ISPRS Journal of Photogrammetry and Remote Sensing 108 (2015) 128–137

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

Road networks as collections of minimum cost paths

Jan Dirk Wegner *, Javier Alexander Montoya-Zegarra, Konrad Schindler

Photogrammetry and Remote Sensing, ETH Zürich, Switzerland

ARTICLE INFO

Article history:
Received 25 June 2014
Received in revised form 1 July 2015
Accepted 9 July 2015

Keywords:
Aerial
Multispectral
Urban
Networks
Extraction
High resolution

ABSTRACT

We present a probabilistic representation of network structures in images. Our target application is the extraction of urban roads from aerial images. Roads appear as thin, elongated, partially curved structures forming a loopy graph, and this complex layout requires a prior that goes beyond standard smoothness and co-occurrence assumptions. In the proposed model the network is represented as a union of 1D paths connecting distant (super-)pixels. A large set of putative candidate paths is constructed in such a way that they include the true network as much as possible, by searching for minimum cost paths in the foreground (*road*) likelihood. Selecting the optimal subset of candidate paths is posed as MAP inference in a higher-order conditional random field. Each path forms a higher-order clique with a type of clique potential, which attracts the member nodes of cliques with high cumulative road evidence to the foreground label. That formulation induces a robust P^N-Potts model, for which a global MAP solution can be found efficiently with graph cuts. Experiments with two road data sets show that the proposed model significantly improves per-pixel accuracies as well as the overall topological network quality with respect to several baselines.

1. Introduction

Despite more than three decades of research, automatic road extraction from remote sensing data remains to a large degree unsolved. Since the initial attempts in the mid-seventies (Bajcsy and Tavakoli, 1976) important progress has been made – see the overview papers (Heipke et al., 1997; Mayer et al., 2006) – but to our knowledge no fully automated road extraction system so far performs at a level that would allow operational use. In practice the extraction or update of roads is at most semi-automatic and requires a significant degree of user interaction, e.g. (Gerke et al., 2004; Zhang, 2004; Helmholz et al., 2012).

Factors that make the task challenging are strong illumination effects, appearance variations due to clutter and shadows, and occlusion by nearby buildings and vegetation. In urban environments these factors are compounded with highly variable road width, density and curvature, which makes the extraction particularly difficult. Even in "planned" cities with a regular grid layout (e.g., many American towns) nearby trees and buildings frequently cast shadows on roads or occlude them altogether. For older or more informally growing cities with irregular, narrow, winding roads the problem becomes much worse.

Road extraction in the presence of noisy and ambiguous low-level image evidence requires strong a priori knowledge. It turns out that formalizing the structural properties of roads (and also other networks, e.g., in medical image processing) in a prior is difficult. Existing models are usually either too restricted to faithfully describe the network, or too complex for stable and efficient inference (see overview in Section 2).

We seek a compromise between these extremes and develop a model of the road network which is on the one hand expressive (e.g., it does not impose a tree structure or require piecewise straight roads), and on the other hand amenable to powerful inference algorithms (i.e., it does not require expensive all-purpose solvers like MCMC or Gibbs sampling). The proposed method follows the recover-and-select strategy: an over-complete collection of potential road segments is generated, which is subsequently pruned to those segments which cover parts of the road network. The segments, which we call *paths*, are found by minimum cost path computation based on local features. The pruning step, in which incorrect paths are suppressed, is formulated as MAP inference in a higher-order conditional random field (CRF), constructed in such a way that it allows for efficient global energy minimization. The conservative *recover* step ensures high completeness (recall), while the *select* step aims to maximize correctness (precision), by explicitly including long-range connections via higher-order CRF potentials.

* Corresponding author.

http://dx.doi.org/10.1016/j.isprsjprs.2015.07.002

To our knowledge, this is the first attempt to combine the classical idea of minimum cost paths with the comprehensive global modeling capabilities of CRFs, for road network extraction in particular and for other loopy, undirected networks in general. A preliminary version of this work appeared in Wegner et al. (2013). In that work local stretches of road were confined to lie on straight line segments. Here, we extend the model to allow for arbitrary (minimum-cost) paths that naturally adapt to more complex road shapes (e.g., sharp bends). Moreover, we present a much expanded experimental evaluation.

2. Related work

Road extraction in rural or suburban areas is often approached in a rule-based fashion, i.e. one attempts to explicitly formulate an exhaustive set of rules for delineating the road network (Doucette et al., 2004; Mena and Malpica, 2005; Poullis and You, 2010; Grote et al., 2012; Ünsalan and Sirmacek, 2012; Miao et al., 2013). Common to all these approaches is a heuristic processing pipeline consisting of multiple sequential or intertwined steps, with a rather large set of parameters that need to be re-tuned for each new scene.

To bridge the gap between low-level road cues and high-level road network layout in a more principled way Stoica et al. (2004) (later followed by Lacoste et al. (2005) and Lafarge et al. (2010)) have introduced marked point processes, a comprehensive probabilistic framework to impose connectivity priors. In Chai et al. (2013)) the authors extend the original idea of sampling line-segments by explicitly modeling junctions with point processes. However, the corresponding objective functions can only be minimized with all-purpose solvers like simulated annealing and/or reversible jump Markov Chain Monte Carlo (RJMCMC). They are thus on one hand computationally very expensive and on the other hand risk not finding a satisfactory optimum (e.g. due to poor mixing of the chain).

All previously mentioned approaches primarily focus on rural and suburban scenes with relatively sparse and mostly unoccluded road networks. Only few works deal with road extraction in more complex urban areas. Hinz and Baumgartner (2003) have developed a detailed heuristic model for roads and their context in scale-space, using evidence from multiple overlapping aerial images. More recently, Youn et al. (2008) combine an orthophoto and airborne laser scanning data to extract wide, largely unoccluded roads that follow a grid pattern. Similar to Hinz and Baumgartner (2003) and Grote et al. (2012) they design a hierarchical framework which constructs longer road pieces from initial segments, but no high-level connectivity is imposed, thus many gaps remain.

Here, we argue that a particularly important property of the road network is its connectedness. Hierarchical bottom-up procedures that iteratively assemble short pieces of road to longer ones must base their decisions primarily on local shape constraints, whereas they account for connectedness at a very late stage (or not at all). It seems more intuitive to view road networks as a collection of smooth, connected long-range paths without strong restrictions on the local shape. Probably the first work to model roads via minimum cost paths is (Fischler et al., 1981). They use an A^*-type algorithm to iteratively find roads based on per-pixel scores generated with a line detector. Since this early attempt various groups have proposed *semi-automated* road tracking approaches, in which single roads (mostly in rural areas) are traced after manual selection of starting nodes. Technical implementations of this idea include Kalman filtering (Vosselman and de Knecht, 1995), extended Kalman filtering and/or particle filtering (Movaghati et al., 2010), heuristic rule-based tracing (Baumgartner et al., 2002), and shortest path computation by dynamic programming (Gruen and Li, 1995; Gruen and Li, 1997; Dal Poz et al., 2010; Dal Poz et al., 2012).

