AUTOMATED CHAIN FOR LARGE-SCALE 3D RECONSTRUCTION OF URBAN SCENES FROM SATELLITE IMAGES

S. Tripodi, L. Duan, F. Trastour, V. Poujad, L. Laurore, Y. Tarabalka

LuxCarta Technology, 460 avenue de la Quiéra, 06370 Mouans-Sartoux, France - stripodi@luxcarta.com

KEY WORDS: 3D Reconstruction, Satellite Images, Stereo Pair, Deep Learning, U-Net.

ABSTRACT:

Automatic city modeling from satellite imagery is a popular yet challenging topic in remote sensing, driven by numerous applications such as telecommunications, defence and urban mamagement. In this paper, we present an automated chain for large-scale 3D reconstruction of urban scenes with a Level of Detail 1 from satellite images. The proposed framework relies on two key ingredient. First, from a stereo pair of images, we estimate a digital terrain model and a digital height model, by using a novel set of feature descriptors based on multiscale morphological analysis. Second, inspired by recent works in machine learning, we extract in an automatic way contour polygons of buildings, by adopting a fully convolutional network U-Net followed by a polygonization of the predicted mask of buildings. We demonstrate the potential of our chain by reconstructing in an automated way different areas of the world.

1. INTRODUCTION

Automatic 3D reconstruction of urban scenes from satellite imagery is a popular yet challenging topic in remote sensing (Bittner et al., 2017, G. Facciolo, 2017). The required accuracy in industrial applications is very high and ever-increasing, which is critical for several fields such as: telecommunications, urban planning, defense, etc. To reconstruct a 3D city model from stereo pairs of satellite images, semi-automatic strategies are typically applied, which are based either on procedural modeling (Vanegas et al., 2010), or on the use of both image processing and machine learning methods to infer scene geometries together with semantics (i.e. distinguish roads, buildings etc.) (Tripodi et al., 2018). In both cases, human interaction still plays a key role, in particular for the rooftop buildings extraction regarding the existing challenges such as occlusion, complex roof-structure, big diversity, small dense buildings, etc. This increases the cost of 3D city models.

In this paper, we propose an end-to-end automated chain for large-scale 3D reconstruction of urban scenes with a Level of Detail 1 (LOD1) of the CityGML formalism, i.e. where buildings are represented as piecewise planar objects with flat roofs and vertical facades (Groger , Plumer, 2012). The proposed chain is inspired by the most recent computer vision and deep learning techniques, and features two principal innovations: 1) We have developed a Digital Terrain Model (DTM) generation method, which uses a novel set of feature descriptors based on multiscale morphological analysis to extract reliable bare-terrain elevations from Digital Surface Models (DSMs). 2) We have developed an algorithm for automatic extraction of contour polygons of buildings. Following the outcomes of the building dense labeling challenge (Huang et al., 2018), we have adopted a U-Net convolutional neural network (Ronneberger et al., 2015) for a building segmentation task. Furthermore, a polygonization algorithm is designed, which processes a mask of buildings to output an ensemble of polygons, where each polygon delineates a building contour.

2. RELATED WORKS

A comprehensive overview of methods for building 3D urban models can be found in the article of Musialski et al. (Musialski et al., 2012). Aerial acquisitions with Lidar scanning (Verdie et al., 2015) or multi-view optical imagery (Hane et al., 2017) has been intensively used so far to reconstruct 3D models on large-scale urban scenes. Because of high acquisition costs and authorization constraints, aerial acquisitions are, however, available only for a limited number of areas in the world. Satellite optical images constitute an excellent alternative to serve as an input data for 3D urban modeling, considering their high-frequency and lower-cost acquisition, as well as continuous worldwide coverage. However, as described in (Poli , Caravaggi, 2013), until very recently the quality of satellite imagery coupled with available methodologies did not allow to produce 3D city models at a high spatial resolution in an automatic way.

A few recent research works have proposed automated methodologies for urban scene reconstruction in LOD 1 from stereo pairs of high-resolution satellite images. The method of (Duan , Lafarge, 2016) has used a semi-global matching technique (Hirschmuller, 2008) to find correspondences in a stereo pair of epipolar images, followed by a joint classification using image radiometry coupled with estimated elevation information to retrieve 3D city models. Even though this method offered a solution for 3D urban reconstruction at a large scale, small geometries could not be captured precisely. Wang and Frahm developed an approach for 3D reconstruction from multi-view stereo satellite images (Wang , Frahm, 2017). They apply a Scale Invariant Feature Transform (SIFT) based (Lowe, 2004) approach to compute fast but reliable 2D feature matches between image pairs, then use an edge-aware interpolation of sparse feature matches followed by bilateral smoothing to obtain dense correspondences, further reconstructed into 3D point clouds. This method has been specifically designed for multi-view stereo data, and gives poor results when applying on one stereo pair only. Furthermore, the reconstructed point cloud does not contain any semantic information, which is essential nowadays for many real-life uses of 3D city models.

