AUTOMATIC ENRICHMENT OF INDOOR 3D MODELS USING A DEEP LEARNING APPROACH BASED ON SINGLE IMAGES WITH UNKNOWN CAMERA POSES

M. Jarząbek-Rychard [a,b] *, H-G. Maas [b]

[a] Institute of Geodesy and Geoinformatics, Wroclaw University of Environmental and Life Sciences, Poland - malgorzata.jarzabek-rychard@upwr.edu.pl
[b] Institute of Photogrammetry and Remote Sensing, Technische Universität Dresden, Germany - hans-gerd.maas@tu-dresden.de

Commission V, WG V/7

KEY WORDS: Building Information Model (BIM), deep learning, object recognition, texture mapping, camera pose estimation

ABSTRACT:

3D building modeling is a diverse field of research with a multitude of challenges, where data integration is an inherent component. The intensively growing market of BIM-related consumer applications requires methods and algorithms that enable efficient updates of existing 3D models without the need for cost-intensive data capturing and repetitive reconstruction processes. We propose a novel approach for semantic enrichment of existing indoor models by window objects, based on amateur camera RGB images with unknown exterior orientation parameters. The core idea of the approach is the parallel estimation of image camera poses with semantic recognition of target objects and their automatic mapping onto a 3D vector model. The presented solution goes beyond pure texture matching and links deep learning detection techniques with camera pose estimation and 3D reconstruction. To evaluate the performance of our procedure, we compare the estimated camera parameters with reference data, obtaining median values of 13.8 cm for the camera position and 1.1° for its orientation. Furthermore, a quality of 3D mapping is assessed based on the comparison to the reference 3D point cloud. All the windows presented in the data source were detected successfully, with a mean distance between both point sets equal to 3.6 cm. The experimental results prove that the presented approach achieves accurate integration of objects extracted from single images with an input 3D model, allowing for an effective increase of its semantic coverage.

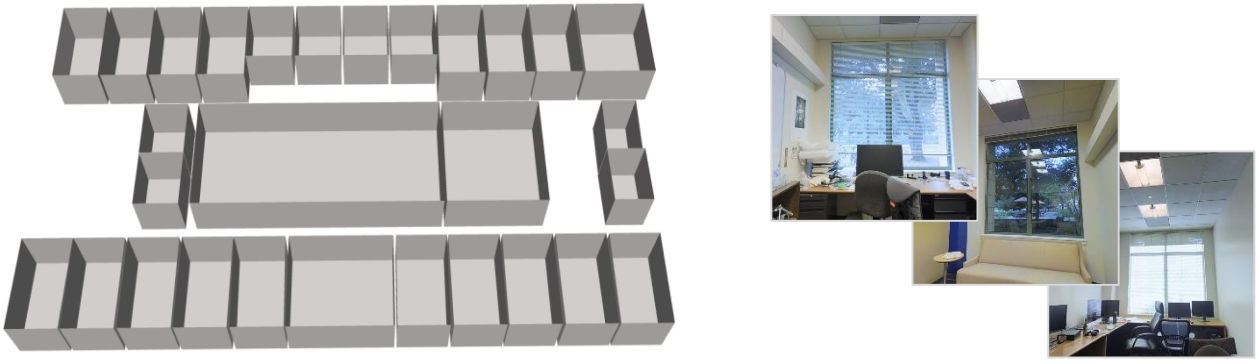
1. INTRODUCTION

Building Information Modeling (BIM) is a diverse and multidisciplinary research subject with steadily increasing interest and demand (Czerniawski and Leite, 2020; Pintore et al., 2020). In recent years, we could observe that the scope of BIM-related topics is not anymore limited to purely professional usage. More frequently it leaves space for various consumer products that rely on realistic 3D models created from spatial data. The intensively growing market of BIM applications triggers the need for algorithms that enable efficient update and interaction with building models.