To our knowledge, minimum cost paths for *automated* road network extraction have not been followed up in recent years in remote sensing, until recently Türetken et al. (2012) tested their method, originally developed for vessel tree extraction in medical imagery, on road networks. In *medical imaging*, many researchers have used minimum cost paths to model 2D and 3D tree structures, e.g. (Li and Yezzi, 2007; Türetken et al., 2011; Benmansour and Cohen, 2011; Bas and Erdogmus, 2011; Zhao et al., 2011). Generally, these approaches first detect local cues, which are then connected to elongated tubes via minimum cost paths. Model-based criteria ensure that all branches fit into a global tree topology, either in a bottom-up or in a top-down fashion. Bottom-up methods try to initially extract only correct network pieces, thereby accepting low completeness, followed by insertion of missing links, e.g. (Bas and Erdogmus, 2011; Wang et al., 2011). Top-down methods proceed the other way round, and first generate an overly complete network, by allowing all potential paths at the risk of a high false alarm rate. Subsequently, erroneous links are pruned to obtain a correct network, e.g. (Li and Yezzi, 2007; Türetken et al., 2012). Bottom-up methods are usually fast to compute iteratively but often fail to bridge large gaps, whereas top-down techniques have problems when it comes to suppressing "shortcuts" through the background. The approach proposed here follows the top-down strategy and aims for high *topological completeness*, i.e. our objective is to extract the complete urban road network *as far as possible*, which is crucial for navigation applications such as personal navigation systems or vehicle routing.

A method similar in spirit to ours is (Türetken et al., 2012), which was extended to graphs with cycles in Türetken et al. (2013). They locally compute road (resp. tube) likelihoods at each pixel (or voxel, if applied to stacks of medical images) and connect seed points via minimum cost paths at several scales. The resulting graph is broken down into short overlapping segments, and a network graph through the set of segments is found with mixed integer programming. Although originally developed for a medical application, the experiments also demonstrate promising performance for suburban road networks in aerial images.

For completeness we also mention a body of literature, starting with (Laptev et al., 2000), that uses the term "road extraction" for the delineation of roads with different variants and extensions of active contour models ("snakes"). For example, Butenuth and Heipke (2012) extend standard snakes to explicitly model the network topology, including junctions, with so-called network snakes. Wang et al. (2011) apply a similar snake formulation to iteratively reconstruct tree-like tubular structures from medical image stacks. However, snakes are a local optimizer, and mainly useful to delineate roads more precisely once their approximate layout is known. We therefore rather see them as a potential geometric refinement *after* extraction.

3. Network extraction

In our approach the road network is thought of as the union of many elongated *paths*. In this way, network extraction can be cast as the search for a set of paths that together cover the entire network. The proposed method follows the recover-and-select strategy:

- In the *recover* step a large, over-complete set of potential *candidate paths* is generated, by finding the most road-like connections between many different pairs of seed points. The aim of candidate generation is high recall, ideally the candidate set covers the entire road network, at the cost of also containing many false positives that do not lie (completely) on roads.

- In the *select* step undesired false positives are pruned from the candidate set to yield a reduced set still covering as much as possible of the network, but with few false positives. This second step is formulated as the minimization of a global higher-order CRF energy, and can be solved to global optimality.

In more detail, our system consists of the following steps: First, an image is segmented into superpixels, which are from then on treated as the smallest entities (nodes) to be labeled. Per superpixel a feature vector is extracted and fed to a binary Random Forest classifier, which assigns each superpixel a unary road likelihood (Section 3.1). Next, promising candidate paths are generated. To that end, superpixels with high road likelihoods are sampled randomly as seed nodes and linked with minimum cost paths. The hope is that *road* superpixels that have high *background* probability, e.g. due to a cast shadow, will be covered by a minimum cost path and thus become member of a connected subset where the majority of superpixels votes for *road*. The superpixels of each candidate path form a higher-order clique in a CRF (Section 3.2).

The potentials of these higher-order cliques are based on the P^N-Potts model of Kohli et al. (2009) that enforces *label consistency* within large cliques, meaning that superpixels within the clique are penalized for deviating from the majority label. In that sense, our method could be seen as an anisotropic "smoothing along the paths".

The resulting CRF energy can be minimized with a graph cut, leading to a global optimum of the binary labeling task. Note that working with the actual long-range paths (cliques) is conceptually different from methods that divide long paths into short segments and classify each segment separately (Türetken et al., 2012). Like (Wang et al., 2011; Türetken et al., 2011) we prefer to work with complete paths, so as not to lose any connectivity information.

3.1. Unary potential

Recall that our smallest entity to be labeled is a superpixel. By a slight abuse of notation, in the following we write $\mathbf{x}$ for both the raw data and the features. Correspondingly, we denote both a particular superpixel from the set S of all superpixels and its features with x_j. Thus, our objective is to assign each superpixel x_j a label $y_j \in \{0, 1\}$, where 1 represents *road* and 0 *background*.

We segment the raw images into superpixels x_j, and train a binary Random Forest classifier (Breiman, 2001) with 20 trees to predict, for each superpixel, the class-conditional (negative log-)likelihoods $E_1(x_j) = -\log P(y_j = 1|x_j)$ for the foreground (*road*) class and $E_0(x_j) = -\log P(y_j = 0|x_j)$ for the *background*. These log-likelihoods form the unary potentials in the conditional random field, i.e. the energies for assigning labels $y_j \in \{0, 1\}$ are then

$$E_u(x_j) = y_j E_1(x_j) - (1 - y_j) E_0(x_j) \tag{1}$$

Although our method could in principle work directly with individual pixels we prefer regular superpixels of an oversegmentation as smallest entities to be labeled, on one hand due to their larger support for feature computation, and on the other hand to reduce the computational burden, both during shortest path generation and during inference. Superpixels sometimes are not correctly aligned with object boundaries, but we do not consider this a major hurdle for our application because the emphasis lies on the correct and complete road network topology rather than on pixel-accurate labeling, which can be achieved in a subsequent refinement step, for example using snakes (Laptev et al., 2000; Butenuth and Heipke, 2012).

As features we use the responses of the color/texture filter bank of Winn et al. (2005), after converting the images to opponent Gaussian color space (Burghouts and Geusebroek, 2009), as well as the height over ground, if available. For each superpixel we record mean and standard deviation of all features, leading to a feature vector of dimension 34 for raw images, respectively 36 if a height channel is available.

An important property of the problem, which is rarely mentioned in the literature, is that the foreground–background labeling problem for line networks is usually very asymmetric: much of the clutter that disturbs the road appearance (overhanging trees, cast shadows, etc.) is also prominent in the diffuse statistics of the background ("everything except roads"), whereas only few things in the background exhibit the comparatively crisp, well-defined statistics of roads. Consequently false negatives in the foreground class (gaps in the road network) are the dominant failure mode of the unary potential that needs to be addressed by a prior, whereas false positives are much less of a problem – see Section 4.2, Figs. 2(a,e,i) and 3(a,e,i,m).

3.2. Minimum cost paths

As explained above, we concentrate on a model to correct false negatives, i.e. superpixels that do actually lie on roads, but have low *road* likelihood. We thus construct an asymmetric prior which attracts superpixels with low road probability to the road class to reduce false negatives, but not the other way round.

Our model represents the network as a union of many minimum cost paths that link two randomly sampled superpixels (Fig. 1), one *start* node x_s and one *end* node x_e. Clearly, the set of all such paths in an image is too large for practical purposes. In order to keep shortest path computation and the subsequent CRF inference tractable, we sample node pairs $\{x_s, x_e\}$ from all superpixels with a road likelihood $\geqslant 0.7$. The minimum cost paths are found with the standard Dijkstra algorithm (Dijkstra, 1959). In order to also cover comparable paths with slightly higher costs due to noise, we use the k-shortest path version of the algorithm to get k mutually exclusive paths per node pair.

Let the path from x_s to x_e be denoted $R_i(s_x, s_e)$, and let its nodes be $\{x_j \in R_i\}$. The cost of a path is

$$C(R_i(x_s, x_e)) = \sum_{j=s}^{e} c(x_j), \tag{2}$$

where in the basic case the individual edge costs are simply $c(x_j) = E_1(x_j)$, stating a preference for paths that pass through nodes of high road likelihood.

