contrast sensitive node weights w_c^i. The intuition is the following: if a superpixel has high *background* likelihood and its features deviate a lot from the other ones in the clique, then it probably belongs to the background, *i.e*. labeling it as *background* should not have a large impact on the energy (for example, think of a small roundabout on a major road). Empirically we found the following weighting scheme to work best: we compute the mean feature vector $\overline{x}_c$ of the clique, and for each superpixel measure the deviation of its feature vector x_i from that mean, using the Euclidean distance $d_c^i = \|\overline{x}_c - x_i\|$ in feature space. With the standard deviation σ_c of all such distances in a clique, the weight is then determined as

$$w_c^i = \begin{cases} w_{max} & \text{if} \quad d_c^i < \sigma_c \\ w_{max}(2 - d_c^i/\sigma_c) & \text{if} \quad \sigma_c < d_c^i < 2\sigma_c \\ 0 & \text{if} \quad d_c^i > 2\sigma_c \end{cases} \quad (3)$$

Along rather homogeneous roads, many feature vectors are very similar and close to the mean, whereas superpixels on vegetation, building features *etc*. stand out. The truncated linear weighting function gives full weight to superpixels within one standard deviation σ_c, a linearly decreasing weight to pixels between σ_c and $2\sigma_c$, and zero weight to nodes outside two standard deviations—effectively removing them from the clique. The contrast sensitive weighting significantly reduces false positives.

3.4. Clique sampling

The remaining problem of the proposed model is that the number of potential cliques is huge. It is infeasible to consider all possible straight chains of superpixels in an image.

The basic insight is the following: most possible cliques will be irrelevant to the problem, because an overwhelming majority of their unaries will have a preference for the *background* class. Since cliques where the *background* label dominates have potential $\psi_c = 0$, such cliques will not change the energy in reasonable regions of the solution space near an energy minimum.

Furthermore, note that many cliques will have (up to the mild influence of the contrast-sensitive weights) the same potential value: consider a *road* segment with, say, two deviating superpixels labeled as *background*. By shifting the line segment along the road, slightly rotating it within the road surface, or shrinking/extending it one can generate a large number of cliques which also have dominant label *road*, the same two *background* members, and similar w_c^i because they are mostly supported by the same pixels.

We thus obtain a representative sample of the relevant cliques (*i.e*. those which could conceivably coincide with a road) in a data-driven fashion. To that end, we return to the definition adopted above: a road network consists of (nearly, locally) straight segments between junctions. Based on this prior we sample cliques either in form of elongated straight segments ("network cliques") or junctions ("junction cliques").

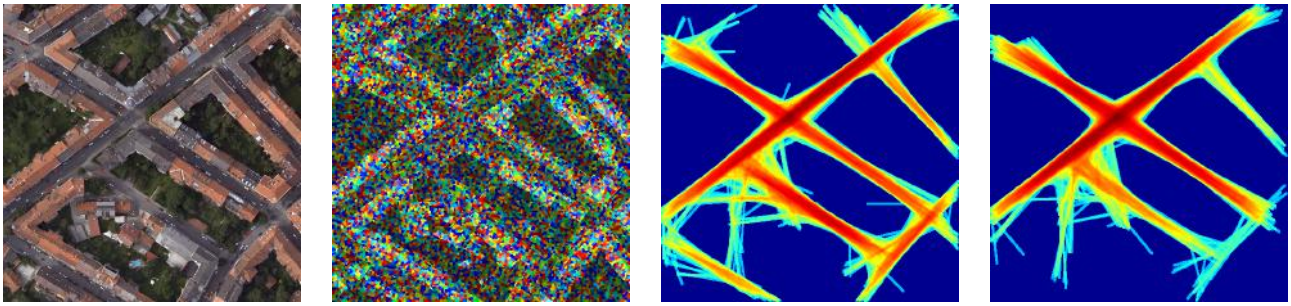
For the network cliques two seed nodes with sufficiently high *road* likelihoods are sampled randomly and connected with a straight corridor. Setting the width of the corridor to the mean super-pixel diameter yields stable performance. All super-pixels whose area lies >50% inside the corridor are considered members of the same clique (Fig. 2).

