IGARSS 2021 - 2021 IEEE International Geoscience and Remote Sensing Symposium | 978-1-6654-0369-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/IGARSS47720.2021.9554162

END-TO-END ROOFLINE EXTRACTION FROM VERY-HIGH-RESOLUTION REMOTE SENSING IMAGES

Wufan Zhao, Claudio Persello, Alfred Stein

Dept. of Earth Observation Science, ITC, University of Twente,
Enschede, The Netherlands
wufan.zhao@utwente.nl, c.persello@utwente.nl, a.stein@utwente.nl

ABSTRACT

Roof shape information is essential for creating 3D building models. However, the automated extracting of roof structures from Earth observation data is a difficult task involving significant uncertainties caused by scene complexity and limited multi-source data coverage. This paper introduces the integrally-attracted wireframe parsing (IAWP) framework to reconstruct building rooflines as a planar graph from remotely sensed images with a single forward pass. We add global geometric line priors through the Hough transform into deep networks to better extract the linear geometric features. We perform experiments on the vectorizing world building (VWB) dataset. The investigated method improves the F-score metrics of corner points/edges by 0.1%/7.7% and 0.6%/1.1%, respectively. Visual comparison results also indicate that the HT-IHT block gives consistent improvements in terms of geometric regularity.

Index Terms— *Building roofline extraction, End-to-end learning, integrated attraction field, Hough-Transformation*

1. INTRODUCTION

Automatic building reconstruction is an essential aspect of capturing and updating spatial data for urban applications, for instance, in urban planning, 3D navigation, and emergency response [1]. The roof structure is essential for creating 3D building models at different levels of detail (LoD). Automated building roof line extraction has remained a challenging task, mainly due to the varying building roof configurations, shadows, geometric distorsions, and overhanging vegetation in the images.

Many research efforts have been conducted to extract rooflines using various Earth observation data, including point cloud data, optical images and Geographic Information System (GIS) vector data, taking advantage of the synergy among all data sources. Fernandes et al. proposed to extract groups of straight lines representing roof boundary sides and roof ridgelines from high-resolution aerial images using corresponding airborne laser scanner (ALS) roof polyhedrons as initial approximations [2]. Alidoost et al. proposed to utilize convolutional neural networks (CNNs) to extract the inherent and latent features from a single image and interpret these as 3D information for building roofline extraction and reconstruction [3]. Due to scene complexity and limited multi-source data coverage, those methods usually involve complex multiple-steps processing pipelines involving feature extraction, fusion, and morphological operations (illustrated in Figure 1). Thus, they are limited to areas with few buildings and cannot scale up to large urban areas.

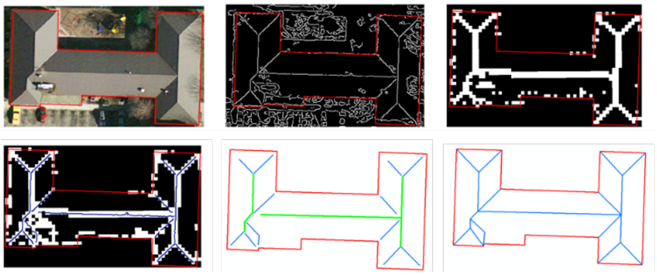

Fig. 1 Traditional hybrid workflow for the extraction of rooflines from Earth observation data (revised from [1]).

In recent years, advances in optical image sensor technology and deep learning algorithms have offered new opportunities to accelerate roof structure extraction and 3D building modeling research. Nauata et al. present an algorithm that uses CNNs to detect geometric primitives and infer their relationships, where integer programming (IP) fuses all the information into a planar graph through holistic geometric reasoning. It first transforms the 2D roof architecture vectorization problem to infer a planar graph from a single RGB image [4]. Jointly detecting meaningful and salient line segments and junctions is challenging as it requires inferring a graph structure with an arbitrary topology. Thus, it is a high-level reconstruction task akin to the floorplan vectorization and wireframe parsing [6], in contrast to low-level reconstruction tasks such as point and line detection. The inference of graph topology of buildings in satellite images is more challenging due to the foreshortening effects through perspective projection. Zhang et al. proposed to apply convolutional message-passing network (Conv-MPN) architecture for roofline structure reconstruction. This method relies highly on preliminary processing (corner detection), and the framework is computationally expensive and inefficient in training and inference [5].

Motivated by the success of recent works in wireframe parsing [6,7], we turn the roofline extraction task into a high-level graph structure reconstruction problem. We aim to directly predict vectorized building roof structure in an end-to-end learnable way. The input is a satellite RGB image. The output is a planar graph depicting both the internal and external roof architecture feature lines.