For a long time, the common concern in indoor modeling for BIM support has been mostly focused on the geometric reconstruction of the main structural building elements: walls, ceilings, and floors (Quintana et al., 2016; Ochman et al., 2019; Nikoohemat et al., 2020). The second, and significantly less covered, core task of as-built BIM creation is object recognition, the process of labeling parts of data or extracted primitives with semantic classes. Although windows and doors constitute a large proportion of indoor objects, many existing studies lack the capabilities to model them (Mura et al., 2014; Thomson and Boehm 2015; Anagnostopoulos et al., 2016). Detection of wall openings in indoor scenes may be a relatively complex task due to the existence of other objects, like pieces of furniture, that cause occlusions. The prevailing approach within the presented in the literature methods that can reconstruct windows objects is to use point clouds obtained by laser scanning, where wall openings appear as holes in a data (Michailidis et al., 2017; Jung et al., 2018; Previtali et al., 2018). In alternative solutions, windows were detected in photogrammetric point clouds (Jarząbek-Rychard and Maas, 2020) or image textures (Tang et al., 2019). Texture mapping, however, often requires tedious human intervention and is usually performed during the reconstruction process, which means that the modeling is not able to accommodate modifications after the model is created. The very complex and frequently changing environment of indoor spaces triggers the need for techniques that allow for an automatic update of existing 3D models, which increases their semantic coverage without the need for costly data capturing and repetitive reconstruction process.

This paper presents a novel approach for semantic enrichment of existing indoor 3D vector models by window objects, based on single RGB images taken at hand (Fig.1b). An important assumption underlying the presented methodology is its feasibility for consumer application, which means that the input images are assumed to have an unknown camera pose. The only given information is an association of the image to the room where it was captured. The innovative idea of the approach is the parallel semantic recognition of the target objects in the image with its automatic positioning in 3D space. Our solution goes beyond pure texture matching and bridges deep learning detection techniques with camera pose estimation and 3D reconstruction. The presented algorithm starts with a joint recognition of window target objects and the spatial layout of the corresponding indoor space. For this purpose, we employ an object detection convolutional neural network Mask R-CNN to semantically recognize desirable building elements and their bounding regions. The detected ceiling region is then processed by a morphological operation, allowing for the extraction of its bounding lines, which are matched with the corresponding 3D

* Corresponding author



(a) (b)

Figure 1. Input data: 3D vector model of a building interior (a), single RGB images of windows (b).

edges in the model. The estimated 2D/3D correspondences allow for the estimation of camera poses formulated as the Perspetive-3-Line (P3L) problem (Xu et al., 2017). Finally, window 3D objects are reconstructed based on the photogrammetric projection of the detected pixels into 3D space, and regularized. To evaluate the performance of our procedure, we compare the estimated values of camera poses and orientation with reference data. Furthermore, a quality assessment of 3D mapping is performed based on the comparison to the reference 3D point cloud. The experimental results prove that the presented approach allows for effective integration of a 3D model with objects detected in 2D single images, achieving an efficient automatic upgrade of the model's semantic level.

2. METHODOLOGY

2.1 Object recognition using deep learning methods

The presented methodologic pipeline of image processing (Fig.2) starts with the detection of chosen indoor objects in RGB images. Besides target windows, the algorithm aims at the recognition of walls and ceiling objects for the subsequent extraction of an indoor space layout (edges of an indoor cube). To perform this task, we select Mask R-CNN (He et al., 2017), a robust deep learning detection framework. Extended from the faster R-CNN architecture (Ren, et al., 2015), Mask R-CNN provides a mask prediction branch composed of a small Fully Convolutional Network for segmenting each Region of Interest (ROI) with simultaneous classification prediction and bounding-box prediction. In the final output, besides the class label and bounding box for each ROI, Mask R-CNN additionally generates a binary mask of each detected object. Example results of indoor object recognition are presented in Fig.3a. At that stage, the

