On the evaluation of higher-order cliques for road network extraction

Javier A. Montoya-Zegarra, Jan D. Wegner, L’ubor Ladický, Konrad Schindler
ETH Zurich, Switzerland

Abstract—The automatic extraction of road networks is an interesting and challenging task. In spite of significant research efforts this problem remains largely open. In our work we attempt to leverage context at two different levels to extract accurate and topologically correct road networks. Local context, in the form of powerful features extracted from large neighborhoods, exploits the layout of road pixels and their co-occurrence with visual patterns along the roads. Global context enforces the connectivity of roads in a network, by grouping individual pixels into longer road segments, modeled as large higher-order cliques. Here, we evaluate different ways of defining these cliques. It turns out that, with modern probabilistic inference techniques, using a smaller number of very large cliques is more efficient than splitting them into a larger number of shorter segments.

I. INTRODUCTION

The automatic extraction of road networks from aerial images is an important and active research topic in the remote sensing community. Road networks are needed for a wide range of applications like mapping, urban planning, navigation systems, etc.. Traditionally, road networks are digitized manually, a time-consuming and tedious task.

There is still no robust method to automatically extract roads from overhead images, in spite of four decades of research, dating back to at least [1]. The main difficulties are (i) the large intra-class variability and (ii) the presence of clutter, caused among others by shadows of trees and buildings, cars, and diverse street furniture.

In our work, we propose a method that incorporates local and global information to model road networks as a collection of road segments. Local context allows one to exploit the visual patterns formed by roads and the objects that surround them (such as for example buildings on either side of the road, or cars parked along its border). Global context helps to assemble road segments to a topologically complete and correct network.

Our method starts by learning a binary classifier that predicts pixel-wise likelihoods for road and background classes. To do so, a set of rich and dense appearance features is extracted from large spatial windows. On top of that, a second classifier is trained to estimate the likelihood that a road of a specific width appears. This results in a $$(x, y, width)$$-volume of road likelihoods (in practice the width is also discretized). Next, following a recover-and-select strategy, we sample a large set of road candidates from the volume (recover-phase). The sampling aims at high recall, hence it generates an overcomplete set that includes a number of spurious road segments.

The set is then pruned to an optimal subset (select-phase) by MAP-inference in a higher-order CRF. More precisely, the pixels of each sampled candidate form a large clique. If within a clique there is enough road evidence, then it is encouraged to label as many of its members as possible as road. If, on the other hand, the cumulative road evidence is low, then it is discouraged to label even some of the member pixels as road.

II. RELATED WORK

Since early attempts such as [1] many approaches have been proposed to automatically extract roads for rural [2], semi-rural [3], or urban areas [4–7].

Heuristic bottom-up grouping methods, e.g. [8, 9], aim at detecting and stitching together a set of road segments detected by means of low-level image processing techniques, such as edge detectors [10, 11], morphological operators [12], or road masks [13]. Drawbacks of this sort of methods are that many parameters need to be tuned again for each new scene and that errors propagate throughout the (greedy) stages of the model.

Marked Point Processes (MPPs) [14, 15], define potential road elements, such as short line segments, as the basic variables in a probabilistic framework, which then allows one to impose high-level topological constraints. To find likely configurations, one has to resort to stochastic optimisation with Reversible-Jump Markov Chain Monte Carlo (RJMC) sampling or similar. Hence, the inference is usually expensive and somewhat brittle.

Another way to model road networks consists in connecting points with high road likelihood via minimum cost paths. Beyond road extraction such approaches are common in the medical imaging domain [16–18] to reconstruct vessel trees or neurons.

Our proposed method combines the minimum cost path idea, used to find road candidates, with a global CRF model to obtain consistent results on the pixel level. The most related works are [6, 19]. Compared to our earlier work [19], the method shown here is not restricted to straight segments, and includes a notion of road width [7]. Also, the present formulation avoids a foreground bias and is better able to remove false positives from the unaries.

In [6], a set of multi-scale local tubularity scores are first computed. Next, seed pixels with high foreground scores are connected to obtain curvi-linear structures that are merged together into a graph.
An optimal subgraph is then found by Integer Programming. The method focuses on the road center-lines rather than on full pixel labeling.

III. MODEL

The overview of our proposed framework is depicted in Figure 1. We model the road network as the union of elongated segments (termed \textit{paths}) and large compact regions (termed \textit{blobs}). Both \textit{paths} and \textit{blobs} come with an associated scale, i.e., we do not only represent road center-lines, but also the local width of the road. We start by training a pixel-wise binary classifier (\textit{road vs. non-road}) that is trained over large spatial neighborhoods using rich appearance features. The large spatial neighborhoods allow us to encode the local co-occurrence of objects.