Junction cliques are star-shaped configurations with three corridors meeting in a central node. They are generated by randomly sampling a central superpixel and three additional ones to define the incoming corridors. Note that any crossing with >3 branches can be represented with multiple junction cliques. Figures 3(a) & (b) show the densities of sampled network and junction cliques, respectively.

In order to reduce the computational burden during inference, we introduce additional criteria to discard irrelevant cliques. First, we only sample seed points from the set of all nodes that have unary *road* probabilities above 0.5. Second, we limit the distance between two seeds; empirically two thirds of the image diagonal works well. Moreover, we discard cliques whose median unary *road* likelihood is below a threshold; for junction cliques, we additionally require a minimum angle between incoming corridors, to avoid pseudo-junctions where two corridors lie on the same road; in our experiments we use a threshold 30°.

4. Experiments

We evaluate our approach on aerial ortho-images from two different urban test sites, generated by dense matching and ortho-rectification. Both consist of 500 × 500 pixels



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

Figure 3. (a) Aerial photo of GRAZ, (b) overlaid with segmentation [26], (c) density of network cliques, (d) and density of junction cliques. Red indicates high density, blue low density.

tiles with a ground resolution of 0.5 m. The first dataset of the city VAIHINGEN, Germany, contains 14 color infrared tiles. 4 tiles were used for training and 10 for testing. The second dataset contains 18 RGB tiles from the city of GRAZ, Austria. 4 tiles were used for training and 14 for testing.[2]

We compare the proposed approach (abbreviated HOP) to three different baselines (Tab. 1 & 2, Fig. 4). The first baseline (RF) is the classification based only on the Random Forest unaries. The second baseline (Potts) also uses the same unaries, but smooths them with a standard first-order contrast-sensitive Potts model. The third baseline (Rules) simply samples promising straight road segments and assigns all their pixels to the *road* class. Sampling follows the same rules also used to generate our higher-order cliques, but with stricter thresholds that yield better performance (the standard clique sampling thresholds conservatively aims to include all potentially useful cliques and would create an excessive number of false positives). This baseline emulates the often used heuristic completion of the road network and allows one to separate the effect of our clique sampling scheme from the effect of the subsequent CRF inference.

We designate all superpixels with $P(road) > 0.5$ as seeds and sample 10 network cliques, respectively 1000 junction cliques, per seed. After pruning unlikely cliques (Sec. 3.4), between 1000 and 5000—depending on image content— remain as higher order potentials. The parameters for our CRF model were determined by grid search and kept constant for all experiments in both datasets: $\alpha = 2$, $\beta = 1$, $\gamma = 0.45$ and $w_{max} = 2$.

4.1. Evaluation measures

We conduct 4-fold cross-validation and employ several different performance metrics to assess our results. First, we compute the *completeness*, *correctness*, and *quality* measures introduced by Wiedemann *et al.* [28] and widely used in the literature on road extraction, *e.g.* [17, 18, 12, 19, 20]. They are defined as follows: Completeness is a variant of recall, which accounts for the fact that road boundaries are noisy and ill-defined. The estimated road segments as well as the ground truth are skeletonized to obtain centerlines. Ground truth centerline pixels are deemed true positives (TP) if they lie within a buffer of width B around the estimated centerline, and false negatives (FN) otherwise. Then, $completeness = TP/(TP + FN)$. Correctness is the equivalent variant of precision: Estimated centerline pixels are TP if they lie within B pixels of the ground truth centerline, or false positives (FP) otherwise, and $correctness = TP/(TP + FP)$. Quality combines both criteria into a single number according to $quality = TP/(TP + FP + FN)$. The buffer width is set to $B = 5$ pixels, corresponding to the narrowest roads we wish to extract.

Additionally, we also assess the pixel-wise segmentation accuracy of our road extraction. To that end we compute the κ-coefficient, widely used as a performance metric in remote sensing. The κ-value measures how much the predicted labels differ from a random label image with the same class frequencies. By measuring the improvement over a chance agreement, as opposed to the one over a 100% wrong result that is measured by the overall accuracy, κ compensates frequency biases.[3]

Although widely used in mapping, these metrics completely disregard the topological correctness of the extracted network, which is crucial for routing and navigation purposes. Consider the case when a short piece of road is misclassified, creating two dead ends. Such a mistake will cause map users to take long detours, while only marginally influencing the completeness (respectively recall). To mea-

[2]The VAIHINGEN data are part of the ISPRS benchmark http://www.itc.nl/ISPRS_WGIII4/tests_datasets.html; The GRAZ data has been kindly provided by Microsoft Photogrammetry, Graz. Data and ground truth will be made available upon publication.