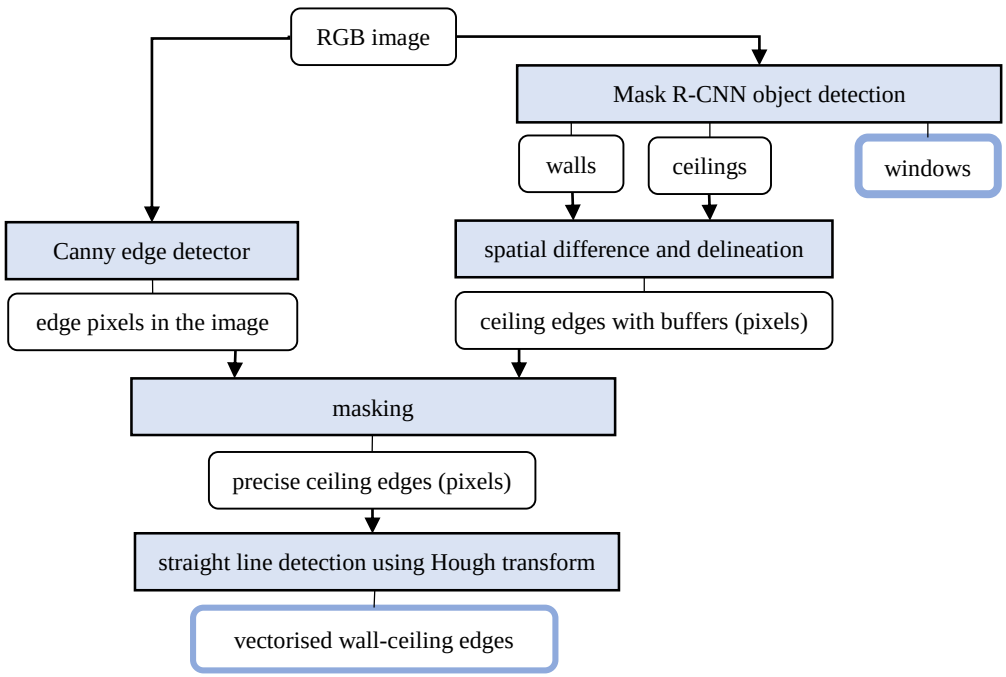


Figure 2. General workflow of the image processing algorithm and its final results: detected window pixels and the 2D vectorized edges of an indoor space.

detected objects may slightly overlap with each other and have irregular outlines. To improve the initial hypothesis of the ceiling boundary, the spatial difference with wall objects is applied. The final binary masks of the detected ceiling and windows are presented in Fig.3c and Fig.3e.

2.2 Extraction of indoor space edges

Although deep learning techniques are powerful tools for region object detection, they show deficits in the precise extraction of its bounding edges, which are crucial for the subsequent estimation of the camera poses. Therefore, in our approach (illustrated in Fig.4) contours extracted from the binary masks provided in the previous step serve as initial information for more accurate extraction of the ceiling-wall edges. To automatically detect pixels belonging to the precise object boundaries in RGB photos, we apply one of the common edge detection methods. For the presented research we choose Canny edge detector. The algorithm, composed of several stages, is much more complex than other detection methods (like e.g. Sobel or Prewitt operator), thus it enables to effectively detect a wide range of edges in the image. In the next step, the binary image with marked edges is masked with a buffer of n pixels (in the following experiments n= 3) around the delineated ceiling object extracted by Mask R-CNN. The resulting region of interest with marked edge pixels serves as an input for the vectorization of edges performed by the