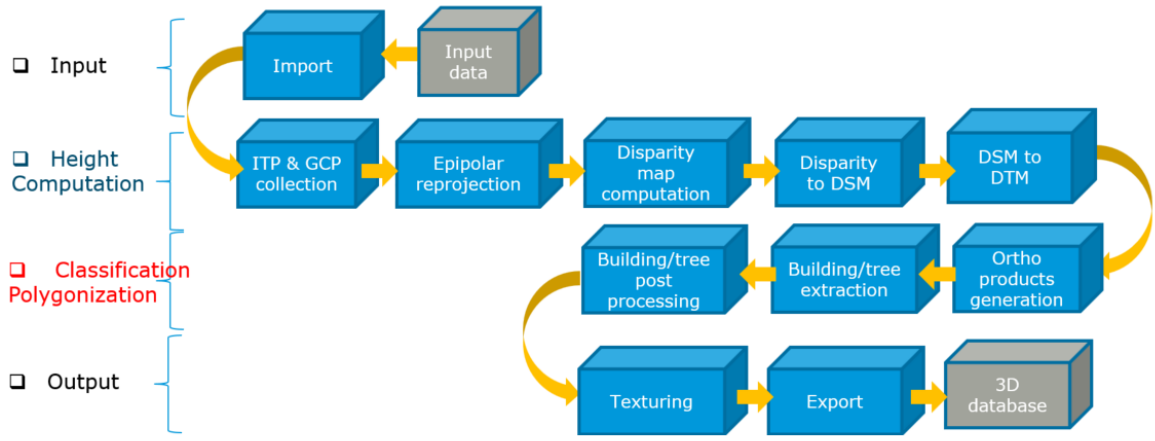


Figure 1. Proposed automated chain for large-scale 3D reconstruction in LOD1.

To reconstruct semantic information, deep learning, in particular convolutional neural networks (CNNs) have been developed at a fast pace and have shown excellent performance for image interpretation. Lately, end-to-end CNN-based architectures have been designed to accurately estimate semantic labels of every pixel for satellite and aerial images, *i.e.* assign each pixel to one of the classes of interest, such as building, tree, car, *etc* (Volpi , Tuia, 2017, Zhu et al., 2017). Thanks to their capability to jointly learn to extract expressive high-level contextual features and conduct the categorization, CNNs have proven being able to generalize to different areas of the earth and to take into account the important intra-class variability encountered over large geographic areas (Huang et al., 2018). As demonstrated in (Huang et al., 2018), U-Net and SegNet are among the most successful CNN architectures for semantic labeling. These algorithms estimate for every pixel a probability to belong to a class of interest. By thresholding these probabilities, a classification mask can be computed, for example a building/non-building mask. To integrate this semantic information within a 3D urban model, the obtained raster mask must be polygonized, so that each polygon corresponds to a building contour. While the common polygonization methods such as the algorithm of Douglas-Peucker (Douglas , Peucker, 1973) are very sensitive to the quality of the input data, mesh-based approximation approaches have been proposed (Tasar et al., 2018), which are more robust but exhibit a high computational complexity. Very recently, a few attempts have been made to design a chain, which learns in an end-to-end fashion to predict building polygons from input aerial images. However, either these models are limited to regress very simple shapes (Girard , Tarabalka, 2018), or results are not accurate enough to be used in real-life applications (Marcos et al., 2018). Thus, the design of the algorithm which would allow to reconstruct contour polygons of buildings with high speed and precision is still an open research topic.

3. PROPOSED CHAIN

The scheme of our automated chain for large-scale 3D reconstruction of urban scenes in LOD1 is illustrated in Figure 1. At the input of the method, a stereo pair of satellite images with the data associated as RPC (Rational Polynomial Coefficients) models (Guo , Xiuxiao, 2006) is given. The developed chain can manage various types of optical satellite images, with different spatial resolutions, such as WorldView, Pléiades, GeoEye, Spot, *etc*. Throughout this article, we illustrate the workflow results by applying it on satellite images with a spatial resolution of 50 cm/pixel, this resolution being sufficient for extracting in an accurate way footprints of buildings. The outcome of the proposed chain presented here is a 3D model in LOD1 represented by a Digital Terrain Model (DTM) and a set of building polygons with the associated height.

This proposed chain consists of two main parts presented in the following sub-sections, respectively:

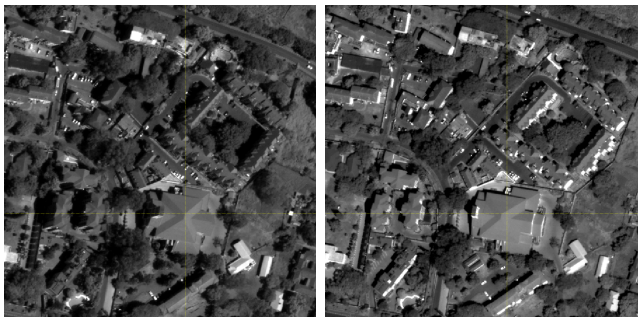
- Height computation, which comprises estimation of a Digital Surface Model (DSM) and a DTM.
- Semantic labeling, comprising the extraction and polygonization of building contours.

The 3D model is then computed by combining height and semantic information, as described in Section 3.3. In the final part of the workflow, we texture the obtained 3D model; however, the texturing procedure is out of scope of this paper.

3.1 Height computation

From the input stereo pairs of images, height information can be efficiently extracted by using an epipolar geometry to estimate a disparity map, followed by a DSM computation. We propose to apply the following steps for this purpose:

1. Adjustment of the RPC model.
2. Construction of epipolar images.
3. Estimation of a disparity map.
4. DSM computation.
5. Generation of a DTM from a DSM.
6. Computation of a Digital Height Model (DHM), which can be obtained by substracting a DTM from a DSM to get the height of the objects above the ground.