This paper makes the following contributions: 1) we investigate the applicability of the integrally-attracted wireframe parsing (IAWP) framework in the context of outdoor building roofline structure reconstruction; 2) we integrate geometric line priors into deep networks for enhanced geometric feature extraction and improved data efficiency by relying on the Hough transform block. Such blocks performs in the Hough domain over the space of all possible image-line parameterizations, which provides the gloal prior knowledge of line features. The IAWP was originally applied for indoor scene wireframe parsing. It is also termed the holistically-attracted wireframe pasing (HAWP).

3. METHODOLOGY

3.1. Integrally-Attracted Wireframe Parser

The IAWP is an end-to-end trainable and fast parsimonious parsing method that can detect a vectorize wireframe in an input image. As illustrated in Fig. 2, the IAWP consists of three components: (i) line segment and junction proposal generation, (ii) line segment and junction matching, and (iii) line segment and junction verification.

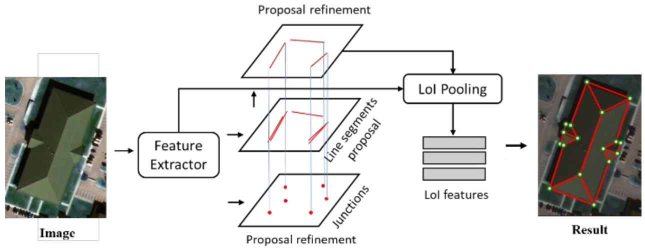

Fig. 2 Architecture of the adopted method (Modified from [7]).

i) Proposal initialization: line segment detection and junction detection. The input image is processed first by the backbone, i.e., the stacked Hourglass network [10], to extract shared deep features. After that, there are two parallel branches to detect junctions and line segments, respectively. For junction detection, the junction mask map and junction offset map are calculated. During training, the binary cross-entropy and the l_1 loss total loss is used. In the inference, the standard non-max-suppression (NMS) is applied to select the top-K junctions as initial junction proposals. For computing line segment proposals, we refer to the original work of HAWP, which derives a 4-D vector field map for line segments. Such a scheme is proved to be more accurate and efficient in reparameterizing the line segments. In addition, to predict the attraction field map (AFM) from the feature map, the network also computes a distance residual map, which is leveraged as an auxiliary supervised signal for learning the position of line segments.

ii) Proposal refinement: line segment and junction matching. The role of this step is to find meaningful alignment between line segment initial proposals and junction proposals. A line segment proposal from the initial set is kept if and only if its two endpoints can be matched with two junction proposals based on Euclidean distance with a threshold. A junction proposal is removed if it does not match any survived line segment proposal after refinement.

iii) Proposal verification: line segment and junction classification. The verification process is to classify the line segments and junctions from the previous refinementstage. Line-of-Interest (LOI) pooling operation is utilized to compute features for a line segment [6]. Geometrically, the proposed wireframe parser is enabled by the holistic 4-dimensional AFM and the "basins" of the attraction field revealed by junctions.

3.2. Hough transform block for global line priors

Recent works add prior knowledge into deep networks that aim to enhance built-in geometric information and reduce the dependency on labeled data [8, 9]. Motivated by those, we add line priors through a trainable Hough transform block into the backbone feature extraction network to better retrieve the building roofs' line features.

Typically, the Hough transform parameterizes lines in polar coordinates as an offset ρ and an angle θ. These two parameters are discretized in bins. Each pixel in the image votes in all line-parameter bins to which that pixel can belong. The binned parameter space is denoted the Hough space, and its local extrema correspond to lines in the image.

We integrate a Hough transform and inverse Hough transform (HT-IHT block) to combine locally learned image features with global line priors. We allow the network to combine information by defining the Hough transform on a separate residual branch.

The HT layer inside the HT-IHT block maps input features to the Hough domain and produce transformed features. This is followed by a set of local convolutions in the Hough domain which are equivalent to global operations in the image domain. The result is then inverted back to the image domain using the IHT layer, and it is subsequently concatenated with the convolutional branch.

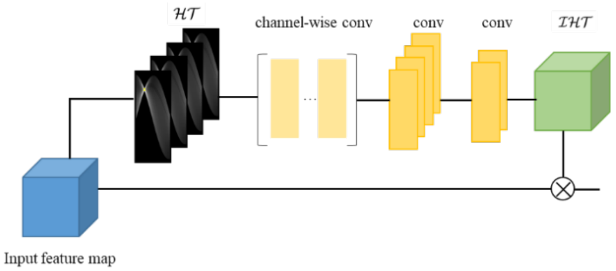

Fig. 3 HT-IHT block.