For the case of road extraction from aerial images, where often also a height value $h(x_j)$ for each superpixel is available from dense stereo or a pre-existing terrain model, we additionally require low slope (height gradient) $\Delta h(x_j) = h(x_{j+1}) - h(x_j)$ between neighboring nodes:

$$c(x_j) = \lambda E_1(x_j) + (1 - \lambda)\left|\Delta h(x_j)\right|, \tag{3}$$

with $\lambda \in [0 \ldots 1]$ a weighting parameter that determines the relative influence of appearance and slope.

Note that paths are instantiated for *all* sampled pairs of start/end nodes, without knowing whether there really exists a connection between them. Although roads may frequently be occluded, we observe that this only occurs for a certain maximum number of consecutive superpixels on a path. In general, $\geqslant 10$ consecutive superpixels with unary road likelihood below 0.5 indicate background. We therefore directly prune unlikely paths that pass through too long stretches of background, and pass only the remaining ones to CRF inference. Each sampled minimum cost path R_i that passes this initial threshold forms a higher-order clique Q_i in a CRF for the selection (inference) step.

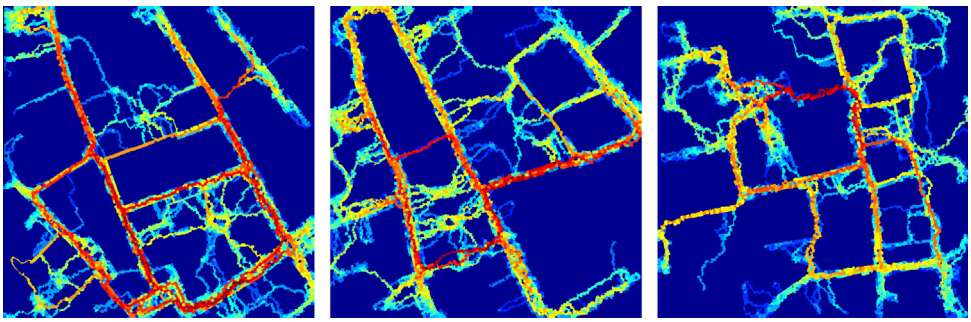

Fig. 1. Densities of minimum cost paths per pixel of GRAZ images shown in Fig. 2. High densities are displayed red, low densities blue.

3.3. Contrast-sensitive node weighting

To measure the goodness-of-fit of individual nodes x_j w.r.t. a given clique Q_i, we introduce weights w_i^j. They are derived by comparing the nodes' appearance to the mean appearance of the clique. To that end, we compute the mean feature vector $\overline{x}(Q_i)$ in the clique i, the Euclidean distance to the mean $|x_j - \overline{x}(Q_i)|$ for each individual node, and the standard deviation $\sigma_x(Q_i)$ of those distances. The variance-adjusted distance $d(x_j, Q_i) = |x_j - \overline{x}(Q_i)|/\sigma_x(Q_i)$ is then used to assign an individual weight to the node x_j:

$$w_i^j = \begin{cases} w_{max} & d(x_j, Q_i) \leqslant 0.5 \\ w_{max}(1 - d(x_j, Q_i)) & 0.5 < d(x_j, Q_i) \leqslant 1 \\ 0 & d(x_j, Q_i) > 1 \end{cases} \quad (4)$$

using a truncated linear weighting function for robustness. The scale w_{max} can be chosen arbitrarily, since it is later rescaled with the path weight λ_{path}, see Eq. (6). We set $w_{max} = 1$. The weights w_i^j can be interpreted as "degrees of clique membership" or as "coupling strengths" between node and clique. They help to better handle cliques which lie largely on roads, but take incorrect "shortcuts" through the background. Such cliques are rather frequent due to the nature of shortest path computation, and if not properly treated can produce false positives. To understand the effect of the weights it is instructive to look at the extreme cases: nodes within $0.5\ \sigma_x(Q_i)$ get the full weight and strongly increase the penalty for labeling the superpixel as *background* (see below). In contrast, nodes further away than $\sigma_x(Q_i)$ get weight zero, effectively removing them from the clique, such that labeling them as *background* incurs no penalty.

3.4. Higher-order CRF model

In their work on the P^N-Potts model Kohli et al. (2008) showed that efficient inference in higher-order CRFs is feasible, if the number of possible states per clique remains low, whereas the absolute clique size is less crucial. Several works have adopted that model for semantic segmentation tasks. Examples include imposing multiple superpixel segmentations as soft constraints on object boundaries (Kohli et al., 2009), modeling long-range texture patterns (Rother et al., 2009), exploiting global co-occurrence statistics of object classes (Ladicky et al., 2010), and concurrently capturing the spatial extent, semantic class, and semantic context of objects (Yao et al., 2012). Few works exist that apply minimum cost paths for the extraction of single "objects" with complex layout. An example is (Vicente et al., 2008), which employs Dijkstra shortest paths within a graph cut framework as connectivity prior for interactive image segmentation, in order to counter the "shrinking bias".

In this paper we use higher-order potentials for road network extraction. A CRF models the posterior $P(\mathbf{y}|\mathbf{x})$ of labels $\mathbf{y}$ depending on observations $\mathbf{x}$ as a Gibbs distribution,

$$P(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp\left(E(\mathbf{x}, \mathbf{y})\right) \quad (5)$$

with $Z(\mathbf{x})$ the partition function which ensures that the probability integrates to 1. The joint Gibbs energy $E(\mathbf{x}, \mathbf{y})$ of unary potentials ψ_j (cf. Eq. (1)), pairwise potentials ψ_b, and higher-order potentials ψ_q of a CRF is (with N and Q the sets of all binary and all higher-order cliques, respectively)

$$E(\mathbf{x}, \mathbf{y}) = \sum_{j \in S} \psi_j(x_j, y_j) + \lambda_{bin} \sum_{n \in N} \psi_b(\mathbf{x}_n, \mathbf{y}_n) + \lambda_{path} \sum_{i \in Q} \psi_q(\mathbf{x}_i, \mathbf{y}_i). \quad (6)$$

Maximum a posterior (MAP) inference then amounts to minimizing $E(\mathbf{x}, \mathbf{y})$ to maximize the posterior probability $P(\mathbf{y}|\mathbf{x})$.

It turns out that moderate smoothing with a 1st-order contrast-sensitive Potts model (Boykov and Jolly, 2001) in addition to the higher-order cliques eliminates small, isolated false positives. We thus include a pairwise term with a low weight $\lambda_{bin} = 0.1$. The pairwise potentials ψ_b compare labels and features of *adjacent* superpixels, whereas the higher-order potentials ψ_q define interactions between *all superpixels contained in a minimum cost path* (cf. Section 3.2). In our case higher-order cliques can reach a size of up to ≈ 300 superpixels.

Only applying standard pairwise potentials to extract roads, which appear as narrow, elongated objects inside a dominant, heterogeneous background, is prone to fail. The local interactions between adjacent superpixels do not carry information about long-range connectivity, and will tend to smooth away thin structures like roads, cf. Section 4.2 and e.g. (Vincente et al., 2008). On the contrary, the higher-order cliques (paths) are designed to fill gaps and improve network completeness. With shortest path sampling we obtain elongated chains of superpixels (Fig. 1), the majority of which in most cases fall on a road. If the overall road evidence of a clique (derived from the unaries) is strong enough, then the higher-order potential drags clique members with less evidence to the foreground. For example, consider a case where long stretches of superpixels with high *road* likelihood on a path are interrupted by short, isolated groups of superpixels with low road likelihood caused by overhanging trees. Because path computation is engineered to follow roads, this is strong evidence that the isolated groups should also be labeled *road*, and consequently the clique potential attracts *all* superpixels of the path to the *road* class. On the contrary, if the majority of superpixels inside a path belongs to*background*, one cannot generally infer that all should be *background*. Therefore, the prior is asymmetric. Note, paths explicitly model foreground, but no prior for superpixels not covered by paths is formulated. In the absence of any path the prior remains neutral (however, the local smoothness prior from the pairwise potentials acts on both foreground and background).