(a) Left image (b) Right image

Figure 2. Closeups of two epipolar images of Nairobi.

3.1.1 Adjustment of the RPC model. RPC camera models encode both the intristic and extrinsic calibrations (Guo , Xiuxiao, 2006). As the RPC models provided by the vendors of satellite images are not accurate enough to georeference an image, the first step of our chain consists in adjusting an RPC model using Ground Control Points (GCPs). For this purpose, Image Tie Points (ITPs) are detected between the pair of images and a reference. The absolute positioning of salient features on the ground (ex., crossroads) can be measured and served as a reference. If an orthoimage over the same area of interest is available, this image can be used as a reference, and ITPs can be computed by using any feature detection algorithm, such as AKAZE, ORB, etc (Tareen , Saleem, 2018). The detected ITPs can be further filtered to keep only the points on the ground, by using one of the following approaches:

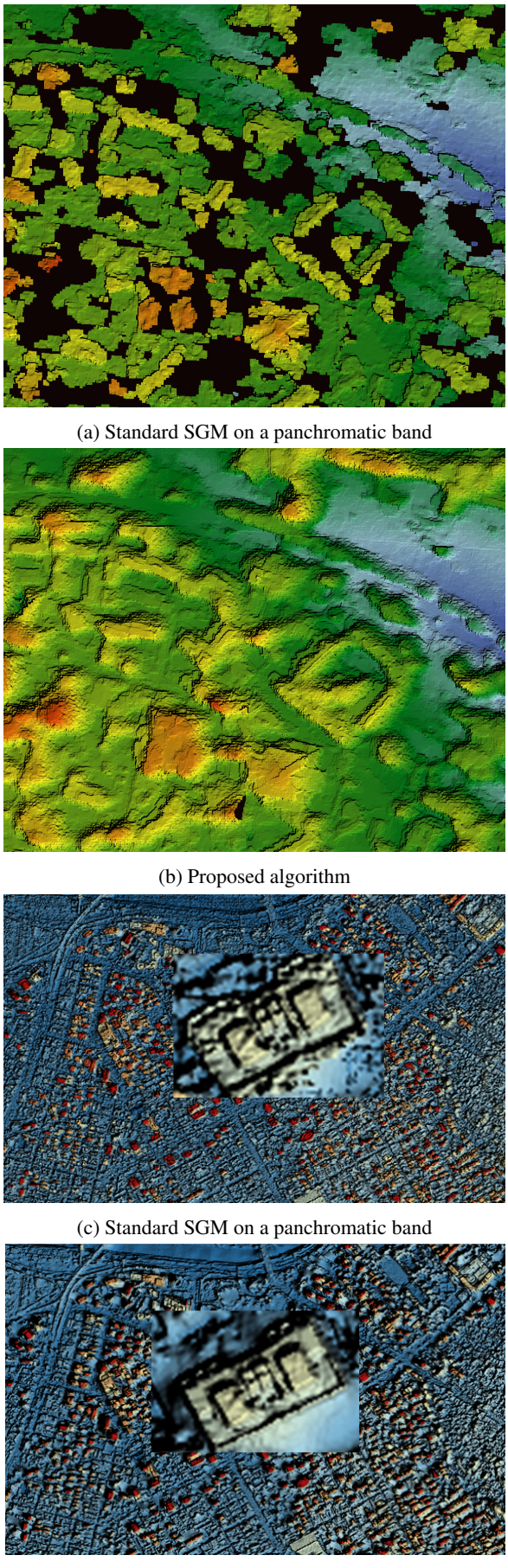
1. By computing locally for each ITP an epipolar image and a disparity map, as described in Sections 3.1.2 and 3.1.3, respectively. The ground points are then estimated from the disparity map and a satellite angle of view.

2. By computing a mask of buildings and trees using deep learning labeling, as detailed in Section 3.2.

Once the GCPs are determined, the RPC model is adjusted by optimizing a polynomial model of order 3, followed by updating RPC coefficients (we refer the reader to (Hartley , Zisserman, 2003) for more details).

3.1.2 Epipolar image reprojection. Epipolar images are stereo pairs in which the left and right images are oriented in such a way that common feature points have the same y-coordinates on both images. This allows to match a pair of images in a faster and more accurate fashion.

To compute epipolar images, ITPs are extracted on the stereo pair of images, by applying one of feature detection methods as described in Section 3.1.1. By using these ITPs and optimizing a polynomial model of order 3, we adjust the RPC model to build two epipolar images in such a way so that the collected ITPs are positioned on the same line. The use of the polynomial of the third order allows us to take into account and model a terrain deformation in an accurate way. Figure 2 illustrates an example of two epipolar images (750×750 closeups of the full processed tiles) reprojected from a stereo pair of Pléiades images acquired over the city of Nairobi, Kenya. In the following figures, we will demonstrate the experimental results on the same Nairobi dataset.



(a) Standard SGM on a panchromatic band

(b) Proposed algorithm

(c) Standard SGM on a panchromatic band

(d) Proposed algorithm

Figure 3. Closeups of DSMs, obtained by applying the standard SGM (Hirschmuller, 2008) and the proposed algorithm.