[3]Formally, $\kappa = \frac{N \sum_i c_{ii} - \sum_i (\sum_j c_{ij} \cdot \sum_j c_{ji})}{N^2 - \sum_i (\sum_j c_{ij} \cdot \sum_j c_{ji})}$, where the c_{ij} are the entries of the confusion matrix and N is the number of pixels. Consider an image with 10% *road* and 90% *background* pixels. A classifier which *always* returns *background* will have 90% overall accuracy, but κ=0%.

	Method	*Qual.*	*Compl.*	*Corr.*	κ
GRAZ	RF	58.0	70.5	76.9	72.7
	Potts	56.2	63.2	**83.4**	71.6
	Rules	39.9	62.3	52.7	57.6
	HOP	**59.9**	**76.9**	73.5	**78.6**
VAIHINGEN	RF	51.7	61.3	78.5	66.6
	Potts	49.2	55.1	**83.5**	64.8
	Rules	49.2	63.4	68.9	66.5
	HOP	**55.6**	**69.4**	75.0	**71.6**

Table 1. Detection performance of road extraction methods: quality, completeness, correctness, κ. All numbers are percentages.

	Method	*Correct*	*2long*	*2short*	*NoConn*
GRAZ	RF	49.7	10.9	0.0	39.0
	Potts	56.1	10.1	1.3	32.6
	Rules	42.0	1.0	56.8	1.0
	HOP	**75.8**	2.6	6.8	14.8
VAIHINGEN	RF	27.1	7.5	0.0	65.4
	Potts	30.6	4.0	0.0	65.3
	Rules	53.6	3.3	26.8	16.3
	HOP	**58.4**	3.9	8.5	29.2

Table 2. Topological performance of road extraction methods: Percentage of paths that are correct, too long, too short, infeasible.

sure how well the topology of the road network has been estimated we define additional error metrics that quantify the topological quality of the extraction results. These are computed in the following way: we randomly sample two points which lie both on the true and the estimated road network, and check whether the shortest path between the two points has the same length in both networks (up to a deviation of 5% to account for geometric uncertainty). We then keep repeating this procedure with different random points and record the percentages of correct, too short, too long and infeasible paths, until these percentages have converged. Infeasible and too long paths indicate missing links, whereas too short ones indicate hallucinated connections.

4.2. Graz

The GRAZ dataset depicts a city center with major roads and large building blocks. RF fails for narrow roads and in cases of occlusion due to cars, shadows or trees (*c.f*. 4, left column, upper two rows). As expected, a contrast-sensitive Potts prior smooths away even more of the road superpixels and is unsuitable for the task. Note that we already use very low weight for the pairwise terms. Stronger smoothing leads to even worse results, *i.e*. even larger parts of the road network are labeled as *background*. Obviously, this method nevertheless achieves the highest correctness, because it aggressively relabels *road* superpixels as *background*, which inherently reduces the false positive rate (Tab. 1).

Heuristically bridging gaps in the Random Forest unaries result (Rules) leads to many false positives, *c.f*. Fig. 4. Consider the falsely detected road (blue stripe) in the upper right part of the first GRAZ image in Fig. 4. The real road happens to be pointing at the inner courtyard near the upper image border, suggesting that the road should be extended (*e.g*. in our framework superpixels on either side of the gap will be sampled as seeds). The proposed HOP model does contain such cliques. Still, no false connection is created, since the unaries strongly attract the superpixels in the gap to the *background* class, and due to the contrast-sensitive node weights they only incur a small penalty for not changing to *road*. The effect occurs more often in the second GRAZ example. Again, the HOP model copes a lot better with the situation. HOP also performs best in terms of topological correctness, outperforming the baselines by 20-34 percent points (Tab. 2). Note that Rules has the highest percentage of too short paths (56.8%) because of the many false positives (*c.f*. Fig. 4).