In many cases paths that largely cover a stretch of road will nevertheless contain a few background superpixels. In such cases the energy should increase gracefully, rather than abruptly with the

Table 1
Pixelwise and topological road extraction results (mean values after cross-validation). All numbers are percentages.

	Method	κ	Qual.	Comp.	Corr.	2long	2short	noC.	Correct
Graz	RF	76	69	80	84	7	3	32	59
	Potts	76	69	76	89	5	3	30	62
	Thresh	68	62	87	69	2	36	1	62
	Paths	**78**	**71**	83	84	6	8	11	**75**
Vaih	RF	73	65	77	81	8	3	40	49
	Potts	73	46	60	67	5	3	42	50
	Thresh	68	61	87	68	1	43	1	55
	Paths	**76**	**68**	81	81	7	11	17	**65**

Bold values represent the top performance.

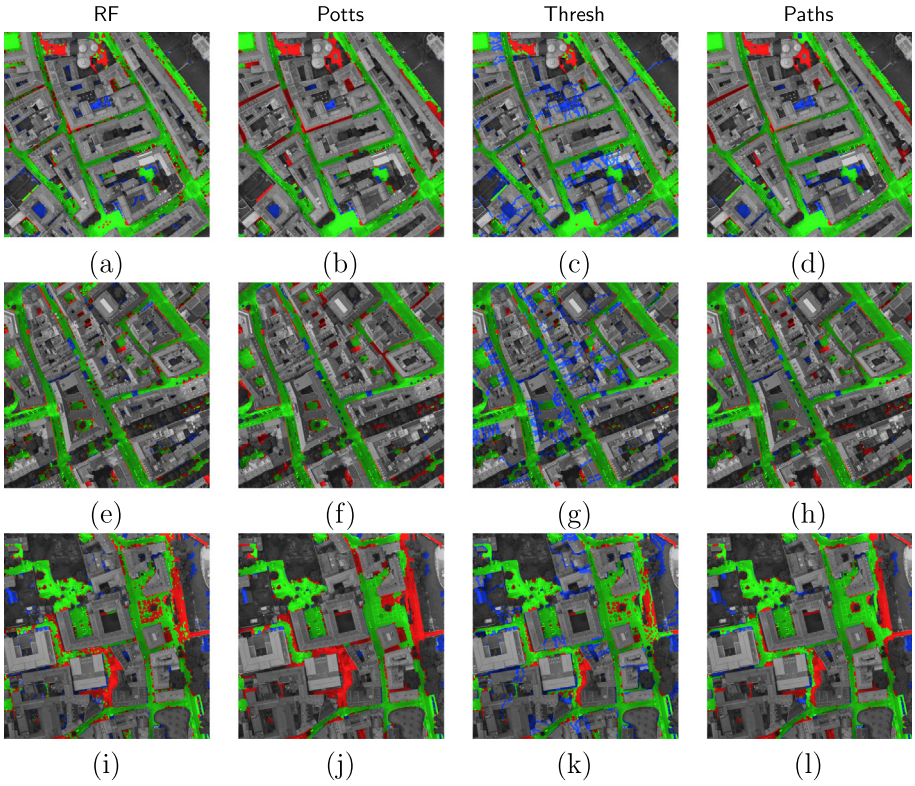

Fig. 2. Road networks extracted in three patches of the Graz orthophoto mosaic. Green true positives, blue false positives, red false negatives.

first deviating superpixel. We therefore employ the robust P^N-Potts model with a linear truncated cost function, with parameters β, γ and the potentials' upper bound α. Together with the asymmetry explained above the higher-order potential $\psi_q(\mathbf{x}_i, \mathbf{y}_i)$ becomes

$$\psi_q(\mathbf{x}_i, \mathbf{y}_i) = \begin{cases} \min\left(\alpha, P_b \cdot \frac{\alpha-\beta}{\gamma} + \beta\right) & \text{if} \quad P_b < P_r \\ 0 & \text{else} \end{cases} \tag{7}$$

where $P_r = \sum(w_i^j \cdot y_j)$ is the weighted sum of *road* superpixels inside a clique, and $P_b = \sum(w_i^j \cdot (1 - y_j))$ is the weighted sum of *background* superpixels. The w_i^j are the weights that act on individual superpixels j to adjust their influence on the potential of a clique i (cf. Section 3.3).

Since our problem only has two labels and the potential (7) is a special case of the robust P^N-Potts model (Kohli et al., 2008), a global minimum of the energy (6) can be found in low polynomial time with a graph cut.

4. Experiments

We evaluate the proposed approach on two different datasets. Both datasets consist of a number of 1000 × 1000 pixel tiles from aerial true orthophotos and corresponding normalized digital surface models (nDSM) from stereo matching, with a ground sampling distance of 0.25 m. The first dataset contains 76 (RGB) tiles covering the entire city center of Graz, Austria, the second one consists of 16 (color infrared) tiles from Vaihingen, Germany.[1]

Raw image tiles are segmented into superpixels with the patch-based energy minimization approach of Veksler et al. (2010). We generate on average 15,000 superpixels per tile, so as to limit graph size.[2]

To obtain candidate paths we randomly sample 1500 pairs of start/end nodes per tile for Graz, and 4000 pairs for the more difficult Vaihingen. These seeds are sampled from all superpixels with high unary road likelihood $P(y_j = 1|x_j) \geqslant 0.7$. The number k of

[1] The Vaihingen data are part of the ISPRS benchmark. The Graz data has been kindly provided by Microsoft Photogrammetry, Graz.
[2] Parameters are found empirically. See Section 4.3 for a parameter study.