3.1.3 Estimation of a disparity map and DSM computation. From two epipolar images, a disparity map is generated using a method based on semi global matching (SGM) algorithm (Hirschmuller, 2008). The SGM technique has proven its efficiency, however it fails under the certain conditions, for example when computing a disparity map for textureless regions. We have designed an algorithm which uses SGM at its core, but includes pre- and postprocessing steps, which make the workflow more robust and thus yield better disparity maps. These improvements include:

1. In addition to an epipolar image pair, we also give at the input the initial disparity map, which we generate by using the elevation data at the best available resolution, such as SRTM[1]. This allows to reduce the search range.

2. Disparity maps are computed for each spectral band of the images, and are then fused into one disparity map. This ensemble approach allows to fill missing values, remove artefacts and increase an elevation accuracy.

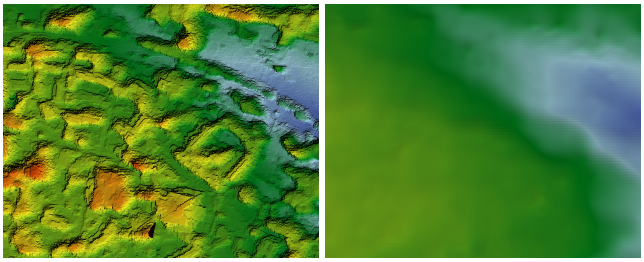
3. Postfiltering is applied to align the edges in the source satellite image and the disparity map.

The DSM is computed from the disparity map and RPC model (Hartley , Zisserman, 2003). Figure 3 shows closeups of DSMs, where the improvements obtained by our algorithm are illustrated when compared to the standard SGM method. While the SGM leaves a lot of non informed values (black pixels in Figure 3a), our method yields a more uniform DSM, where each pixel has an informative value (see Figure 3b). Furthermore, our algorithm provides more accurate contours (see a comparison on Figures 3c and 3d).

3.1.4 Generation of a DTM from a DSM. In Section 3.1.3, we discussed about the generation of DSMs which include various geographical information presented in the image scene, such as ground, trees, buildings, mountains, *etc*. In particular, buildings and trees are among the most interesting above-ground semantic classes for 3D modeling of urban scenes. The heights of these objects can be obtained by subtracting a DTM from a DSM, where the DTM represents the bare-earth elevation. We have developed a new DTM generation algorithm (Duan et al., 2019) that consists of two steps: classification and surface interpolation.

The proposed method applies a novel set of feature descriptors to classify all pixels in the DSM into four classes: flat-ground, above-ground-objects, slopes and other. The feature descriptors are computed by applying a multi-scale morphological profile analysis, and classifying each pixel by observing the changes between adjacent profiles. Our approach is inspired by (Pesaresi , Benediktsson, 2001), but adapted to elevation maps by using multiscale erosion and opening operations (instead of opening and closing, as in (Pesaresi , Benediktsson, 2001)). The proposed classification scheme extracts reliable bare-terrain pixels, is robust to noise from the DSM and adapts well to local reliefs in both flat and highly mountainous areas.

From the estimated bare-terrain elevations, we reconstruct the final DTM by applying a least-squares smooth embedding approach to interpolate the surface. We adapted the spectral affine-kernel embedding (SAKE) algorithm proposed in (Budninskiy et al., 2017), and developed its simpler expression that we call LAKE, as the Spectral solve in the original approach is replaced by a faster Least-squared solve in our specific application. As it turns out, this approach is particularly appropriate to construct a DTM, as a bare terrain can be thought of as a smooth two-dimensional embedding in 3D of a surface given as a few scattered elevations. The SAKE method does not suffer form typical oscillations of interpolation methods based on differential equations. Our experiments run on worldwide scenarios show the potential of our algorithm to produce reliable and accurate large-scale DTMs from a large variety of DSMs of various qualities and spatial resolutions from satellite data. Figure 4 shows an example of the DTM generated by our method.



(a) DSM (b) DTM

Figure 4. Example of the generated DTM from a DSM.

3.2 Building classification and polygonization

In the previous section, we have explained how our chain estimates the height information from a stereo pair of satellite images. In this section, we describe our methodology aiming at extracting contour polygons of buildings. Our algorithm comprises two steps:

1. Semantic labeling.

2. Polygonization of building contours.

The first step aims at assigning each pixel of the satellite image to a building or a no/building class. In the most recent works, a U-Net convolutional neural network architecture has exhibited the highest performance; in particular, it has shown the best results in the challenge of building contour segmentation from aerial images (Huang et al., 2018). The

