https://doi.org/10.2352/EI.2024.36.17.AVM-114

Fusing segmentation and domain-knowledge model to extract intersection topologies from aerial/terrestrial orthographic images

Julien A. Vijverberg[1,2]***, Bart Beers***[1]***; Peter H. N. de With***[2]***;***
[1]***Cyclomedia Technology B.V.; Zaltbommel, The Netherlands***
[2]***Eindhoven University of Technology, SPS-VCA group of Electr. Eng.; Eindhoven, The Netherlands***

Abstract

Automated extraction of intersection topologies from aerial and street-level images is relevant for Smart City traffic-control and safety applications. The intersection topology is expressed in the amount of approach lanes, the crossing (conflict) area, and the availability of painted striping for guidance and road delineation. Segmentation of road surface and other basic information can be obtained with 80% score or higher, but the segmentation and modeling of intersections is much more complex, due to multiple lanes in various directions and occlusion of the painted stripings. This paper addresses this complicated problem by proposing a dualistic channel model featuring direct segmentation and involving domain knowledge. These channels are developing specific features such as drive lines and lane information based on painted striping, which are filtered and then fused to determine an intersection-topology model. The algorithms and models are evaluated with two datasets, a large mixture of highway and urban intersections and a smaller dataset with intersections only. Experiments with measuring the GEO metric show that the proposed late-fusion system increases the recall score with 4–7 percentage points. This recall gain is consistent for using either aerial imagery or a mixture of aerial and street-level orthographic image data. The obtained recall for intersections is much lower than for highway data because of the complexity, occlusions by trees and the small amount of annotated intersections. Future work should aim at consolidating this model improvement at a higher recall level with more annotated data on intersections.

1. Introduction

Despite the promised reduction of traffic by the advent of remote working, self-driving cars and environmental awareness, the pressure on (urban) traffic infrastructure is still growing. This compromises both travel times and human safety and leads to increased environmental issues. To increase safety and effective traffic flow, city management has a growing interest in Smart City concepts, where people and traffic management are carefully monitored for optimizing travelling conditions and improving traffic flow. Essential elements of road infrastructures are the intersections, which are crucial in supplying traffic to, through and from the cities and which are attractive points for measuring traffic throughput. Better insights on the structure of intersections and their traffic throughput can help to improve any of the abovementioned issues. Accurate information on intersection topologies can be readily combined with video cameras and induction loop measurements, so that it is important to automatically derive the topology of the intersection. This topology can then be combined with local measurements to provide the information required for optimizing traffic flow and safety.

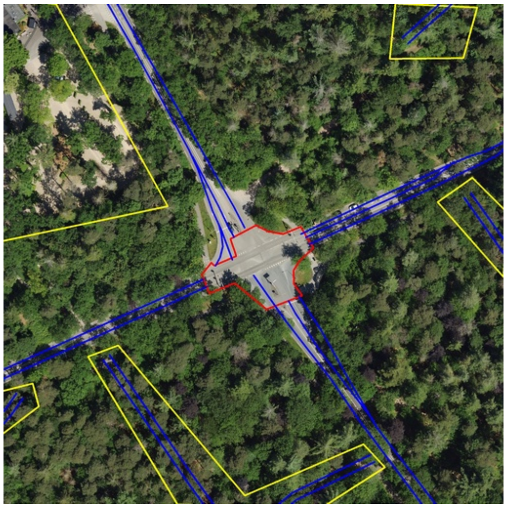


Figure 1. *Example of annotated drive lines (blue), the conflict area (red) and ignore zones (yellow).*

This paper attempts to derive the topology of an intersection in automated form using orthographic aerial and terrestrial images. The research aims to find this topology in either one of the two data sources, of which the images are captured by mapping companies. For the purpose of mapping the world, mapping companies often opt to record during favorable capturing conditions: no rain, no snow and with sufficient daylight. However, the dataset images are captured every 5 m, which means approximately 3 frames per second at typical urban cruising speed of 50 km/h, so that the vehicle and the camera are rarely in a static position and thus always moving. Because of the low frame rate and the moving position of the camera, traffic flow cannot be measured and should be obtained from other cameras and sources. This implies to investigate methods for detecting the road infrastructure, which can be derived from lane detection and line paintings (*i.e.* those lines which define lane boundaries) and road surfaces (*i.e.* the drive-able road surface, not parking areas) to ultimately identify the intersection topology.