Overall, the HOP model clearly extracts the road network most faithfully. There are two main failure modes: *(i)* if long *continuous* stretches of road are not detected by the unaries, the model in its current form will not be able to remedy this, since no cliques can be constructed which contain enough road evidence (*e.g*. Fig. 4, bottom row, top right corner). We point out that the requirement $P_b < P_r$ in eqn. (2) does *not* mean that only roads with $>50\%$ support in the unaries are detected: all that is required is that for every edge of the road network there exist enough *subsegments* with at least 50% support. *(ii)* dead ends undetected by the unaries cannot be repaired, since cliques are only sampled between seed pixels (*e.g*. Fig. 4, first row, top left corner). Another possible cause of failure is due to the definition of a "road". The proposed prior is designed for roads that are elongated and/or connect junctions. Short, isolated regions like courtyard parking lots, which are often also labeled as *roads* (as is the case in our ground truth), are not covered and are neither encouraged nor discouraged.

4.3. Vaihingen

The VAIHINGEN road network has a different, more irregular and complex structure. There are many short and narrow roads, with cast shadows and overhanging trees. Building shapes and sizes are more diverse than those of GRAZ, making road extraction more challenging. Quantitative results are given in the bottom half of table 1.

The HOP model outperforms all three baselines on all performance measures except correctness, again because of the standard Potts model's excessive bias towards the *background* class (Tab. 1 & 2). Qualitative results in Fig. 4 confirm that the Potts model clearly performs worst, in spite of the high correctness numbers. The results conform with our assertion that pairwise smoothing potentials are counterproductive for the extraction of thin structures.

The Rules baseline again erroneously cuts across objects. Consider the center of the first VAIHINGEN image where blue stripes indicate that buildings and tree rows surrounded

by road and parking lots are misclassified. The effect is observable once more in the second VAIHINGEN image. HOP overcomes various challenging situations not solved by RF and Potts. For example, consider the narrow road in the bottom left of the first VAIHINGEN image. RF wrongly labels many superpixels as *background*, and consequently the road is cut into several small pieces. Another situation leading to misclassification by RF (and Potts) can be observed on the major road diagonally crossing the image from bottom center to upper right. High buildings and trees cast dark shadows on the road resulting in false negatives. The HOP prior extracts the entire road successfully.

HOP also again shows the highest topological correctness, improving between 5 and 31 percent points over the baselines. Note that Rules has the lowest percentage of "no connections" because it generates many wrong roads across the background (*c.f.* Fig. 4), which is reflected in the high rate of too short paths (26.8%).

5. Conclusion and outlook

We have formulated an effective and efficient model for road network extraction based on the P^N-Potts model. The proposed method significantly outperforms standard baselines, both in terms of segmentation accuracy and of a newly introduced measure of topological correctness.

At present, the prior kicks in only if the dominant label in a clique is *road*. In future work, we plan to extend it such that even less evidence is required, if its distribution (especially alignment) suggests the presence of a road. We also plan to include additional cues in the prior, *e.g.* the canyon-shaped profile of roads in a height map, if they are surrounded by buildings or trees.

Moreover, we have manually determined the parameters for clique sampling. It seems feasible to learn these parameters automatically from the vast amounts of existing digital map data, and to eventually develop a probabilistically motivated sampling scheme without hard thresholds. In the same vein one could also try to learn more complex patterns in the road network, such as rectangular loops or the T-junctions at highway exits. It is still unclear where the sweet spot lies between a tractable and efficiently computable model and one that closely approximates the highly complex distribution of local road patterns.

Finally, the proposed model is not *per se* limited to roads in overhead images, and it would be interesting to adapt it to other domains that exhibit linear network structures, such as blood vessels or neurons in medical imagery, or material inspection (crack detection) in aerospace engineering.

References

[1] R. Bajcsy and M. Tavakoli. Computer recognition of roads from satellite pictures. *IEEE T. Systems, Man, and Cybernetics*, 6(9):623 – 637, 1976.

[2] Y. Boykov and M.-P. Jolly. Interactive organ segmentation using graph cuts. In *MICCAI*, 2000.

[3] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *PAMI*, 23(11):1222–1239, 2001.

[4] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[5] G. J. Burghouts and J.-M. Geusebroek. Material-specific adaptation of color invariant features. *Pattern Recognition Letters*, 30(3):306–313, 2009.