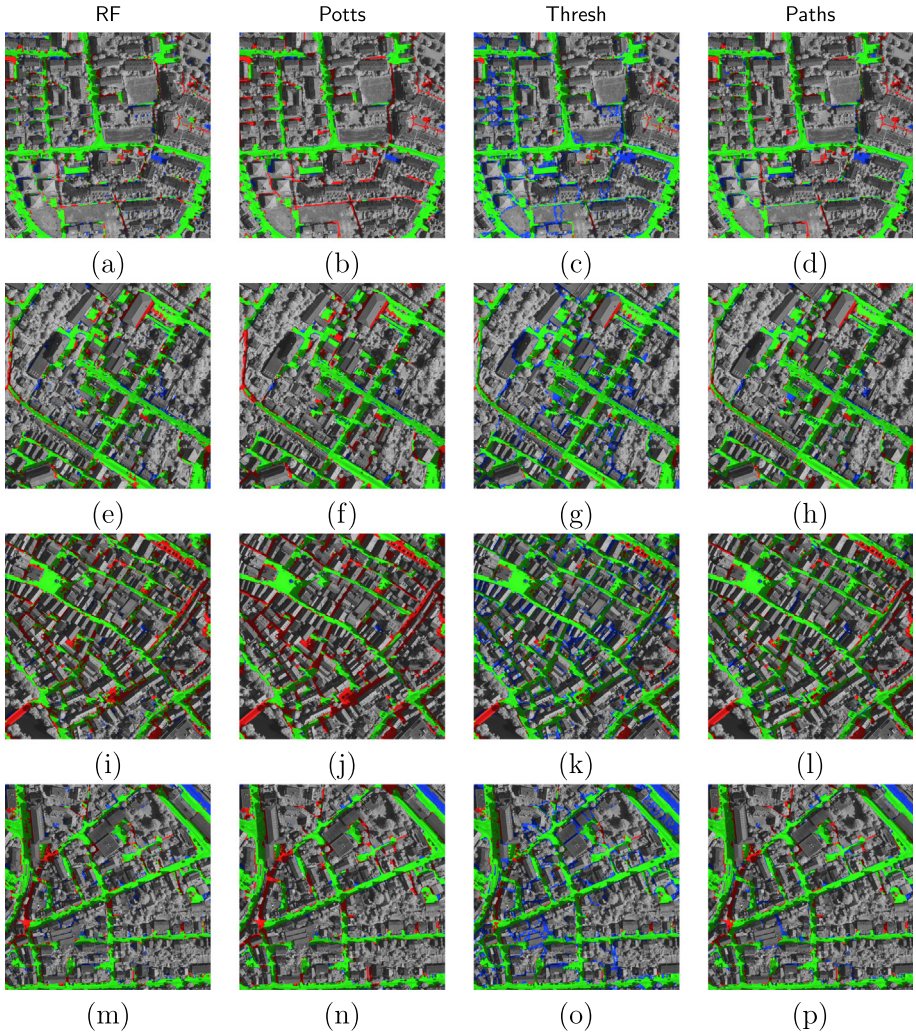

Fig. 3. Road networks extracted in three patches of the VAIHINGEN orthophoto mosaic. Green true positives, blue false positives, red false negatives.

shortest paths per start/end pair is set to $k = 4$ for both data sets resulting in a total of 6000 (GRAZ), respectively 16,000 (VAIHINGEN) paths per image. We give high weight $\lambda = 0.9$ to the unary energy and low weight $1 - \lambda = 0.1$ to the height gradient because the terrain height is already implicitly contained in the unaries as a feature. However, keeping the height gradient in the equation helps avoiding gross errors particularly if the unary classifier (i.e., Random Forest) is uncertain at buildings that have similar appearance (color, texture) as streets, but different height.

A bit of pairwise smoothing to remove noise inside the background and on wide roads proved beneficial, as long as the higher-order potentials dominate. We thus weight the binary term ψ_b with $\lambda_{bin} = 0.1$ and the higher-order term ψ_q with $\lambda_{path} = 1.0$.

4.1. Error measures

We report the standard metrics *quality*, *completeness*, and *correctness* (Wiedemann et al., 1998) commonly used in literature on road extraction (e.g., Laptev et al., 2000; Doucette et al., 2004; Mayer et al., 2006; Hu et al., 2007; Mnih and Hinton, 2010; Mnih and Hinton, 2012). Another popular metric is the κ-value to assess pixel-wise labeling accuracy. It quantifies how much the predicted labels differ from a random image with the same label counts, thus measuring the improvement over chance, whereas overall accuracy measures the improvement over a 100% incorrect result. Hence κ avoids biases due to uneven class distribution.[3]

A topologically correct network is essential for navigation purposes, but that fact is not well captured by the described metrics: even a small gap can lead to lengthy detours, while not having much impact on completeness. We thus use an additional metrics based on shortest paths between randomly sampled, correctly labeled road pixels, for which a road connection exists in ground truth. In case the extracted network has the correct topology, the predicted and actual path lengths should be (nearly) identical. Incorrect shortcuts result in too short paths (*2short*), incorrect gaps in the extracted network cause too long paths (*2long*), or they disconnect the network into disjoint parts with no connection at all (*noC*). We repeatedly sample pairs of seed pixels and compare the path lengths between the ground truth and our estimate until the percentages of all error types have converged. A tolerance of 5% of the path length is used to account for geometric uncertainty.

[3] $\kappa = \frac{N\sum_i c_{ii} - \sum_i (\sum_j c_{ij} \cdot \sum_j c_{ji})}{N^2 - \sum_i (\sum_j c_{ij} \cdot \sum_j c_{ji})}$, with c_{ij} the entries of the confusion matrix and N the number of pixels. E.g., for an image with 10% *road* and 90% *background* pixels a result without a single *road* pixel has 90% overall accuracy, but $\kappa = 0\%$.

4.2. Results

For quantitative analysis, we run 7-fold cross-validation for GRAZ with a 11/65 training/testing split. In case of VAIHINGEN we conduct 4-fold cross-validation with 4/12 training/testing split.

We compare the proposed method (Paths) to three baselines: the raw unaries (RF), only the standard contrast-sensitive pairwise Potts model without path cliques (Potts), and a rule-based version of the path prior, which directly thresholds the cost-weighted path density (Thresh). Evaluation results are given in Table 1, example results are depicted in Fig. 2 (GRAZ) and Fig. 3 (VAIHINGEN).

The two data sets have somewhat different characteristics. GRAZ has major roads and large blocks of buildings, whereas VAIHINGEN has narrower roads often shaded by trees or even completely occluded in the town center. Narrow and/or occluded roads (by trees or cast shadow) are typical situations where RF and Potts fail. RF misclassifies road parts with shadows (Figs. 2(a and e) and 3(e and i)) and Potts tends to smooth away narrow pieces of road (Figs. 2(b, f, j) and 3(b, j, n)) although we already gave a very low weight to the pairwise potentials.

Consequently Potts performs worst on VAIHINGEN with respect to *quality*, even below the raw unaries. On the GRAZ data set with its wider roads, the pixelwise accuracy of Potts is on par with RF, while its topological correctness (*correct*) is slightly higher. Our proposed method (Paths) resolves these situations much better and extracts many narrow and occluded portions of the road network, leading to 2–3% gains in κ and *quality* compared to the second best method. Particularly large improvements of 10–13% are achieved with respect to topological correctness (*correct*). However, due to its asymmetric nature it does produce some additional false positives, which is reflected in the number of *2short* paths.

To separate the performance of the clique sampler from the global CRF cost function we relabel superpixels covered by paths (that contain less than 10 consecutive superpixels with unary road likelihood below 0.5) as roads (Thresh). Clearly, minimum cost paths without a global CRF model do not improve the unary result but rather produce many false positives (see column Thresh in Figs. 2 and 3).

Nonetheless, Paths still exhibits a number of typical failures. Labeling errors in dead end roads cannot be corrected, unless by chance a seed is sampled at their very end. Moreover, minimum cost paths between different start/end nodes tend to use the same superpixels when passing through the same stretch of road or the same junction (those which have the highest road likelihood). As a result, superpixels with low road likelihood, often on the border of a wide road or crossing, may be missed by the path sampler and then cannot be recovered. Choosing a high number k of mutually exclusive shortest paths between two seeds reduces this effect, but k cannot be set arbitrarily high because this would force too many paths through the background and increase the number of false positives near road boundaries. If a scene contains both very narrow and very wide roads, like in the bottom row of Fig. 2, not all errors in the unaries (red parts center and right in Fig. 2(i)) can be corrected by Paths (remaining red parts Fig. 2(l)).

For a visual comparison to a state-of-the-art method, we extract roads in the EPFL-dataset of suburban RGB ortho-images, consisting of screenshots from Google Earth without height information (Türetken et al., 2013), see Fig. 4. Images and results of (Türetken et al., 2013, hereafter termed EPFL) were provided by the authors.