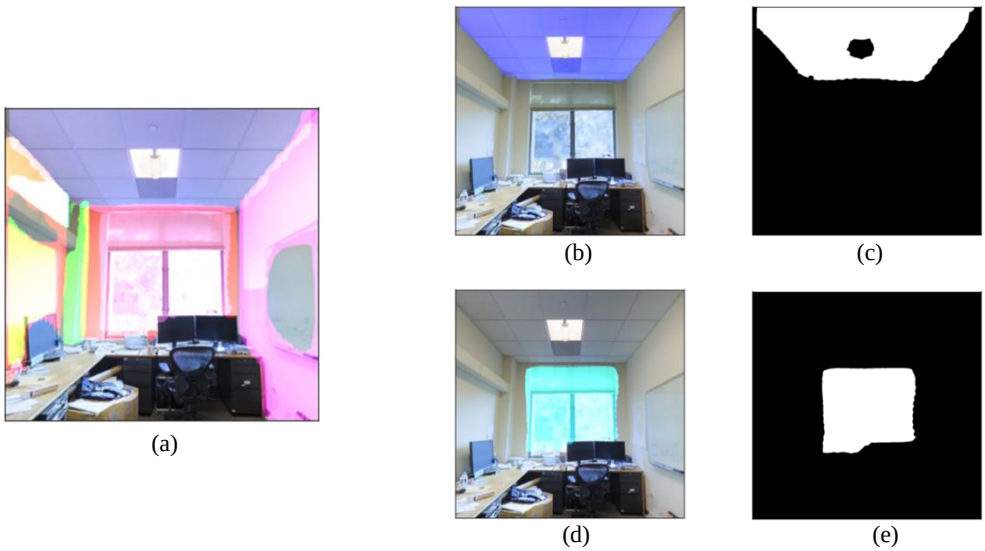


Figure 3. Object detection in RGB image using Mask R-CNN: extracted objects (windows, ceilings, walls), partially overlapped (a), detected separated objects and their corresponding masks: ceiling (b) and (c) and window (d) and (e).

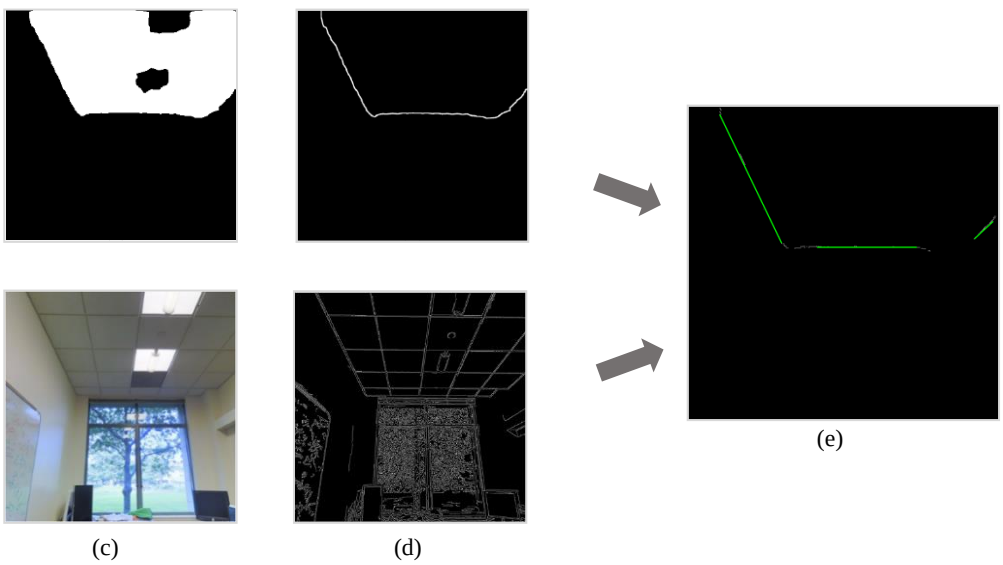


Figure 4. Edge detection process: ceiling mask extracted by R-CNN (a), detected outer contour (b), corresponding RGB image (c), extracted image edges (d), edges derived by masking image edges using detected outer contour (white) and the final straight lines detected by Hough transform (green) (e).

Hough transform. Straight-line segments and their corresponding line equations are detected by finding peaks in the Hough space of the binary image. According to the assumption underlying the Hough transform, the algorithm returns numerous hypotheses of line positions, based on the same pixels connected in multiple configurations. Thus, to reduce the search space for potential pixels belonging to the estimated line, the number of peaks in the Hough space is limited to 3, and the part of the image within the buffer of 3 pixels from the extracted line is excluded from the processing. The finally extracted ceiling edges are shown in Fig.4e. To establish correspondences between the edges of a 3D model and the edges extracted from the image, the lines are sorted according to their coordinates and stored as a topological chain.