Automated detection of road infrastructure has been intensively explored over the past decade. Some lane detection methods [1], [2] propose to detect painted striping and drive lines from the vehicle-mounted cameras for autonomous vehicles. This type of work mostly focuses on forward-looking cameras, while using left, right and backward-looking cameras which should give more information. Furthermore, these methods mainly detect lanes for

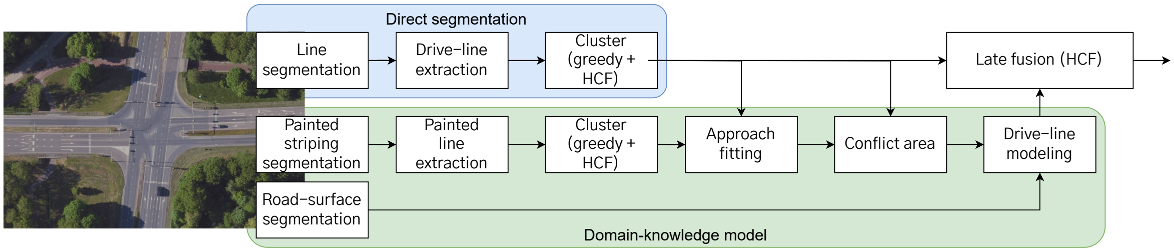


Figure 2. *Overview of the intersection topology modeling, depicting the individual segmentation functions and domain-knowledge functions. The final result provides a geo-referenced topology model of the intersection.*

the current vehicle position, but do not combine detections across frames. He and Balakrishnan[3] are training a U-Net that can detect drive lines directly. Although they gracefully published their code and data, we could not reproduce the same results as the authors reported. Liao *et al.* [4] propose MapTR, a transformer model to detect painted striping lines in an end-to-end fashion, but do not yet extract drive lines. Xu *et al.* [5] have recently presented CenterLineDet which does detect drive lines and attempts to connect consecutive regions.

Our previous work [6] has derived the drive lines from the painted striping and road edges, but this results into several rule-based post-processing steps and does not account for merging and splitting lanes. This method offers insufficient performance because many intersections in the dataset do not have painted striping, while edge-of-road features have been found to be very noisy. Moreover, the approach does not solve branching and merging lanes. In [7], we have used segmentation networks to find the drive lines directly by learning line segmentation and line extraction. Since this initial work was based on aerial imagery, one of the failure cases included trees covering the roads in aerial views. This problem can be circumvented by using images captured by terrestrial vehicles.

The first problem that is addressed in this work, is the occlusion of the road in aerial views. More specifically, we will look into the usage of Street-Ortho images [8], which are orthographic, top-view images generated from a terrestrial vehicle to avoid occlusions from the air. The second problem is based on the possible combination of domain knowledge with extra information resulting from lane striping. The domain knowledge can be extracted from images and involves information on the average width of lanes, road surface identification and special striping for crossings. This information is combined with the direct extraction of drive lines to construct a better model of the intersection topology. The aforementioned problem aspects have not been addressed in other work thus far.

The contributions of this paper include two aspects. (1) A method for late-fusion of the drive-line data by direct segmentation within the images and the domain knowledge defined by conventional indications. (2) The combined use of aerial and Street-Ortho data in order to improve drive-line extraction performance.

The remainder of this paper is organized as follows. Section first provides an overview of the extraction pipeline and zooms in on its stages of particular interest in subsequent sections. Section presents the experimental results. Section discusses the results and presents conclusions.

2. Method

This section describes the employed method to extract drive lines for intersection topologies. The first subsection describes an overview of the complete method. After a subsection on the data preparation, we describe the two channels separately, *i.e.* direct segmentation and the domain-knowledge model, but mostly report on the improvement details.

2.1 Overview of complete topology derivation

Figure 2 depicts a block diagram with the flow of the information-extraction model. The diagram shows a dual-channel approach, where the upper channel involves direct segmentation functions and the lower channel combines domain knowledge of the topology with striping information. These two channels are combined at the end of the modeling trajectory. The inputs are the geographical location of the intersection and the aerial and/or Street-Ortho image(s). From the image(s), binary segmentation masks are extracted from which lines are extracted. These painted striping and drive-line lines are merged with Highest Confidence First (HCF) algorithm, *i.e.* a greedy, hierarchical algorithm reported in [6]. The separate approaches of the intersection are fitted on these merged lines. Next, the conflict area, *i.e.* the area where drive lines from different approaches cross each other, is estimated, in order to filter the painted striping lines and obtain reliable results. These filtered painted striping lines are then used to extract drive lines [6]. Finally, the drive lines extracted from the drive-line mask and those derived from the directly detected painted striping are combined via the above-mentioned greedy merging algorithm (HCF). The final result for the intersection is the intersection model, containing the number of approach lanes, the type of crossing and the total complexity of the intersection. Finally, the intersection model is converted to a geo-referenced model to enable integration with other downstream functions or automation.