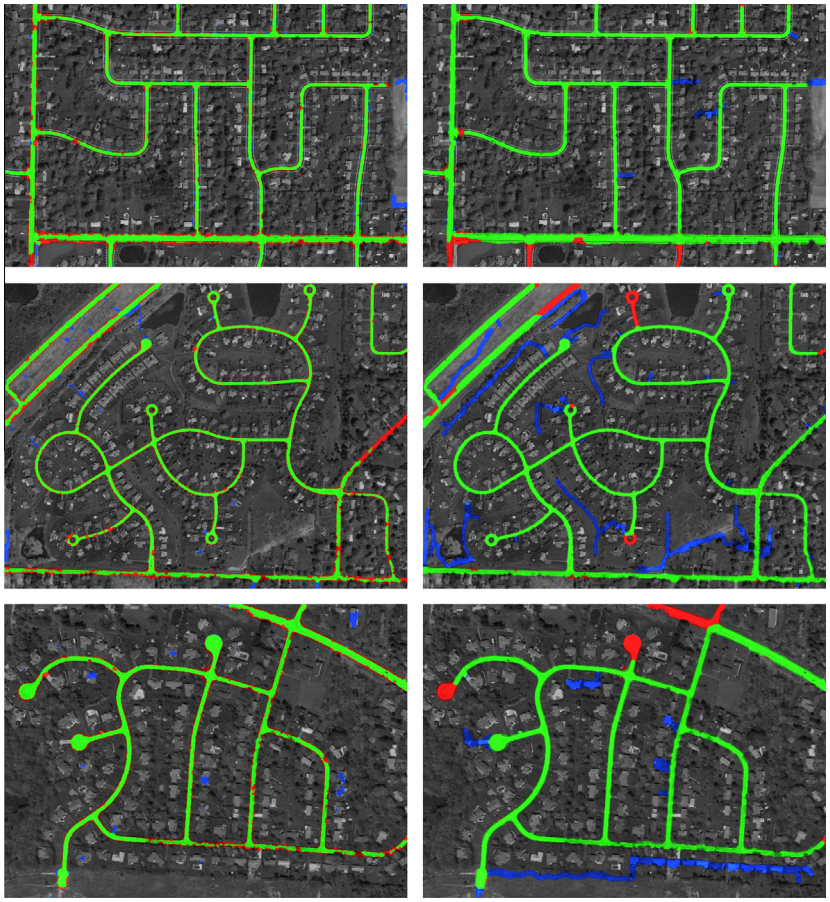

Fig. 4. Example results for EPFL data set obtained with our method (left) and the EPFL method (right). Green true positives, blue false positives, red false negatives.

Note that the EPFL results were achieved using gray-scale versions of the RGB ortho-images, whereas our approach needs color images as input. To adapt our approach to this data set without height channel, we drop the height from all steps of our method for the comparison: (i) the Random Forest uses no height features; (ii) path computation does not use the height gradient for the edge cost between adjacent superpixels; and (iii) the contrast-sensitive Potts potentials are computed without the height gradient. The visual comparison shows that our method produces significantly fewer gross errors, in particular it manages better to avoid false positives, see Fig. 4. For correct roads, EPFL does give smoother road boundaries than our method without post-processing. This could be expected, since on the one hand it operates on pixels rather than superpixels; and on the other hand it explicitly estimates the road width as part of the per-pixel "tubularity", which works well for unoccluded roads with nearly constant width, as depicted in the data set.

4.3. Parameter study

In order to quantify the sensitivity of the method to different parameter settings we perform a study in which we vary several crucial parameters. We test each parameter separately within a reasonable range and keep all others fixed. Tests are performed on a representative train/test split of VAIHINGEN, i.e. the one that is closest to the cross-validation average (cf. VAIH in Table 1). We test different parameters for path sampling, for path pruning, and for the truncated linear node weighting function. Additionally, we assess the impact of an alternative superpixel segmentation algorithm on the results. For all tests we record pixel-wise (κ) as well as topological error metrics (*quality* and the percentage of *correct* paths of the topological measure, cf. Section 4.1).

Two different parameters have to be set for path sampling (Section 3.2): the amount of start–end node pairs *sep* per image tile and the number k of mutually exclusive shortest paths per node pair. We evaluate $sep \in (1000, 2000, 3000, 4000, 5000)$ (Fig. 5(a)) and $k \in (1, 2, 4, 6, 8)$ (Fig. 5(b)). For *sep* this corresponds to a minimum number $k \cdot sep = 4 \cdot 1000 = 4000$ and maximum number of $k \cdot sep = 4 \cdot 5000 = 20,000$ paths, while the total path number for tests with k ranges from $k \cdot sep = 1 \cdot 4000 = 4000$ to $k \cdot sep = 8 \cdot 4000 = 32,000$. It turns out that different settings of *sep* (Fig. 5(a)) and k (Fig. 5(b)) do not significantly change κ and *quality* in general. However, they do have an influence on the percentage of topologically *correct* connections. Especially the extreme setting $k = 1$ misses important connections and significantly decreases topological correctness.

Pruning of unlikely paths before CRF inference (Section 3.2) is governed by a maximum allowed number *prune* of consecutive superpixels with unary road likelihood <0.5. Paths with longer stretches of background are discarded directly and not passed to inference. We test for a range of lengths $prune \in (5, 10, 15, 20, 25)$, see Fig. 6(a). The tests indicate that *prune* should be set $\geqslant 10$ in order not to drop too many promising paths at an early stage of the process.

The truncated linear weighting function (Section 3.3) for a node's clique membership needs lower and upper bounds (Eq. (4)). All nodes with $d(x_j, Q_i) \leqslant l \cdot \sigma_x(Q_i)$ receive w_{max}, whereas nodes with $d(x_j, Q_i) > u \cdot \sigma_x(Q_i)$ receive $w_i^j = 0$. We start from a very conservative setting of $(l, u) = (0.0, 0.5)$ and incrementally increase l and u in steps of 0.5 to $(l, u) = (2.0, 2.5)$, Fig. 6(b). While

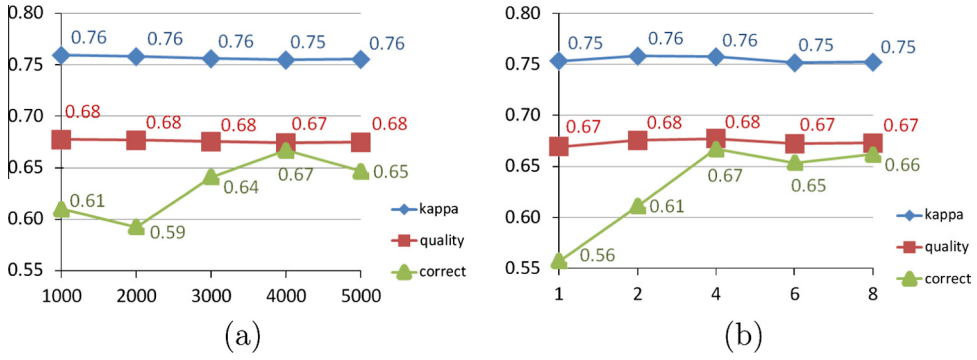

Fig. 5. Results of the parameter study (percentages on vertical axis, parameters on horizontal axis). (a) Path sampling: Evaluation of start–end node pairs *sep* (with k fixed to 4) and (b) evaluation of mutually exclusive paths k per node pair (with *sep* fixed to 4000). All numbers are percentages.

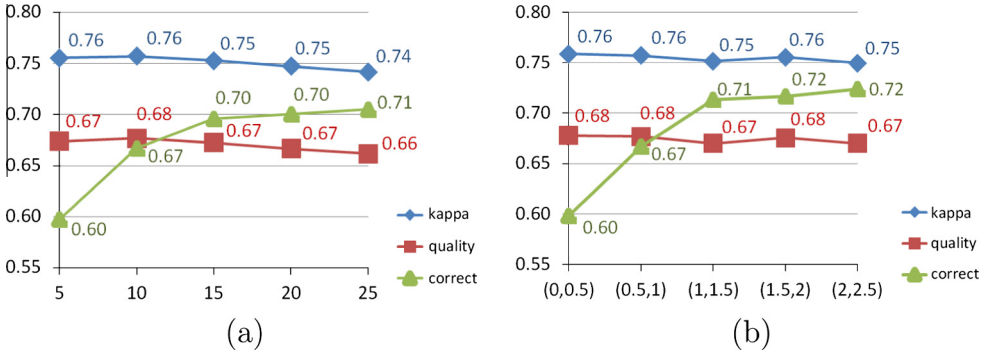

Fig. 6. Results of the parameter study (percentages on vertical axis, parameters on horizontal axis). (a) Path pruning: Evaluation of the maximum number *prune* of consecutive superpixels with unary road likelihood below 0.5 per path. (b) Node weighting: Evaluation of different settings of (l, u) for the truncated linear node weighting function.