2.3 Camera pose estimation from line correspondences

Camera pose estimation is an important step in a broad range of applications, related to augmented reality, robotics, and computer vision. The pose of a calibrated camera is usually determined by analysing *n* correspondences between 3D reference features and their 2D projections. Contrary to the well-studied Perspetive-n-Point (PnP) problem that utilizes point features (e.g., Lepetit et al., 2009), solutions dedicated to line features (known as, Perspective-n-Line problem, PnL) remains challenging. For the determination of the camera pose in 3D space there are six degrees of freedom. Thus, to get a finite set of solutions, at least three correspondences should be given, since each correspondence offers two dimensions of constraint on the pose parameters. In our research, because of the limited features of the input 3D model (vertices and edges of a cuboid) and their visibility in the images, only the line-based solutions can be applied. Furthermore, due to the clutters that usually occlude wall-floor edges, and shadows that hinder detection of vertical edges, we limited the set of the corresponding lines to 3 ceiling-wall edges. Using the minimum possible size of the correspondence set, we thus face the special case of the PnL, i.e. the Perspetive-3-Line (P3L) problem (described in detail by Xu et al., 2017). Given a calibrated camera and three reference lines $L_i (i \in \{0,1,2\})$ with their corresponding 2D projections in the image space as l_i, the camera pose can be determined based on these 3D/2D correspondences ($L_i \Leftrightarrow l_i$). The problem involves two tasks: estimating the rotation matrix $\boldsymbol{R}$ and the translation vector $\boldsymbol{t} \in R^3$. In general, it involves solving nonlinear equations of 8-th order polynomials. It is then well known that the solution to the P3L problem is not uniquely determined. On the other hand, the potential solutions obviously differ from each other, which makes it possible to choose the real one based on the roughly given initialization - in the presented experiment, we use for this purpose the initial camera location calculated as the centre of gravity of the vertices in each cuboid (indoor space).

2.4 3D reconstruction of the detected objects

The core idea of 3D window object reconstruction is based on the photogrammetric projection of the detected window pixels on the given 3D vector model. The projection is performed through the collinearity equations, along with the previously extracted exterior orientation parameters of the camera. A set of 3D window points is computed by the intersection of walls (provided by the input model) with viewing rays assigned with the pixels of windows binary masks, detected by Mask R-CNN (Section 2.1). 3D coordinates of window corner points are estimated by the best fitting rectangle algorithm (assuming window edges to be vertical and horizontal in 3D space) performed on the projected window contour points derived by the convex hull method. Due to the many occlusions of the lower part of the window (typically computer screens and other things placed on the desktops), its bottom outline is often very irregular. Therefore, the extracted lower edge of the window is shifted to the lowest detected window point. As the result of fitting objects to the computed 3D points, we obtain rectangles of varying sizes, which are not always properly aligned. In the last step of the presented pipeline, window vector models are subjected to global regularization. The objects are grouped according to two regularities constraints: same shape and same vertical alignments. The final windows layout is obtained by enforcing positional and shape changes according to median values calculated for each group.

3. RESULTS AND DISCUSSION

To verify the performance of the presented approach, we use part of the data belonging to the Stanford 2D-3D-Semantic Dataset (2D-3D-S) (Armeni et al., 2016). The dataset provides a variety of mutually registered modalities from 2D, 2.5D, and 3D domains, with instance-level semantic and geometric annotations. The data are collected in 6 large-scale indoor areas that originate from 3 different buildings of mainly educational and office use. For our experiment we choose three subsets of the data: i) 3D point cloud of an indoor area, which served as a base for 3D modeling of an input indoor model, ii) RGB images providing additional texture information for this area (with visible windows), with an association to the corresponding building spaces, iii) RGB images belonging to other buildings, used for transfer learning by Mask R-CNN. Since our image-model matching procedure is based on a single photo without exterior orientation parameters, the images belonging to the second subset have to present 3 visible ceiling-wall edges for correspondences matching. Such images were not provided for 3 out of the 23 offices. Besides the listed subsets, we also use camera information as reference data for the verification of the pose estimation algorithm performance. The initial indoor model subjected to the refinement is reconstructed according to the modified 3D modeling approach described in (Jarząbek-Rychard and Borkowski, 2016) adapted for indoor scenes. As direct input data, the reconstruction algorithm uses only wall points, according to the semantic classification provided by the benchmark. The model is stored as a list of *x,y,z* coordinates of wall vertices and their topological relations (connecting edges).