As shown in Fig. 3, the input feature map, coming from the previous convolutional layer, learns local edge information and is combined on a residual branch with line candidates, detected in global Hough space. The input feature map is transformed channel-wise to the Hough domain through the HT layer into multiple HT maps. The result is then filtered with 1D channel-wise convolutions. Two subsequent 1D convolutions are added for merging and reducing the channels. The output is converted back to the image domain by the IHT layer. The deep hough transform, which applied to all unique lines in an image, is order-agnostic in both the feature space and the parametric space, making it highly parallelizable. Upon implementation, we replace the hourglass blocks with the HT-IHT block, and the parameters of the HT-IHT block are much less than hourglass block. The Hough transform provides the prior knowledge about global line parameterizations, while the convolutional layers learn the local gradient-like line features.

3.3. Loss function

The overall workflow is trained end-to-end with the following loss function,

$$L = L_{LS} + L_{Junc} + L_{Ver}$$

where L_{LS}, L_{Jun} and L_{Ver} are the losses for the line segments, junctions, and vertices, respectively. The channel-wise ℓ1 norm is used for computing line segments loss. The L_{Jun} calculates the weighted sum of losses from the junction mask map and the junction offset map. We use binary cross-entropy loss in the verification module. Denote by and L_{Ver} the loss computed on the sampled LoIs. A more detailed descriptions of the loss functions can be found in [7].

4. EXPERIMENTAL RESULTS

4.1. Datasets and Evaluation Metrics

4.1.1. Datasets.

We perform experiments on the vector world building (VWB) [4] dataset to validate our method. The images are part of the SpaceNet Public Dataset, and they cover the cities of Atlanta, Paris, and Las Vegas. The images have a spatial resolution of 30 cm. The interior and exterior building edges are annotated as 2D planar graphs. The building instances in images are cropped and resized to the size of a 256 × 256 image patch. The entire dataset includes 1601 training and 400 testing samples.

4.1.2. Evaluation Metrics.

For a fair comparison, we follow the accuracy evaluation settings used in other wireframe parsing and roofline extraction works [5, 6]. Instead of directly using the vectorized representation of line segments, heatmaps are used, which are generated by rasterizing line segments for both parsing results and the ground truth. We reported the measures heatmap based precision (P), recall (R), and F-score (F) for junctions and line segments (edges), respectively.

4.2. Implementation Details

Our method is trained using the Adam optimizer with a total of 100 epochs on a single GeForce RTX 2080 Ti GPU device. The learning rate, weight decay rate, and batch size are set to 4×10^{-4}, 1×10^{-4} and 6, respectively. The learning rate is divided by 10 at the 25-th epoch. We kept other hyper-parameter settings the same as with original IAWP.

4.3 Results and Discussion

We conducted the experiments using IAWP and IAWP-HT, respectively and compared them with Conv-MPN.

4.3.1. Quantitative results

Table 1 summarizes the results and comparisons in terms of the evaluation metric stated in Section 4.1.2. The IAWP and IAWP-HT achieved better performances on most metrics. Specifically, compared with Conv-MPN, IAWP improves the F-score metrics of corner points and edges by 0.1% and 7.7%, respectively. Moreover, the Hough transformation further improves this result by 0.6% and 1.1%, respectively, indicating that the HT-IHT block enhances the geometric feature detection. Conv-MPN shows a slightly higher value in precision metric for junctions since it uses Faster R-CNN [11] to pre-extract the corner points, while our method does not require a pre-processing step. In terms of efficiency, both methods run roughly two times faster than Conv-MPN which applies convolutional message passing for graph feature volumes update, and only using 1/4 GPU memory. This proves its increase in effectiveness in inferring the planar graphs of building roofline structure.

Tab. 1 Extraction results on the VWB dataset

		Conv-MPN	IAWP	IAWP-HT
Juction	P	77.9	76.4	77.2
	R	80.2	82.1	82.3
	F	79.0	79.1	79.7
Edge	P	56.9	60.8	61.5
	R	60.7	73.2	74.7
	F	58.7	66.4	67.5

4.3.2. Visual comparison

Figure 4 shows the planar graph reconstruction results by the IAWP based methods. Both IAWP and IAWP-HT are able to extract complex building roof structure with both interior and exterior edges without relying on any hand-crafted constraints or preliminary vertices detection. The corner points and connection relationships between corner and line segments can be well reconstructed for complex roof structures. Adding the HT block allows the network to be more sensitive to the roof's linear details by combining local and global image information.

We further notice several failure cases during the experiment due to missing detection and incorrect graph relation inference. Future research will aim to further improve the approach by training on other datasets and adding multiple data sources (e.g., a digital surface model).