[6] P. Felzenszwalb and D. Huttenlocher. Efficient belief propagation for early vision. *IJCV*, 70(1), 2006.

[7] M. Fischler, J. Tenenbaum, and H. Wolf. Detection of roads and linear structures in low-resolution aerial imagery using a multisource knowledge integration technique. *Computer Graphics and Image Processing*, 15:201 – 223, 1981.

[8] C. Galleguillos, A. Rabinovich, and S. Belongie. Object categorization using co-occurrence, location and appearance. In *CVPR*, 2008.

[9] S. Gould, J. Rodgers, D. Cohen, G. Elidan, and D. Koller. Multi-class segmentation with relative location prior. *IJCV*, 80(3):300–316, 2008.

[10] C. Heipke, H. Mayer, and C. Wiedemann. Evaluation of automatic road extraction. In *3D Reconstruction and Modeling of Topographic Objects*, 1997.

[11] S. Hinz and A. Baumgartner. Automatic extraction of urban road networks from multi-view aerial imagery. *ISPRS J. Photogrammetry and Remote Sensing*, 58:83 – 98, 2003.

[12] J. Hu, A. Razdan, J. C. Femiani, M. Cui, and P. Wonka. Road network extraction and intersection detection from aerial images by tracking road footprints. *IEEE TGRS*, 45(12):4144 – 4157, 2007.

[13] P. Kohli, L. Ladicky, and P. H. S. Torr. Robust higher order potentials for enforcing label consistency. In *CVPR*, 2008.

[14] V. Kolmogorov. Convergent tree-reweighted message passing for energy minimization. *PAMI*, 28(10), 2006.

[15] C. Lacoste, X. Descombes, and J. Zerubia. Point Processes for unsupervised line network extraction in remote sensing. *PAMI*, 27(10):1568 – 1579, 2005.

[16] L. Ladicky, C. Russell, P. Kohli, and P. H. Torr. Graph cut based inference with co-occurrence statistics. In *ECCV'10*.

[17] 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*, 12:23 – 31, 2000.

[18] H. Mayer, S. Hinz, U. Bacher, and E. Baltsavias. A test of automatic road extraction approaches. In *IAPRS*, volume 36(3), pages 209 – 214, 2006.

[19] V. Mnih and G. E. Hinton. Learning to detect roads in high-resolution aerial images. In *ECCV*, 2010.

[20] V. Mnih and G. E. Hinton. Learning to label aerial images from noisy data. In *ICML*, 2012.

[21] J. Shotton, J. Winn, C. Rother, and A. Criminisi. TextonBoost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *ECCV*, 2006.

[22] R. Stoica, X. Descombes, and J. Zerubia. A Gibbs Point Process for road extraction from remotely sensed images. *IJCV*, 57(2):121 – 136, 2004.