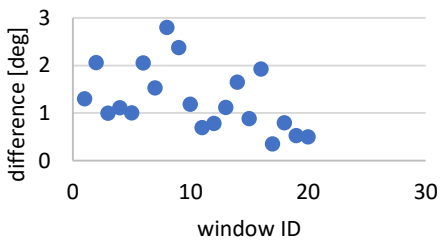


Figure 5. Angular differences between the estimated camera orientation and the reference data.

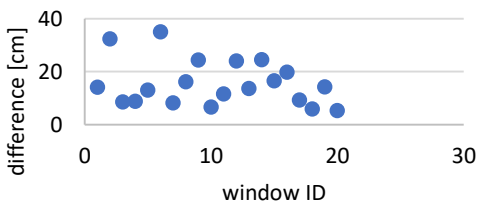


Figure 6. 3D shifts between the estimated camera position and the reference data.

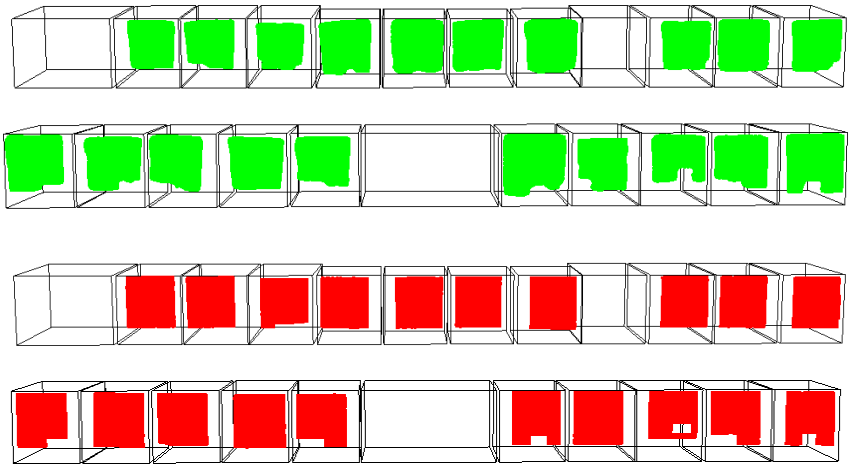


Figure 7. Detected window pixels projected on the 3D model (dense 3D points) (green), and reference 3D point cloud (red).

To train Mask R-CNN for the detection of chosen indoor objects (windows, ceilings, walls) we use a network with a backbone Res-Net-50, pretrained on the open deep learning dataset ImageNET (Deng et al., 2009) and apply transfer learning techniques. The dataset consists of 100 images, with corresponding ground-truth annotations with the chosen 3 semantic classes. We split the data into 80 images for training and 20 images for validation. The test dataset consists of 20 images captured in the target building and directly serving for its further enrichment procedure. As an optimization algorithm we used Stochastic Gradient Descent (SGD), the learning rate is set to 0.005, the weight decay to 0.0005, and the momentum to 0.9. The method was implemented in PyTorch and processed using free NVIDIA Tesla P100 GPU provided by Google Colab. The obtained results serve as a base for the subsequent 2D/3D line correspondences and the estimation of camera poses (described in Section 2.2 and 2.3).