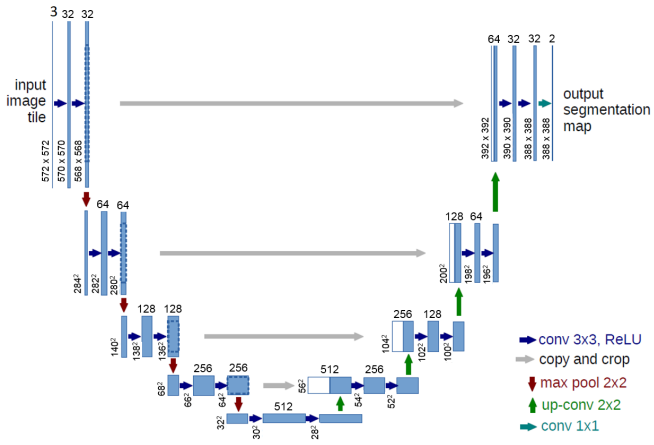
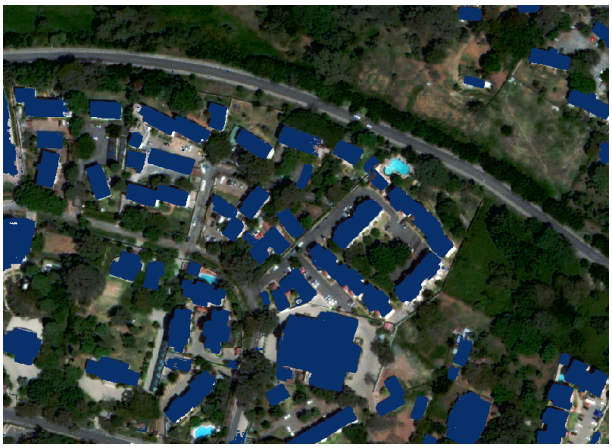


Figure 5. U-Net architecture described in (Huang et al., 2018) and used in our chain.

[1] https://www2.jpl.nasa.gov/srtm/index.html



(a) Classification

(b) Polygonization

Figure 6. Classification and polygonization of buildings.

U-Net architecture (Ronneberger et al., 2015) is built upon the fully convolutional network and is modified by adding skip connections between the downsampling path and the upsampling path, which aim at preserving the local information and thus better outlining contours of objects.

Inspired by the winning method of the Inria aerial image labeling benchmark (Huang et al., 2018), we have adopted the original U-Net architecture from (Ronneberger et al., 2015), with a single major modification: we used half as many filters at each layer (the modified architecture is illustrated in Figure 5). For example, 32 filters are used instead of 64 in the first-level convolutional layers, 64 filters instead of 128 filters in the second-level layers, etc. As explained in (Huang et al., 2018), the reduced number of filters in U-Net model decreases the risk of overfitting to the training data.

The main bottleneck of applying deep learning for accurate labeling tasks is the availability of training data set, comprising satellite images and the corresponding masks delineating objects of the thematic classes we are interested in. Providing geodata for more than 20 years, we have a huge quantity of ground-truth data. Using these data allowed us to train a generic model able to predict good classification of the building rooftops for the majority of use cases. It has been tested on thousands of square kilometers, giving very good and useful results in an industrial context.

Figure 6a illustrates an example of building labeling results. U-Net provides as output a classification mask in a raster format; however, a vector of each building is asked by our customers. As discussed in Section 2, polygonization of raster masks is not a trivial problem to solve. For example, building polygons have several constraints, including the fact that in the majority of cases building angles are right. We have developed an algorithm for polygonization of building contours, which consists of two steps: First, a naive polygonization of the mask of every building is performed, using the GDAL[2] tool gdal_polygonize. Then, we perform a polygon simplification, by searching for a compressed polygon with the best quality/complexity ratio, *i.e.* with the minimum number of vertices within a specified tolerance of en error. We have adapted a method of (Gribov , Bodansky, 2004) to our task, where we reinforced the right angle constraint. Figure 6b depicts the results of our polygonization algorithm, where it can be seen that the shapes of buildings are accurately preserved. The designed polygonization technique works very well because the U-Net succeeds in yielding accurate building masks. In the case of the moderate prediction results, a more sophisticated algorithm similar to (Tasar et al., 2018) would be needed to reconstruct contours of buildings.

Figure 7. Example of 3D reconstruction in LOD1.

3.3 3D Reconstruction

In the previous sections, we have described our workflow for the extraction of height information and building footprints. To reconstruct 3D model in LOD1, we have to assign a height for each building. For this purpose, we first compute a DHM, by subtracting a DTM from a DSM, as mentioned in Section 3.1.4. Then, for each building polygon we collect all pixel values of a DHM included in the considered polygon. We compute the median of this collected set of values, and assign the resulting median value as the height of the corresponding building; this allows to remove artefacts presented in the DHM.

As a result, we have for each building its footprint and its relative height with respect to the DTM. The DTM also allows us to georeference in an accurate way each building. Figure 7 shows an example of the 3D urban reconstruction by combining the obtained footprints and corresponding heights.

[2]https://www.gdal.org

4. EXPERIMENTS

We have illustrated an example of the application of our automated chain on the stereo pair of Pléiades images over Nairobi throughout the paper. We have tested our chain in production on more than thirty cities over the total area of thousands of square kilometers, for different kinds of urban area: residential, industrial and dense. In this section, we complete the validation of our chain. We first describe the dataset used for training the U-Net network. We then compare our reconstruction results with the ground-truth and with other state-of-the-art automatic methods.

4.1 Dataset

To train a neural network model which would be generic to work well on new unseen cities, the training data must be large enough and contain a big variety of representative samples. For this purpose, we have composed a training dataset set containing images acquired by different satellites over different types of cities (dense, industrial, residential areas and city centers). We have chosen the following data to cover varied cases:

- Images have been acquired by 3 types of satellites: Pléiades, WorldView and GeoEye. We have uniformized the image sampling at 50 cm/pixel of spatial resolution.
- 15 cities across 5 continents are present: Sydney (AU) , Melbourne (AU), Leibnitz (AT), Amostra (BR), Bobo-Dioulasso (BF), Toronto (CA), Vancouver (CA), Santiago (CL), Zagreb (HR), Seoul (KR), Ciudad (MX), Lagos (NG), Luna (RO), New York (US), Tashkent (UZ).
- The total dataset covers around 500 km^2.