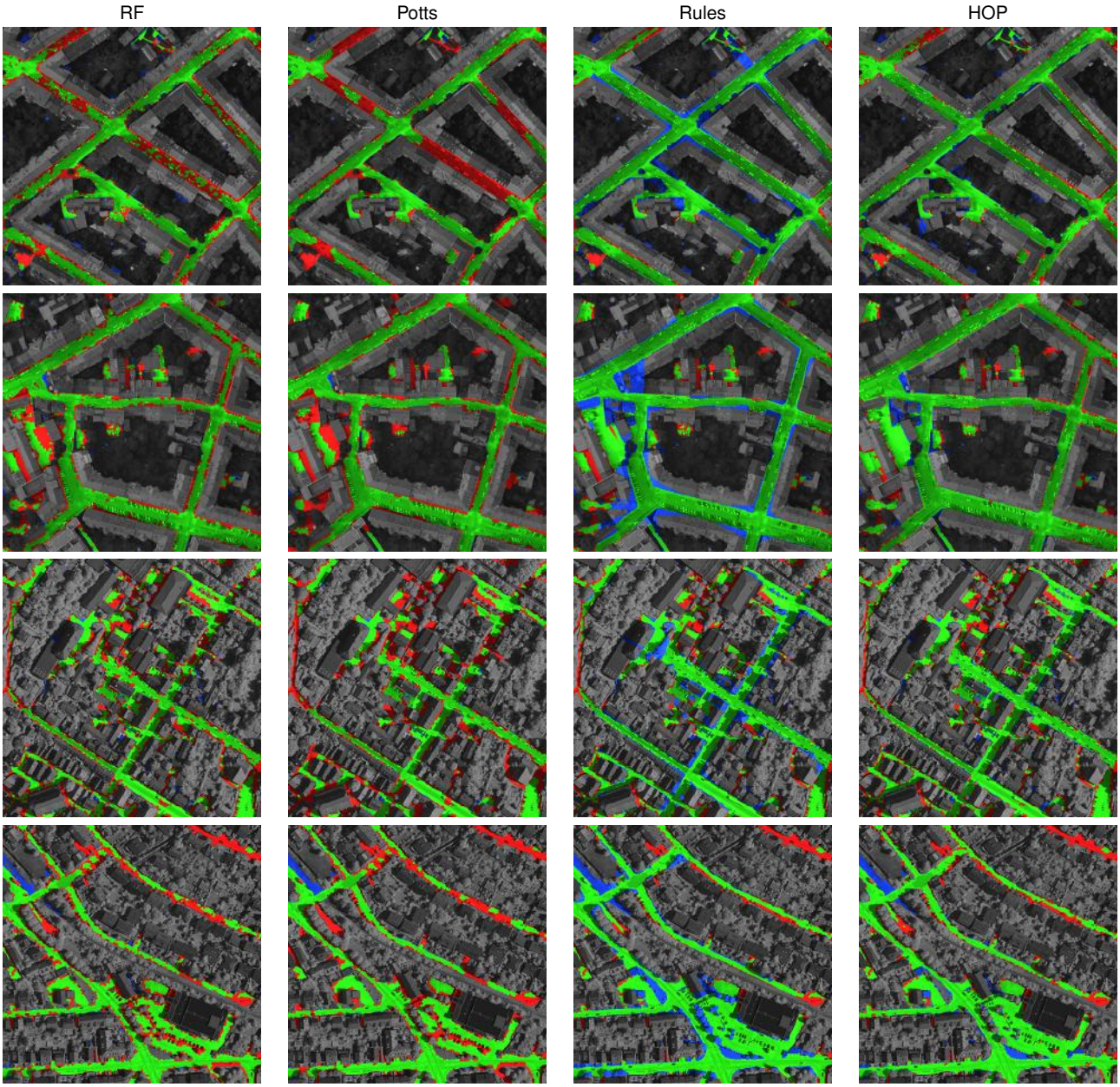


Figure 4. Road networks extracted in four different aerial photos (GRAZ upper two rows, VAIHINGEN lower two rows). True positives are displayed green, false positives blue, and false negatives red.

[23] A. Torralba, K. Murphy, and W. Freeman. Contextual models for object detection using boosted random fields. In *NIPS'05*.

[24] F. Tupin, B. Houshmand, and M. Datcu. Road detection in dense urban areas using SAR imagery and the usefulness of multiple views. *IEEE TGRS*, 40(11):2405 – 2414, 2002.

[25] F. Tupin, H. Maître, J.-F. Mangin, J.-M. Nicolas, and E. Pechersky. Detection of linear features in SAR images: Application to road network extraction. *IEEE TGRS*, 36(2):434 – 453, 1998.

[26] O. Veksler, Y. Boykov, and P. Mehrani. Superpixels and supervoxels in an energy optimization framework. In *ECCV'10*.

[27] S. Vincente, V. Kolmogorov, and C. Rother. Graph cut based image segmentation with connectivity priors. In *CVPR'08*.

[28] C. Wiedemann, C. Heipke, H. Mayer, and O. Jamet. Empirical evaluation of automatically extracted road axes. In *CVPR Workshops*, 1998.

[29] J. Winn, A. Criminisi, and T. Minka. Object categorization by learned universal visual dictionary. In *CVPR*, 2005.

[30] J. Yao, S. Fidler, and R. Urtasun. Describing the scene as a whole: Joint object detection, scene classification and semantic classification. In *CVPR*, 2012.

[31] J. Zhao and S. You. Road network extraction from airborne LiDAR data using scene context. In *CVPR Workshops*, 2012.