2.2 Data preparation - Blending and artifacts

The data preparation solely consists of blending of the Street-Ortho imagery with the aerial imagery. Unfortunately, Street-Ortho imagery cannot be produced at all desired locations where the research is conducted. This limitation mainly affects the highway scenes, which are only used for training in this paper. However, there are a couple of intersections with entries or exits to the highway, that miss the imagery on one or a few approaches of the intersection. In order to exploit this data maximally, the Street-Ortho image is overlaid on top of the corresponding aerial image to construct a proper image. The alignment of the images depends solely on the accuracy of the positioning and calibration of the aerial and terrestrial cameras: No further processing is performed to improve the alignment. Figure 3 shows visual examples of aerial and blended imagery. When comparing Scene 1 aerial to its blended counterpart, it is readily clear that in the blended image the roads under the trees are visible. However, there are some visual artifacts in the Street-Ortho image which occur less in the aerial data, including artifacts around moving vehicles and the illumination differences caused by driving variable trajectories across the intersection at different times of the day. Scene 2 also shows some Street-Ortho artifacts on the right approach due to illumination differences. A final note of interest is that the painted striping in the top approach in Scene 1 is still poorly visible due to poor illumination conditions under the trees.

2.3 Direct Segmentation

The direct segmentation method extracts drive lines directly from the image. It consists of three steps: segmentation, drive-line extraction and drive-line clustering with HCF.

A Segmentation

The segmentation function-modeling is based on U-Net [9] for each feature type. An initial comparison with the more modern SWIN transformer [10] for drive-line segmentation has confirmed the selection of U-Net for the segmentation of the desired line information.

Data augmentation steps of the training data include horizontal and vertical flipping, adding Gaussian noise, contrast variations, 180 degrees rotation. It should be noticed that, likely due to the consistent quality of the professional source images, the data augmentation techniques like sharpening, blurring and adding noise are reduced in probability of occurrence by a factor 3 to 20 compared to the typical levels reported in literature. However, the augmentation for the drive-line segmentation has been modified at one particular point. We have found empirically that randomly cropped images should contain a minimum amount of ground-truth information (currently at least 1%), since the training data is rather unbalanced due to the limited thickness of drive-lines and painted striping lines in the images.

In the data, conflict areas and ignore zones are specifically defined and annotated. The conflict area of an intersection is the region where drive lines are crossing on that intersection. An ignore zone is an area which is ignored for the analysis for avoiding uncertainties in decision making. Any pixel within a conflict area or ignore zone does not contribute to the loss function or validation scores during training. Ignore zones are defined in areas where typically several misclassifications of pixels occur and which do not contribute to the intersection-topology modeling, such as residential areas. Finally, we have employed individual networks for automated analysis of the blended and aerial data.

B Drive-line extraction

Initial line extraction from the drive-line mask, is now performed using the Hough transform on tiles of 64×64 pixels, instead of using the other methods investigated in previous work [7]. In the preliminary studies of the work, we have found that there is yet insufficient ground-truth annotations for painted striping to train an LCNN [11]. Furthermore, the proposed work on NEFI [12] appears to yield unpredictable and poor results on our segmentation masks. Employing the Hough transform on small-sized tiles, such as 64×64 pixels, has resulted in better performance, despite that it is limited to only detecting straight lines.

2.4 Domain-knowledge model

The domain-knowledge model adds general, prior knowledge about intersections with the objective to increase the accuracy. The model uses the same segmentation model and the associated training as described above, to create segmentation masks of the painted striping and the road surface. For extracting painted striping lines, line extraction (going from pixels to lines) and merging with the HCF algorithm is executed as in the direct segmentation branch of the drive lines. Similar to our earlier work [6], the approaches and conflict areas are fitted from the painted striping lines. The following subsections describe how drive lines are added to these algorithms and how the novel extensions are implemented in the overall framework. Finally, the domain-knowledge model extracts the drive lines as the lines between consecutive painted striping lines.