These date have been manually labeled to outline building polygons.

4.2 Experimental results

In this section, we validate the results of this chain, in particular reconstruction of building contours, by comparing the polygons reconstructed by our method and state-of-the-art approaches. We have focused our evaluation on taking a closer look at a reconstruction of building polygons, because it is the critical point of our automated workflow.

4.2.1 Comparison with the state-of-the-art methods. We compare the building polygons produced by our framework with the manually-labeled ground truth, as well as with the results of two state-of-the-art approaches:

1. Our U-net-based classification results, polygonized by using the algorithm of Douglas-Peucker (Douglas , Peucker, 1973).
2. The method described in (Duan , Lafarge, 2016), see Section 2 for a summary.

Figure 8 shows the results of these different methods using the same test image over Nairobi as in Section 3. The polygonization based on a Douglas-Peucker algorithm gives an approximation of the contour too rough to be applicable in an industrial context. In particular, polygon simplification can deform too much the shape of buildings and the produced polygons lack geometric regularity (right angles are not well reconstructed). The method of (Duan , Lafarge, 2016) fails if the estimated disparity map is not accurate. The consequence of this are missing buildings or very imprecise contours. Therefore, this algorithm is not optimal for use cases like low buildings, but performs better on the city center with tall buildings as described in (Duan , Lafarge, 2016).

Our method gives very good results: both classification and polygonization yield coherent outputs to be used in an industrial context. The building contours are well regularized and respect the contours of source images in most cases. The designed chain allows us to drastically reduce the manual correction, which we apply at the end to provide the results with a very high precision, as asked by our customer. The mean Intersection Over Union score (Csurka et al., 2013) between the results obtained by our automated framework, and the final corrected results provided to the clients, is 94.25%, where 75.12% of the buildings reconstructed by our chain have remained intact.

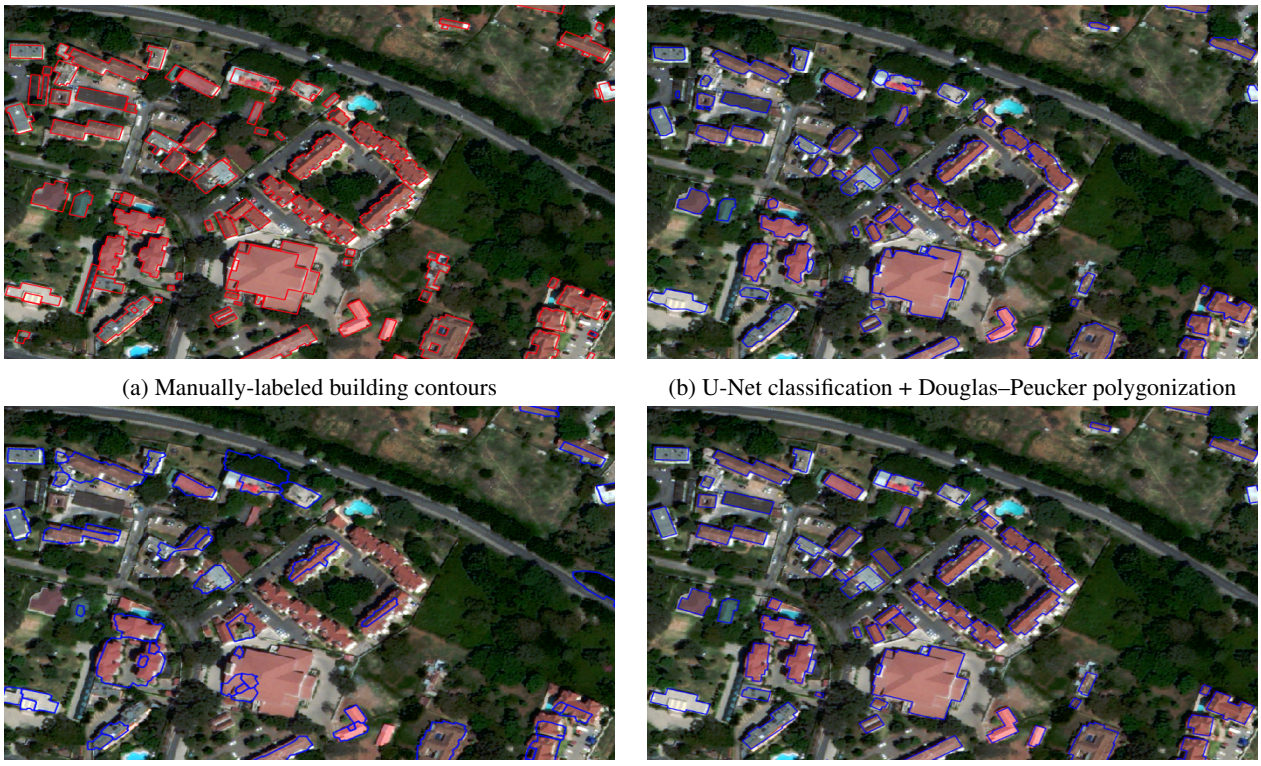
4.2.2 Qualitative analysis. In the previous section, the developed chain has proven its efficiency to provide building contours from a stereo pair of satellite images. In this section, we demonstrate the robustness of our workflow. The critical point of the chain is the capacity of our deep learning method to handle the maximum of use cases. Figure 9 shows the results of our framework, notably reconstructed building contours, on different types of urban areas:

- *Residential area:* Small individual houses are very well detected and small details of the building shapes are preserved. We can also remark that swimming pools are not confused with buildings.
- *Industrial area:* Large buildings are well detected, as well as small buildings. The U-Net architecture show its efficiency to handle different sizes of objects thanks to its different layers.
- *Very dense area:* This kind of areas is very well managed by our automated chain. Even if manually it is not easy to see and separate each individual building, the result is convincing enough and useful from an industrial point of view.

The illustrated examples prove that our model generalizes well to classify different kinds of buildings. Our model has not seen during its training the cities used for evaluation in all figures of this paper; it thus performs very well on new "unseen" cities. Moreover, only a very small part of our data archive was sufficient for training the network, confirming that the U-Net does not need huge volumes of training data.

5. CONCLUSIONS

In this paper, we have presented our automated chain for large-scale 3D reconstruction of city scenes in LOD1 from stereo pairs of satellite images at very high resolution (50 cm/pixel). The developed chain consists of two main parts: extraction of the height information (DSM and DTM) and reconstruction of building polygons. We have described how we extract height information in a robust way and how the deep learning method has helped us to handle the critical point of this chain: extraction of building contours.



(a) Manually-labeled building contours

(b) U-Net classification + Douglas–Peucker polygonization

(c) Method of (Duan , Lafarge, 2016)

(d) Our method

Figure 8. Evaluation of the building polygon extraction algorithm. Comparison with the state-of-the-art.

The proposed chain has been tested on several thousands of km^2 in an industrial context and has proven its efficiency. As shown in the experimental section, it succeeds in providing accurate reconstruction results for different types of urban areas. The use of the deep learning method has allowed us to solve the problem of semantic labeling in new "unseen" cities, that the classical approaches could not solve with an acceptable precision for all use cases.

The developed framework also works very well for 3D reconstruction of natural environments, in particular building 3D maps of trees; however, this topic is out of scope of this paper. In future, we plan to extend our automated chain to enable 3D urban reconstruction in LOD2, where additionally to the information provided in LOD1, for every building a geometrically simplified roof shape is reconstructed.

REFERENCES

Bittner, K., Cui, S., Reinartz, P., 2017. Building Extraction from Remote Sensing Data using fully convolutional Networks. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-1/W1, 481–486.

Budninskiy, M., Liu, B., Tong, Y., Desbrun, M., 2017. Spectral Affine-Kernel Embeddings. *Computer Graphics Forum*, 36, 117-129.

Csurka, G., Larlus, D., Perronnin, F., Meylan, F., 2013. What is a good evaluation measure for semantic segmentation?. *BMVC*, 27.

Douglas, D.H., Peucker, T.K., 1973. Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *Cartographica*, 10, 112–122.

Duan, L., Lafarge, F., 2016. Towards large-scale city reconstruction from satellites. *Computer Vision – ECCV 2016*, Springer International Publishing, Cham, 89–104.

Duan, L., Desbrun, M., Giraud, A., Trastour, F., Laurore, L., 2019. Large-scale dtm generationfrom satellite data. *The IEEE Conference on Computer Visionand Pattern Recognition (CVPR) Workshops*.

Facciolo, G., Franchis, C.D., Meinhardt-Llopis, E., 2017. Automatic 3d reconstruction from multi-date satellite images. *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 1542–1551.

Girard, N., Tarabalka, Y., 2018. End-to-end learning of polygons for remote sensing image classification. *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, 2083–2086.