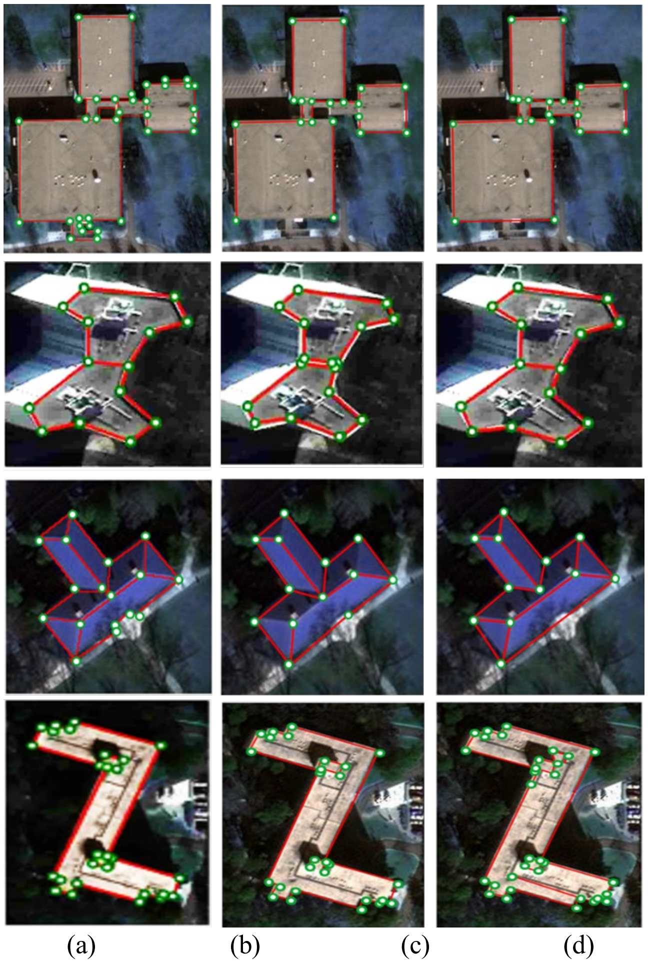

Fig. 4. Example results from different methods. (a) Ground truth (b) IAWP (c) IAWP-HT

5. CONCLUSIONS

This paper introduces a integrally-attracted wireframe parser framework for end-to-end building roofline extraction from very-high-resolution remote sensing images. We exploit geometric priors based upon the Hough transform for improving line feature detection. The results on VWB datasets show that IAWP-based methods perform better than competing models (i.e., Conv-MPN) in quantitative evaluations. The HT-IHT block gives consistent improvements in terms of precision and geometric regularity on visual comparison. Both methods further indicate a considerable improvement in computational resource-saving and training efficiency. The wireframe would enable richer architectural modeling and analysis for broad applications in urban visualization and planning.

6. REFERENCES

[1] Zheng Y, Weng Q, Zheng Y. A hybrid approach for three-dimensional building reconstruction in indianapolis from LiDAR data. Remote Sensing. 2017 Apr;9(4):310.

[2] Fernandes VJ, Dal Poz AP. A Markov-random-field approach for extracting straight-line segments of roofs from high-resolution aerial images. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. 2016 Sep 7;9(12):5493-505.

[3] Alidoost F, Arefi H, Tombari F. 2D Image-To-3D Model: Knowledge-Based 3D Building Reconstruction (3DBR) Using Single Aerial Images and Convolutional Neural Networks (CNNs). Remote Sensing. 2019 Jan;11(19):2219.

[4] Nauata N, Furukawa Y. Vectorizing World Buildings: Planar Graph Reconstruction by Primitive Detection and Relationship Inference. In European Conference on Computer Vision 2020 Aug 23 (pp. 711-726). Springer, Cham.

[5] Zhang F, Nauata N, Furukawa Y. Conv-mpn: Convolutional message passing neural network for structured outdoor architecture reconstruction. InProceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2020 (pp. 2798-2807).

[6] Zhou Y, Qi H, Ma Y. End-to-end wireframe parsing. InProceedings of the IEEE International Conference on Computer Vision 2019 (pp. 962-971).

[7] Xue N, Wu T, Bai S, Wang F, Xia GS, Zhang L, Torr PH. Holistically-attracted wireframe parsing. InProceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2020 (pp. 2788-2797).

[8] Han Q, Zhao K, Xu J, Cheng MM. Deep Hough Transform for Semantic Line Detection. arXiv preprint arXiv:2003.04676. 2020 Mar 10.

[9] Lin Y, Pintea SL, van Gemert JC. Deep Hough-Transform Line Priors. InEuropean Conference on Computer Vision 2020 Aug 23 (pp. 323-340). Springer, Cham.

[10] Newell A, Yang K, Deng J. Stacked hourglass networks for human pose estimation. InEuropean conference on computer vision 2016 Oct 8 (pp. 483-499). Springer, Cham.

[11] Ren S, He K, Girshick R, Sun J. Faster r-cnn: Towards real-time object detection with region proposal networks. IEEE transactions on pattern analysis and machine intelligence. 2016 Jun 6;39(6):1137-49.