A. Approach fitting

Using the painted striping lines after filtering, the purpose of the succeeding approach fitting is to find the approaching lanes of the intersections.

A RanSaC-like approach is adopted in which the algorithm randomly selects one pivot line, which will be considered a valid painted striping line for an approach lane in the current iteration. All lines on that side of the intersection with approximately the same angle as the pivot line are projected onto a specific line, that is orthogonal to the pivot line at the intersection. The projections can be used in a 1D clustering problem to determine which lines are close to the pivot line in terms of distance. Evidently, close lines are clustered. The support for the proposed approach lane is then computed by summing the lengths along the trajectories of all lines in sufficiently large clusters. If the support is insufficient, the algorithm randomly selects another line as pivot line. Assuming there was sufficient support, the algorithm randomly selects another line from the remaining, non-clustered lines and repeats the procedure for a maximum amount of iterations.

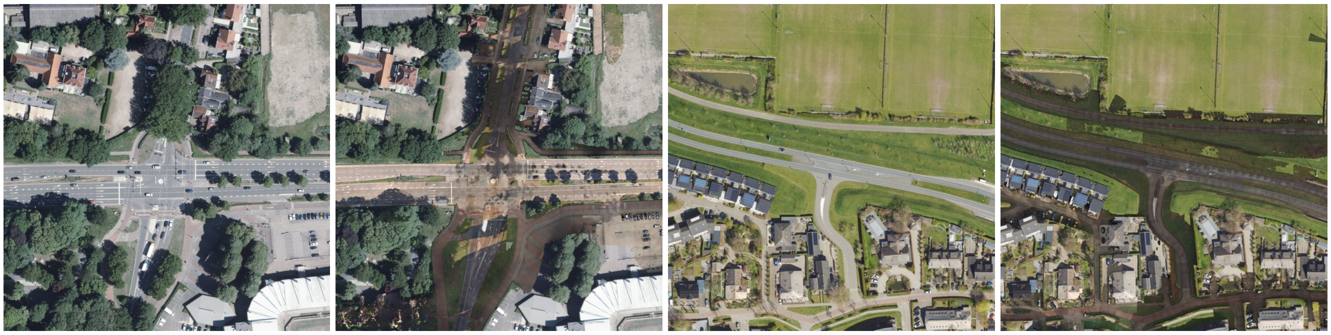
The algorithm is executed in parallel both for painted striping lines and for drive lines. The use of drive lines in this algorithm is similar to the previous approach: The drive lines are clustered around any kind of pivot line and other candidate drive lines are clustered using the projection method, optimizing the support, etc.

B. Conflict area estimation

The method of finding the conflict area is based on the intersection points of the painted striping lines from different approach lanes. The conflict area is defined by finding the hull around these intersection points. By mixing drive lines and painted striping lines, extra intersection points can be found by mixed crossings of drive lines and painted striping lines.

C. Late fusion HCF

Late fusion is the final stage of the modeling process and combines the output of the direct segmentation channel and the domain-knowledge model channel. In the fusion, also the HCF algorithm is applied, but with a different setting of the algorithm: the weights are tuned more towards clustering longer line strings



(a) Scene 1 - Aerial (b) Scene 1 - Blended (c) Scene 2 - Aerial (d) Scene 2 - Blended

Figure 3. *Examples of aerial and blended images illustrating the advantage of avoiding occlusions compared to aerial imagery, but also other artifacts.*

together.

3. Experimental results

A. Brief data description

The dataset comprises aerial and Street-Ortho images of size 2048×2048 pixels, which corresponds to a spatial resolution of approximately 10 cm per pixel. The dataset is split in 38 urban intersection scenes for training and 29 scenes for validation, where intersections from the same municipality are grouped together. Furthermore, the training set also includes 78 highway scenes for additional learning material, which could be easily annotated for ground-truth information. As mentioned earlier, the Street-Ortho images are sometimes not available for highway scenes or parts of the other scenes. For all scenes, drive lines and road surfaces are available in annotated form as ground truth, where the road surfaces are partly auto-generated with scores above 80%. For 46 scenes, also painted striping annotations are available.

B. Initial segmentation

Table 1 shows the IoU score and the recall for the U-Net segmentation stage, which includes both the intersection as the highway scenes for training and validation. Note that the validation results for painted striping are not presented, because the amount of training data is so low, although visual verification on new data shows that the segmentation appears to be of sufficient quality. In all these cases, it is clear that both IoU score and recall improve when using blended images albeit with a variable number of percentage points. More specifically, the recall of the road surface improves 7 percentage-points and the drive line 4 percentage points. The data amount is here much higher than for intersections only, so that a well above 80% scores are obtained.