At a second stage, we train a classifier on the obtained road likelihoods using statistics about their local distributions. This generates a \((x,y,width)\)-volume of road likelihoods in which the expected road widths are discretized across the scales. Note that, this second round of classification has the benefit of correcting and smoothing out the road scores obtained from the context-classifier. Furthermore, estimating the road widths rather than the center-lines only generates a richer model of the road network. In order to generate our road candidates, we sample a set of putative \textit{paths} and \textit{blobs} from the resulting volume of road likelihoods. \textit{Paths} maximize the cumulative road evidence and are sampled by connecting random pairs of road-like pixels. \textit{Blobs} are generated by scanning the top scales and by looking for the maximal cumulative road likelihood.

In order to cover as much as possible of the road network, we follow a recover-and-select strategy: an over-complete set of road \textit{candidates} is generated that covers as much of the underlying road network as possible. Subsequently, the subset of road candidates that best explains the road evidence is selected within an energy minimization framework. Each candidate is mapped to a higher-order clique of a CRF with a robust \(P^m\)-Potts potential that encourages clique members to take on the same label. This CRF can be solved to global optimality with the mincost/maxflow algorithm.

A. Local context-aware road scores

In order to obtain the pixel-wise binary road scores (\textit{road vs. non-road}), we adopt the multi-feature extension [20] of Texton-boost [21]. A set of appearance features (SIFT [22], local ternary patterns [23], textons [24], and self-similarity [25]) are densely extracted. Each feature type is (soft-)quantized into a corresponding dictionary of 512 visual words. To include context, a large spatial window \((160 \times 160 \) pixels) is centered over each pixel and word histograms are extracted over a fixed set of 200 rectangles \((4 \times 4 \text{ to } 80 \times 80 \) pixels). The extracted histograms are then concatenated into a single feature vector. After classification with Boosting, the class scores are converted to road and background probabilities \(S = (S_{\text{road}}, S_{\text{bg}})\) with a sigmoid mapping.

B. 3D Road likelihoods

To predict the presence of a road of specific width, we use a second round of classification. The road scores obtained in the previous step from the context classifier are mapped to a scale-space representation by applying a scale-normalized Laplacian of Gaussian [26]. Each scale is associated with a discrete road width. Next, we compute a set of statistical measures such as: mean, median, and standard deviation on the pixel-wise LoG responses to train a Random Forest on ground-truth road widths. This allows us to predict the local likelihood for each possible road width resulting in a volume \(L(x,y,width)\) of road likelihoods.

C. Sampling of road candidates

To extract a set of representative road candidates, we sample elongated segments (termed \textit{paths}) and large compact regions (termed \textit{blobs}) from the road likelihood volume \(L(x,y,width)\). At this stage, we aim for high completeness: the union of all sampled road candidates should cover as much as possible of the road network even at the cost of low correctness.

Elongated \textit{paths} are sampled by connecting random pairs of pixels with enough road likelihood in \(L(x,y,width)\) inside the road likelihood volume. These start and end pixels are connected with the Fast-Marching (FM) 3D algorithm [27]. It is important to notice that the sampled paths have an explicit road width assigned at each pixel, which changes smoothly. The smooth transition of road widths is warranted by the FM energy function which prefers smooth transitions in the scale cube, rather than abrupt changes. Compact \textit{blob} regions are added to cover large road regions without a clear direction such as parking-places, roundabouts, etc. For that purpose, we scan the top scales (above the maximum road width) of the \(L(x,y,width)\) volume for local maxima and perform non-maxima suppression locally.

D. Efficient global selection of road candidates

During the last step, we select the subset of road candidates that best explain the road evidence. Since some of the previously sampled road candidates will overlap or (partially) pass through background, it is necessary to add a correction step that slightly modifies candidates where required, but prefers changing them as little as possible.

This correction and selection step is formulated as an energy minimization problem \(E = \sum_u E_u(x_u) + \sum_Q E_p(Q)\) over all pixels \(x_u\) of the image. More precisely, each of the sampled road candidates is mapped to a clique \(Q\), composed by its corresponding member pixels. The pixel-wise unarys \(E_u(x_u) = -\log(S)\) are derived from the negative log-likelihoods of the road scores (from the original classifier). The higher-order potentials \(E_p(Q) = \min(\alpha, N_k \cdot \frac{|Q|}{\gamma^\beta} + \beta)\) are robust \(P^n\)-Potts \([28]\), which encourage all member pixels to have the same label. The \(\{\alpha, \beta, \gamma\}\) parameters define the linear truncated function that establishes how the energy increases as the number of pixels deviate from the dominant label, and \(N_k\) is the count of pixels that take on label \(k\). If within a clique \(Q\) there is enough road evidence, then the clique members (pixels) are drawn to the road class. As a result, gaps of \textit{false negatives} traversed by a clique are corrected. On the other hand, if there is not enough total road evidence, then the clique members are drawn to the background. In this way \textit{false positives} not connected to the road network (e.g., inner courtyards) are removed.
IV. EXPERIMENTAL RESULTS

We evaluate our experiments on the Graz data set which comprises 64 aerial images covering an urban region. Buildings are big and usually contain inner-yards or inner parking places. Road networks are mainly formed by major avenues connected with secondary streets. The roads change slowly both in width and in curvature. Furthermore, some objects appear on roads such as trees and cars. During the experiments the following split was considered: 30 images were used for training, 12 for validation, and 22 for testing.