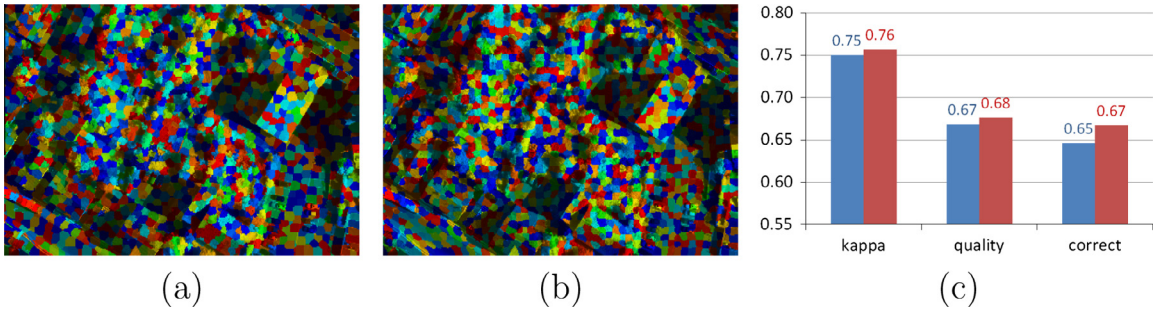

Fig. 7. Superpixel segmentations overlaid to image generated with the methods of (a) (Veksler et al., 2010), (b) SLIC (Achanta et al., 2012), and (c) comparison of quantitative results (SLIC blue, ours red) (percentages on vertical axis).

one could also vary the slope by changing l and u differently, we found this to have only minor influence. Too conservative values for (l, u) weaken the clique membership of too many superpixels and counteract the intended effect of the cliques, hence they yield inferior topological correctness. Larger values preserve the cliques better and unleash the potential of the model.

Finally, we compare our default way of generating superpixels, the energy-based approach of Veksler et al. (2010), with the SLIC method of Achanta et al. (2012), Fig. 7. The method of Veksler et al. (2010) has a small advantage of one or two percent points over SLIC for all measures. Note, however, that SLIC is faster to compute.

Overall, the proposed scheme is rather robust against different parameter settings. It should be noted, though, that some settings have considerable influence on the computation time. In particular, increasing the number of minimum cost paths slows the method down (how much depends strongly on the implementation and hardware used, since the computation trivially parallelises over different paths). On the other hand, higher values for *prune* do lead to a larger CRF with more large cliques, but that only marginally increases the run-time for inference.

5. Conclusions

We have proposed a model for long-range network structures in images, in which the network is seen as a collection of (partially overlapping) curvilinear paths. Putative pieces of the network are found by minimum cost path search and then pruned to a set that covers the road network, through MAP inference in a CRF. The paths correspond to higher-order cliques and their potentials are designed in such a way that they allow for efficient inference in spite of large clique sizes. In our experiments the proposed method outperforms several natural baselines, both in terms of labeling accuracy and w.r.t. topological correctness. Our model is not restricted to 2D network extraction, and its adaptation to 3D networks often encountered in medical image stacks appears straight-forward.

Still, a lot of prior knowledge remains unused and there are ample opportunities to improve network extraction. For example, crossings and T-junctions are not yet explicitly incorporated, but are strong evidence for networks. Simple star-shaped "junction cliques" in our experience only lead to small improvements (Wegner et al., 2013), so better ways to use evidence from junctions should be found. False negatives still remain on very narrow roads and near the boundaries of wide roads or crossings. These can be better recovered by (i) moving to a per-pixel classification (instead of superpixels) and (ii) by explicitly estimating the road width instead of tuning k, as proposed in our recent paper (Montoya et al., 2014). In Montoya et al. (2015) we show that our framework is flexible and can be extended to incorporate a higher-order prior for building extraction. While per-pixel classification improves numbers slightly, the superpixel approach presented here runs generally faster. Using superpixels instead of single pixels significantly speeds up path computation and inference, because the graph size (the number of nodes) per image is significantly reduced (from 1 million to ≈17,000).

Finally, minimum cost path sampling is at present done independently of the CRF. It appears feasible to include the path search directly into the CRF via pairwise potentials derived from the per-edge cost. Such a solution would certainly be more elegant and principled, albeit at the cost of more complicated inference.

References

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S., 2012. SLIC superpixels compared to state-of-the-art superpixel methods. IEEE Trans. Pattern Anal. Mach. Intell. 34 (11), 2274–2282.

Bajcsy, R., Tavakoli, M., 1976. Computer recognition of roads from satellite pictures. IEEE Trans. Syst. Man Cybernet. 6 (9), 623–637.

Bas, E., Erdogmus, D., 2011. Principal curves as skeletons of tubular objects. Neuroinformatics 9, 181–191.

Baumgartner, A., Hinz, S., Wiedemann, C., 2002. Efficient methods and interfaces for road tracking. In: ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. 34(3B), pp. 28–31.

Benmansour, F., Cohen, L.D., 2011. Tubular structure segmentation based on minimal path method and anisotropic enhancement. Int. J. Comput. Vis. 92, 192–210.

Boykov, Y., Jolly, M., 2001. Interactive graph cuts for optimal boundary and region segmentation of objects in n-d images. In: IEEE International Conference on Computer Vision.

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.

Burghouts, G.J., Geusebroek, J.-M., 2009. Material-specific adaptation of color invariant features. Pattern Recogn. Lett. 30 (3), 306–313.

Butenuth, M., Heipke, C., 2012. Network Snakes: graph-based object delineation with active contour models. Mach. Vis. Appl. 23, 91–109.

Chai, D., Förstner, W., Lafarge, F., 2013. Recovering line-networks in images by junction-point processes. In: IEEE Conference on Computer Vision and Pattern Recognition.

Dal Poz, A., Gallis, R., da Silva, J., 2010. Three-dimensional semiautomatic road extraction from a high-resolution aerial image by dynamic-programming optimization in the object space. IEEE Geosci. Remote Sens. Lett. 7 (4), 796–800.

Dal Poz, A., Gallis, R., da Silva, J., Martins, É.F.O., 2012. Object-space road extraction in rural areas using stereoscopic aerial images. IEEE Geosci. Remote Sens. Lett. 9 (4), 654–658.

Dijkstra, E., 1959. A note on two problems in connexion with graphs. Numer. Math. 1, 269–271.

Doucette, P., Agouris, P., Stefanidis, A., 2004. Automated road extraction from high resolution multispectral imagery. Photogram. Eng. Remote Sens. 70 (12), 1405–1416.

Fischler, M., Tenenbaum, J., Wolf, H., 1981. Detection of roads and linear structures in low-resolution aerial imagery using a multisource knowledge integration technique. Comput. Graph. Image Process. 15, 201–223.

Gerke, M., Butenuth, M., Heipke, C., Willrich, F., 2004. Graph-supported verification of road databases. ISPRS J. Photogram. Remote Sens. 58, 152–165.

Grote, A., Heipke, C., Rottensteiner, F., 2012. Road network extraction in suburban areas. Photogram. Rec. 27 (137), 8–28.

Gruen, A., Li, H., 1995. Road extraction from aerial and satellite images by dynamic programming. ISPRS J. Photogram. Remote Sens. 50 (4), 11–20.