C. Line-detection segmentation metrics

Using segmentation metrics is a possible way for evaluating the line-detection results. To create the segmentation masks, the ground-truth and detected lines are rendered as lines with 5 pixels in width. The quality of these masks can be compared using standard evaluation metrics for segmentation, *i.e.* recall and precision scores.

Table 1. Performances for drive-line and road-surface segmentation using aerial and blended images on a combined dataset of highways and urban crossings.

	Aerial		Blended	
	IoU	Recall	IoU	Recall
Road Surface	0.80	0.81	0.84	0.88
Drive line	0.85	0.55	0.86	0.59

Table 2. Measured GEO and segmentation metrics for the direct segmentation method and the late-fusion approach using aerial or blended imagery of intersections only.

Input	Algorithm	GEO		Segmentation	
		recall	prec.	recall	prec.
Prev. work [6]		0.29	0.69	n.a	n.a.
Aerial	Direct	0.50	0.47	0.25	0.26
Aerial	L-fusion	0.54	0.40	0.29	0.24
Blended	Direct	0.45	0.55	0.23	0.29
Blended	L-fusion	0.52	0.46	0.28	0.28

D. Line-detection GEO metric

Following other work on metrics [3], the evaluation includes performance in terms of GEO recall and precision. First, the GEO metric prescribes to sample the ground-truth lines and then the detected lines in equidistant points. For this, 0.25 m is adopted for the sampling distance similar to recent related work [3]. Second, using an optimal assignment [13] to assign ground-truth points within a certain range to detected points, allows to determine true positives, false negatives etc., and hence recall and precision scores. During conversations with users of the algorithms, it became clear that positional accuracy is not essential for their traffic-control applications, therefore we select a threshold of 1 m for the distance between ground-truth points and detected points.

E. Line-detection results

Both image sources (blended vs aerial) are used for testing with respect to: (a) whether drive lines are extracted directly (only using drive-line segmentation, line extraction and line merging) or (b) late-fusion results. Table 2 lists the achieved performance in terms of the GEO metric and segmentation metrics. The obtained results show that higher segmentation metrics are also coupled to higher GEO metrics in nearly all cases, except one for which we have no explanation. The results show that using the proposed pipeline typically results in higher GEO recall (*e.g.* 0.54 instead of 0.5 for aerial) at the cost of lower precision (*e.g.* 0.40 instead of 0.47). Furthermore, when using blended source data, the obtained results show higher precision with lower recall than when using aerial imagery as input. For example, the late-fusion result achieves a precision of 0.40 for aerial images and 0.46 for blended images, for nearly the same recall values of 0.54 and 0.52, respectively. The measured recall scores are lower for segmentation, since the implied distance threshold of 5 pixel line models between neighboring lines is 2.5 pixels, corresponding to approximately 25 cm, compared to the range threshold of 1 m used in the GEO metric.

F. Execution time

The dual-channel approach takes significantly more time (20 minutes) than than the direct segmentation approach (2.5 minutes). This increase is mostly due to the iterative algorithms *Clustering* and *Approach Fitting* as described in Section , although both are implemented in python code with few optimizations.

4. Discussion and Conclusions

In this paper, we have presented a dualistic segmentation approach for drive lines on lanes of roads and direct segmentation of painted stripes on roads for creating a intersection-topology model, using either aerial or combined aerial/street-level orthographic images as input. The late-fusion system is based on direct segmentation of drive-lines and a domain-knowledge model that derives drive lines from painted stripe lines. In this way, a more richer input to the final determination of drive lines can be exploited.

The experiments on pixel-level segmentation of road surface and drive lines when sufficient data is available, using a combined dataset of highways and intersections, results in recall scores 80–88% for road surface and 50–59% for drive lines.

Using only the intersection dataset, reduces all scores drastically because the total dataset size is more than halved and much more complex in terms of lane structures and it has more occlusions by trees which corrupts the delineation of roads.

For the intersection data only, the GEO metric results show that the proposed late-fusion system increases the recall score with 4–7 percentage points at the cost of a lower precision, albeit at a much lower recall value of about 50-60%. This is less harmful for this application, since manual removal of false positives costs less time than adding false negatives manually. The higher recall is explained by the fact that the late-fusion system incorporates domain knowledge as well as direct segmentation results. This recall improvement is consistent when using conventional segmentation metrics.