The evaluation of the camera pose estimation is performed based on the comparison with the reference data, computed for each image matched with the 3D model. The quality analysis is built on the differences calculated for two indicators: angular camera orientation (Fig.5) and 3D camera position (Fig.6). The absolute values of angular discrepancies oscillate between 0.3° and 2.7°, giving the median value of 1.1°. The median planar displacement of the camera position is equal to 13.8 cm, with the values in the range of 5-35 cm. The estimated exterior orientation parameters enable to project the detected window pixels on the 3D walls of the input 3D model. To assess the performance of the automatic 3D mapping, the computed 3D window points are compared to the semantically classified reference 3D point cloud. The mean approximate distance between both point sets is equal to 3.6 cm. There is one outlier at 64 cm, while for most of the points the difference is close to 0 cm. The visual comparison is presented in Fig.7. Except for one case, when the window is falsely extended to the edge of the wall, all projected window points form proper layout and shapes, even preserving the shape of occlusions. The main difference between the results and the reference set is in the sharpness of window boundaries. The final building model enriched in window vector models, after regularization of their global layout, is shown in Fig.8. The visual assessment demonstrates that the proposed method can preserve a high accuracy fit between the objects detected in the images and the input indoor model.

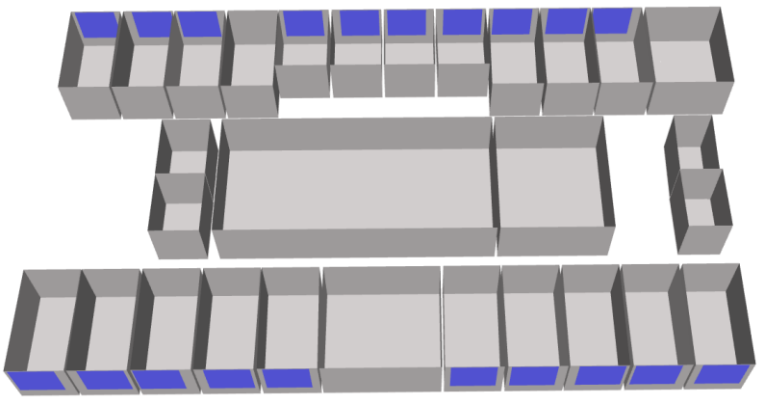


Figure 8. Final results: input 3D model automatically enriched in window vector objects.

4. CONCLUSION

This paper presents an innovative approach to the automatic upgrade of existing indoor 3D models using up-to-date semantic information extracted from single RGB images with unknown camera poses. The new insight of the research covers the parallel estimation of the image exterior orientation parameters with recognition of target objects and their 3D reconstruction. To solve these problems, we propose a novel methodology that employs deep convolutional neural networks together with projective photogrammetry. Although the presented experiments are focused on the detection and modeling of window objects, the algorithms behind can be easily adapted for other planar objects visible in indoor scenes. The evaluation of the procedure for extraction and matching of the correspondences between input 3D model and images shows that the method allows estimating camera poses with a median value of 13.8 cm for the camera position and 1.1° for its orientation. The quality assessment of 3D object mapping, revealing 3.6 cm of the mean approximate distance, indicates a high correlation between extracted object points and the reference data. In this study, we confirmed that the proposed approach can achieve effective integration of vector 3D models with objects detected in single images acquired in indoor scenes. The underlying methodology so far assumes that the internal camera parameters are available from a priori calibration. In future work, we aim at the comparison of the performance of the presented pose estimation approach against other available in the literature methods dedicated to indoor application. We also plan to increase the applicability of the method allowing for the use of the images captured by smartphone cameras, which often show instable interior orientation (Elias et al., 2020). Furthermore, we intend to extend the scope of BIM-related information by adding the possibility to detect other indoor objects (e.g. doors, radiators, furniture), and investigating thermal infrared images as an additional data source.

ACKNOWLEDGEMENT

The project is financed by the Polish National Agency for Academic Exchange as part of the Mieczysław Bekker Programme.

REFERENCES

Anagnostopoulos, I., Belsky, M., Brilakis I., 2016. Object boundaries and room detection in as-is BIM models from point cloud data, *Proceedings in the International Conference on Computing in Civil and Building Engineering, ISCCBE, Osaka, Japan*, pp. 968–974.