As quality measures we evaluate the classification accuracy at pixel-level (F1-score) and the topology of the extracted road networks (with the metric of [19]). Given an extracted road network and a reference network (ground-truth) we sample aleatory Dijkstra paths between the road pixels that are present in both images. We then measure what fraction of the connecting paths have the correct length within 5% tolerance and are respectively too short (2short), too long (2long), or completely infeasible (noC).

We start by first comparing the achieved results of our method and three baselines. As first baseline (Winn+) we apply the super-pixel-based method presented in [19] that combines likelihoods from a Random Forest classifier (unary term) with an asymmetric higher-order $P^n$-Potts potential applied to a graph of straight cliques and junction cliques (prior term). The second baseline (Context) corresponds to the raw classification obtained from the context-aware unaries (Subsection III-A). The third baseline (RawPath) corresponds to the promising road candidates that were sampled from the volume of road-likelihoods (Subsection III-C). This third baseline helps us to separate the effect of our road sampling scheme from the effect of the subsequent CRF inference. As it can be noticed our method (Ours: full clique) outperforms the two baselines in both F1-score and topological measure.

We also evaluated the effect of the size of the higher-order cliques and partitioned them into a set of overlapping “sub-cliques”. Each sub-clique will then form a new higher-order clique and the set of new cliques is independently evaluated during inference. Additional parameters of our approach include therefore the sub-clique length (in pixel units) and their stride/overlap length. The cliques were partitioned over their center-line. In our evaluations, we considered different values and investigated their influence on the final performance of our approach. We considered different clique lengths: 150, 300, 600 and for each length, we used different strides corresponding to $\frac{1}{2}$, $\frac{1}{4}$, and $\frac{1}{8}$ of the clique length. For each of the possible combinations, we also recorded the average CPU-time used during inference. In Figure 2, the last two pairs of bars compare the accuracy of reducing the size of the cliques against using full-length cliques. As it can be shown, there is a slight improvement in the F1-score and topological measure when using full-length cliques. The big gain however is in that the processing time gets reduced by 30 min by using larger cliques. Our approach also outperforms the other three baselines including the recent work of [19] (Winn+). Furthermore, Figure 3 summarizes the F1-score and topological measure obtained by different clique sizes given that the stride was fixed (1/4 of the clique length). We can notice from the results that for the three different configurations the accuracy measures are pretty stable. However, what is important to note is that by using shorter cliques (e.g. 150) the CPU-time used during inference almost doubles the time needed for larger cliques (e.g. 600). This suggests that larger cliques should be preferred over shorter ones as the accuracy measures are equivalent, however, at the cost of having a more efficient computation.

Finally, to evaluate the impact of the stride percentage, we fixed the clique length to 600 as suggested from the previous result. In Figure 4, we report results by using stride values of: 150, 300, and 450. It can be seen that as the size of the stride increases, the processing time gets reduced and topological measures stay stable. These results suggest that one should not recoil of using long-range cliques although they correlate many variables. In fact, it turns out that they decrease computation time. The reasons are that (i) it is rather the number of cliques than the number of variables per clique that lead to additional computation time (for higher-order $P^n$-Potts potentials) and (ii) long-range cliques explain long-range context in a less redundant way than short, overlapping cliques do.
In this work, we have investigated different ways to model road context. Our approach leverages context at two different levels. Local context is achieved by using a set of powerful features extracted from large image neighborhoods. The global context enforces the connectivity of longer road segments.

After evaluating different strategies of defining these cliques it turns out that long-range cliques are less redundant than shorter segments and help to reduce computational time for inference while F1-score and topological measures remain stable.

**V. CONCLUSIONS AND FUTURE WORK**

In this work, we have investigated different ways to model road context. Our approach leverages context at two different levels. Local context is achieved by using a set of powerful features extracted from large image neighborhoods. The global context enforces the connectivity of longer road segments which are modeled as higher-order cliques within a Conditional Random Field. After evaluating different strategies of defining these cliques it turns out that long-range cliques are less redundant than shorter segments and help to reduce computational time for inference while F1-score and topological measures remain stable.

**REFERENCES**