A further discussion is required on the low absolute values of the scores, even with the domain-knowledge incorporated in the late-fusion system. The high scores of 80% as reported by [3] can be reproduced for the mixed highways-intersections data, but not for intersections only because this data is much more complex. For example, a considerable amount of intersections in our dataset is partially occluded by trees so that the striping of lanes in the ground truth of aerial views is regularly corrupted. The planting of trees along roads is commonly applied along virtually all roads, which is part of the Dutch culture. This makes accurate modeling of complex topologies with multiple lanes a difficult task.

Future work should focus on more intersection data and specifically developing ground-truth data for blended imagery, not only for aerial imagery. Finally, it is interesting to combine and compare our complete model involving the domain-knowledge model with recently reported direct extraction approaches such as CenterLineDet [5].

Acknowledgment

This work has been funded by the Dutch government and supported by ITEA as part of the ITEA3 18036 SMART Mobility Project.

References

[1] Z. Wang, W. Ren, and Q. Qiu, *Lanenet: Real-time lane detection networks for autonomous driving*, Accessed:2022-5-5. [Online]. Available: `https://arxiv.org/pdf/1807.01726.pdf`.

[2] Y. Guo, G. Chen, P. Zhao, W. Zhang, J. Miao, *et al.*, “Gen-lanenet: A generalized and scalable approach for 3d lane detection,” in *European Conf. on Computer Vision*, 2020, pp. 666–681.

[3] S. He and H. Balakrishnan, “Lane-level street map extraction from aerial imagery,” in *IEEE/CVF Winter Conf. on Applications of Computer Vision (WACV)*, 2022, pp. 1496–1505. DOI: `10.1109/WACV51458.2022.00156`.

[4] B. Liao, S. Chen, X. Wang, T. Cheng, Q. Zhang, *et al.*, “MapTR: Structured modeling and learning for online vectorized hd map construction,” in *International Conference on Learning Representations*, 2023.

[5] Z. Xu, Y. Liu, Y. Sun, M. Liu, and L. Wang, “Centerlinedet: Centerline graph detection for road lanes with vehicle-mounted sensors by transformer for hd map generation,” in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 3553–3559. DOI: `10.1109/ICRA48891.2023.10161508`.

[6] J. A. Vijverberg, B. J. Beers, and P. H. N. de With, “Towards automatic inference of layouts of traffic intersections for smart cities,” in *GEOProcessing*, 2022, pp. 43–46.

[7] J. A. Vijverberg, B. J. Beers, E. Bondarev, and P. H. N. de With, “Drive-line extraction from aerial images,” in *GEOProcessing*, 2022, pp. 43–46.

[8] *Street ortho lidar*, Accessed:2023-9-10. [Online]. Available: `https://www.berryvansomeren.com/posts/street_ortho`.

[9] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, Springer, 2015, pp. 234–241.

[10] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, *et al.*, “Swin transformer: Hierarchical vision transformer using shifted windows,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[11] Y. Zhou, H. Qi, and Y. Ma, “End-to-end wireframe parsing,” in *ICCV 2019*, 2019.

[12] M. Dirnberger, T. Kehl, and A. Neumann, “NEFI: Network extraction from images,” *Sc. Reports*, vol. 5, 2015.

[13] H. W. Kuhn, “The hungarian method for the assignment problem,” *Naval research logistics quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.

Author Biography

Julien Vijverberg received his MSc in Electrical Engineering at Eindhoven University of Technology (TU/e) and pursues his PhD in Cyclomedia under supervision of Peter de With at the TU/e.

Bart Beers received his MSc in Geodesy from Delft University of Technology in 1979 and his PhD from the same university in 1995 on the subject of a photogrammetric system. He is one of the founders of Cyclomedia and has continuously worked on technological innovations in the field of Mobile Mapping.

Peter H. N. de With is Full Professor of the Video Coding and Architectures group in the Department of Electrical Engineering at Eindhoven University of Technology. He worked at various companies and was active as senior system architect, VP video technology, and business consultant. He is an IEEE Fellow and member of the Royal Holland Society of Sciences, has (co-)authored over 600 papers on video coding, analysis, architectures, and 3D processing and has received multiple paper awards. He has been a program committee member of several IEEE conferences and holds some 30 patents.