Gruen, A., Li, H., 1997. Semi-automatic linear feature extraction by dynamic programming and LSB-snakes. Photogram. Eng. Remote Sens. 63 (8), 985–995.

Heipke, C., Mayer, H., Wiedemann, C., 1997. Evaluation of automatic road extraction. In: ISPRS Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. 32, pp. 47–56.
Helmholz, P., Becker, C., Breitkopf, U., Büschenfeld, T., Busch, A., Braun, C., Grünreich, D., Müller, S., Ostermann, J., Pahl, M., Rottensteiner, F., Vogt, K., Ziems, M., Heipke, C., 2012. Semi-automatic quality control of topographic data sets. Photogram. Eng. Remote Sens. 78 (9), 959–972.
Hinz, S., Baumgartner, A., 2003. Automatic extraction of urban road networks from multi-view aerial imagery. ISPRS J. Photogram. Remote Sens. 58, 83–98.
Hu, J., Razdan, A., Femiani, J.C., Cui, M., Wonka, P., 2007. Road network extraction and intersection detection from aerial images by tracking road footprints. IEEE Trans. Geosci. Remote Sens. 45 (12), 4144–4157.
Kohli, P., Ladicky, L., Torr, P.H.S., 2008. Robust higher order potentials for enforcing label consistency. In: IEEE Conference on Computer Vision and Pattern Recognition.
Kohli, P., Ladicky, L., Torr, P.H.S., 2009. Robust higher order potentials for enforcing label consistency. Int. J. Comput. Vis. 82 (3), 302–324.
Lacoste, C., Descombes, X., Zerubia, J., 2005. Point Processes for unsupervised line network extraction in remote sensing. IEEE Trans. Pattern Anal. Mach. Intell. 27 (10), 1568–1579.
Ladicky, L., Russell, C., Kohli, P., Torr, P.H., 2010. Graph cut based inference with co-occurrence statistics. In: European Conference on Computer Vision.
Lafarge, F., Gimel'farb, G., Descombes, X., 2010. Geometric feature extraction by a multimarked point process. IEEE Trans. Pattern Anal. Mach. Intell. 32 (9), 1597–1609.
Laptev, I., Mayer, H., Lindeberg, T., Eckstein, W., Steger, C., Baumgartner, A., 2000. Automatic extraction of roads from aerial images based on scale space and snakes. Mach. Vis. Appl. 12, 23–31.
Li, H., Yezzi, A., 2007. Vessels as 4-D curves: global minimal 4-D paths to extract 3-D tubular surfaces and centerlines. IEEE Trans. Med. Imag. 26 (9), 1213–1223.
Mayer, H., Hinz, S., Bacher, U., Baltsavias, E., 2006. A test of automatic road extraction approaches. In: ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. 36(3), pp. 209–214.
Mena, J., Malpica, J., 2005. An automatic method for road extraction in rural and semi-urban areas starting from high resolution satellite imagery. Pattern Recogn. Lett. 26, 1201–1220.
Miao, Z., Shi, W., Zhang, H., Wang, X., 2013. Road centerline extraction from high-resolution imagery based on shape features and multivariate adaptive regression splines. IEEE Geosci. Remote Sens. Lett. 10 (3), 583–587.
Mnih, V., Hinton, G.E., 2010. Learning to detect roads in high-resolution aerial images. In: European Conference on Computer Vision.
Mnih, V., Hinton, G.E., 2012. Learning to label aerial images from noisy data. In: International Conference on Machine Learning.
Montoya, J., Wegner, J., Ladicky, L., Schindler, K., 2014. Mind the gap: modeling local and global context in (road) networks. In: German Conference on Pattern Recognition (GCPR), pp. 212–223.
Montoya, J., Wegner, J., Ladicky, L., Schindler, K., 2015. Semantic segmentation of aerial images in urban areas with class-specific higher-order cliques. In: ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. II, pp. 127–133.
Movaghati, S., Moghaddamjoo, A., Tavakoli, A., 2010. Road extraction from satellite images using particle filtering and extended Kalman filtering. IEEE Trans. Geosci. Remote Sens. 48 (7), 2807–2817.
Poullis, C., You, S., 2010. Delineation and geometric modeling of road networks. ISPRS J. Photogram. Remote Sens. 65, 165–181.
Rother, C., Kohli, P., Feng, W., Jia, J., 2009. Minimizing sparse higher order energy functions of discrete variables. In: IEEE Conference on Computer Vision and Pattern Recognition.
Stoica, R., Descombes, X., Zerubia, J., 2004. A gibbs point process for road extraction from remotely sensed images. Int. J. Comput. Vis. 57 (2), 121–136.
Türetken, E., Benmansour, F., Andres, B., Pfister, H., Fua, P., 2013. Reconstructing loopy curvilinear structures using integer programming. In: IEEE Conference on Computer Vision and Pattern Recognition.
Türetken, E., Benmansour, F., Fua, P., 2012. Automated reconstruction of tree structures using path classifiers and mixed integer programming. In: IEEE Conference on Computer Vision and Pattern Recognition.
Türetken, E., González, G., Blum, C., Fua, P., 2011. Automated reconstruction of dendritic and axonal trees by global optimization with geometric priors. Neuroinformatics 9, 279–302.
Ünsalan, C., Sirmacek, B., 2012. Road network detection using probabilistic and graph theoretical methods. IEEE Trans. Geosci. Remote Sens. 50 (11), 4441–4453.
Veksler, O., Boykov, Y., Mehrani, P., 2010. Superpixels and supervoxels in an energy optimization framework. In: European Conference on Computer Vision.
Vicente, S., Kolmogorov, V., Rother, C., 2008. Graph cut based image segmentation with connectivity priors. In: CVPR'08.
Vincente, S., Kolmogorov, V., Rother, C., 2008. Graph cut based image segmentation with connectivity priors. In: IEEE Conference on Computer Vision and Pattern Recognition.
Vosselman, G., de Knecht, J., 1995. Road tracing by profile matching and Kalman filtering. In: Automatic Extraction of Man-Made Objects from Aerial and Space Images. Birkhäuser Verlag, Basel, pp. 265–274.
Wang, Y., Narayanaswamy, A., Roysam, B., 2011. Novel 4-D open-curve active contour and curve completion approach for automated tree structure extraction. In: IEEE Conference on Computer Vision and Pattern Recognition.
Wegner, J.D., Montoya-Zegarra, J.A., Schindler, K., 2013. A higher-order CRF model for road network extraction. In: IEEE Conference on Computer Vision and Pattern Recognition.
Wiedemann, C., Heipke, C., Mayer, H., Jamet, O., 1998. Empirical evaluation of automatically extracted road axes. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops.
Winn, J., Criminisi, A., Minka, T., 2005. Object categorization by learned universal visual dictionary. In: IEEE Conference on Computer Vision and Pattern Recognition.
Yao, J., Fidler, S., Urtasun, R., 2012. Describing the scene as a whole: joint object detection, scene classification and semantic classification. In: IEEE Conference on Computer Vision and Pattern Recognition.
Youn, J., Bethel, J.S., Mikhail, E.M., Lee, C., 2008. Extracting urban road networks from high-resolution true orthoimage and lidar. Photogram. Eng. Remote Sens. 74 (2), 227–237.
Zhang, C., 2004. Towards an operational system for automated updating of road databases by integration of imagery and geodata. ISPRS J. Photogram. Remote Sens. 58, 166–186.
Zhao, T., Xie, J., Amat, F., Clack, N., Ahammad, P., Peng, H., Long, F., Myers, E., 2011. Automated reconstruction of neuronal morphology based on local geometrical and global structural models. Neuroinformatics 9, 247–261.