Armeni, I., Sener, O., Zamir, A. R., Jiang, H., Brilakis, I., Fischer, M., Savarese, S., 2016: 3D semantic parsing of large-scale indoor spaces. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1534–1543.

Czerniawski, T., Leite, F., 2020: Automated digital modeling of existing buildings: A review of visual object recognition methods. Automation in Construction 113, 103131.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009: Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255.

Elias, M., Eltner, A., Liebold, F., Maas, H. G., 2020: Assessing the influence of temperature changes on the geometric stability of smartphone-and raspberry Pi cameras. *Sensors*, 20(3), 643.

He, K., Gkioxari, G., Dollar, P., Girshick, R., 2017: Mask R-CNN. *IEEE Trans. Pattern Anal. Mach. Intell.* 42 (2), 386–397.

Jarząbek-Rychard M., Borkowski A., 2016: 3D building reconstruction from ALS data using unambiguous decomposition into elementary structures. *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 118, 1-12.

Jarząbek-Rychard M., Maas H.-G., 2020: Supervised Detection of Façade Openings in 3D Point Clouds with Thermal Attributes. *Remote Sensing,* Vol. 12 (3) No. 543, 1-17.

Jung, J., Stachniss, C., Ju, S., Heo, J., 2018: Automated 3D volumetric reconstruction of multiple-room building interiors for as-built BIM. *Advanced Engineering Informatics*, *38*, 811-825.

Lepetit, V., Moreno-Noguer, F., Fua, P., 2009: EPnP: An accurate O(n)solution to the PnP problem," *IJCV*, vol. 81, no. 2, 155–66.

Michailidis, G. T., & Pajarola, R., 2017: Bayesian graph-cut optimization for wall surfaces reconstruction in indoor environments. *The Visual Computer*, *33*(10), 1347-1355.

Mura, C., Mattausch, O., Villanueva, A.J., Gobbetti, E. Pajarola, R., 2014: Automatic roomdetection and reconstruction in cluttered indoor environments with complex roomlayouts, Comput. Graph. 44, 20–32.

Nikoohemat, S.; Diakité, A.A.; Zlatanova, S.; Vosselman, G., 2020: Indoor 3D reconstruction from point clouds for optimal routing in complex buildings to support disaster management. *Automation in Construction*., 113, 103109.

Ochmann, S., Vock, R., & Klein, R., 2019: Automatic reconstruction of fully volumetric 3D building models from oriented point clouds. *ISPRS journal of photogrammetry and remote sensing*, *151*, 251-262.

Pintore, G., Mura, C., Ganovelli, F., Fuentes-Perez, L., Pajarola, R., Gobbetti, E., 2020: State-of-the-art in Automatic 3D Reconstruction of Structured Indoor Environments. EUROGRAPHICS, Vol. 39, No.2

Previtali, M., Díaz-Vilariño, L., and Scaioni, M., 2018: Towards automatic reconstruction of indoor scenes from incomplete point clouds: door and window detection and regularization. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.,* XLII-4, 507–514.

Thomson, C., Boehm, J., 2015: Automatic geometry generation from point clouds for BIM, *Advanced Engineering Informatics* 38, 811–825.

Quintana, B., Prieto, S, Adán, A., Vázquez, A.S., 2016: Semantic scan planning for indoor structural elements of buildings, Adv. Eng. Inf. 30, pp. 643–659.

Ren, S., He, K., Girshick, R. B., Faster, Sun,J., 2015: R-CNN: towards real-time objectdetection with region proposal networks. *CoRR abs/1506.01497.*

Tang, S.; Zhang, Y.; Li, Y.; Yuan, Z.; Wang, Y.; Zhang, X.; Li, X.; Zhang, Y.; Guo, R.; Wang, W., 2019: Fast and Automatic Reconstruction of Semantically Rich 3D Indoor Maps from Low-quality RGB-D Sequences. *Sensors,* 19, 533

Xu, C., Zhang, L., Cheng, L., Koch, R., 2017: Pose Estimation from Line Correspondences: A Complete Analysis and a Series of Solutions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, 1209-1222.