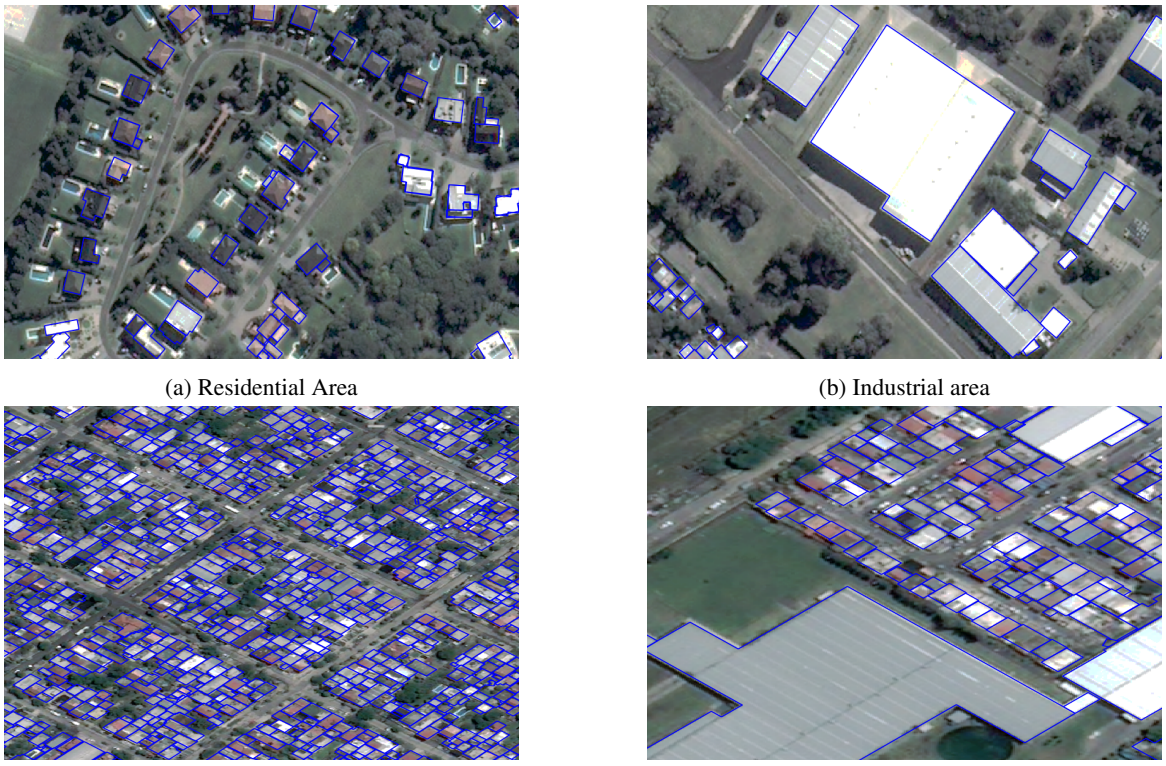
Gribov, A., Bodansky, E., 2004. A new method of polyline approximation. *Structural and Syntactic and and Statistical Pattern Recongnition and ser. Lecture Notes in Computer Science*, 3138, 504–511.

Groger, G., Plumer, L., 2012. CityGML interoperable semantic 3D city models. *Journal of Photogrammetry and Remote Sensing*.

Guo, Z., Yuan, X., 2006. On RPC model of satellite imagery. *Geo-spatial Information Science*, 9, 285–292.

Hane, C., Zach, C., Cohen, A., Pollefeys, M., 2017. Dense semantic 3D reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39.

Hartley, R., Zisserman, A., 2003. *Multiple view geometry in computer vision*. Cambridge university press.



(a) Residential Area (b) Industrial area

(c) Dense Area (d) Industrial and Dense Area

Figure 9. Reconstructed building contours for different types of urban areas.

Hirschmuller, H., 2008. Stereo Processing by Semiglobal Matching and Mutual Information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30, 328-341.

Huang, B., Lu, K., Audebert, N., Khalel, A., Tarabalka, Y., Malof, J., Boulch, A., Saux, B. Le, Collins, L., Bradbury, K., Lefevre, S., El-Saban, M., 2018. Large-scale semantic classification: outcome of the first year of inria aerial image labeling benchmark. *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, 6947–6950.

Lowe, D., 2004. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60, 91-110.

Marcos, D., Tuia, D., Kellenberger, B., Zhang, L., Bai, M., Liao, R., Urtasun, R., 2018. Learning deep structured active contours end-to-end. *2018 IEEE Conference on Computer Vision and Pattern Recognition*, 8877–8885.

Musialski, P., Wonka, P., Aliaga, D., Wimmer, M., van Gool, L., Purgathofer, W., 2012. A survey of urban reconstruction. *EUROGRAPHICS 2012 State of the Art Reports*, 1–28.

Pesaresi, M., Benediktsson, J.A., 2001. A new approach for the morphological segmentation of high-resolution satellite imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 309-320.

Poli, D., Caravaggi, I., 2013. 3D modeling of large urban areas with stereo VHR satellite imagery: lessons learned. *Natural Hazards*, 68.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Springer, 234–241.

Tareen, S.A.K., Saleem, Z., 2018. A comparative analysis of sift and surf and kaze and akaze and orb and and brisk. *iCoMET*.

Tasar, O., Maggiori, E., Alliez, P., Tarabalka, Y., 2018. Polygonization of binary classification maps using mesh approximation with right angle regularity. *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, 6404–6407.

Tripodi, S., Duan, L., Tasar, O., Tarabalka, Y., Clerc, S., d'Andon, O. Fanton, Trastour, F., Laurore, L., 2018. Automatic and robust chain for urban reconstruction from satellite imagery. *Phi-week Workshop of ESA*.

Vanegas, C.A., Aliaga, D.G., Beneš, B., 2010. Building reconstruction using manhattan-world grammars. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 358–365.

Verdie, Y., Lafarge, F., Alliez, P., 2015. LOD generation for urban scenes. *ACM Transactions on Graphics*, 34.

Volpi, M., Tuia, D., 2017. Dense semantic labeling of sub-decimeter resolution images with convolutional neural networks. *International Journal of Computer Vision*, 55.

Wang, K., Frahm, J., 2017. Fast and accurate satellite multi-view stereo using edge-aware interpolation. *2017 International Conference on 3D Vision (3DV)*, 365–373.

Zhu, X.X., Tuia, D., Mou, L., Xia, G.-S., Zhang, L., Xu, F., Fraundorfer, F., 2017. Deep learning in remote sensing: a comprehensive review and list of resources. *IEEE Geoscience and remote sensing magazine*, 5.