BUILDING DETECTION AND ROOF TYPE CLASSIFICATION IN LOW RESOLUTION PHOTOGRAMMETRIC POINT CLOUDS FROM AERIAL IMAGERY

Maria Axelsson, Jörgen Karlholm, Ulf Söderman

Swedish Defence Research Agency (FOI), Linköping, Sweden

ABSTRACT

Building footprint detection and three-dimensional (3D) reconstruction from aerial imagery is of interest in many urban applications e.g. city planning, state cadastral inspection, environmental simulations, and radio transmission simulations. In applications where large land areas need to be covered regularly to update data it is not practical to use laser scanning or acquire aerial images with high resolution and large image overlaps. In these applications the reconstructed photogrammetric point cloud may have low resolution with less building details. We present two deep learning approaches that support the applications mentioned above. One example where semantic segmentation is applied to detect building footprints and another example of roof type classification using a deep convolutional neural network pre-trained using RGB data. Results are shown for a low resolution dense photogrammetric point cloud generated using multi-view stereo reconstruction of standard overlapping aerial images from nationwide data collection in Sweden.

1. INTRODUCTION

Today, multi-view stereo reconstruction of 3D geometry from two-dimensional (2D) images is well studied and used in large scale applications. Dense photogrammetric point clouds of large areas can be generated from highly overlapping aerial images and are common in city modeling, see e.g. [1]. These point clouds contain both height information and spectral information which makes them attractive for reconstructing buildings in 3D as they can be texture mapped using the spectral information. Approaches based on imagery from e.g. aerial images with small ground sampling distance or laser scanning, provide point clouds with high resolution and many resolved building details. However, in some applications it is not practical to acquire images with large image overlaps and small ground sampling distance or use laser scanning. One such example is when large areas needs to be covered regularly to keep the point cloud up to date, e.g. when generating a 3D map for an entire country on a regular basis. In this application the point cloud can be generated using aerial images with smaller image overlaps and imaged at larger distances. This gives a photogrammetric point cloud with low resolution, which is more challenging to use for building detection and reconstruction than high resolution photogrammetric point clouds or a point cloud from laser scanning.

In this paper we present two deep learning approaches for urban applications. One method using semantic segmentation for building footprint detection (Section 2) and a method for automatic classification of the most common model building shapes, ridge roofs and flat roofs, using deep convolutional neural networks (CNN) (Section 3). We show results using a relatively low resolution photogrammetric point cloud, a digital surface model, generated using multi-view stereo reconstruction from high altitude aerial imagery with relatively small image overlap. The aerial images are overlapping in both the flight direction with 60% and in cross direction with 25%. A dense photogrammetric point cloud is generated from the images using multi-view stereo with Semi-Global Matching and fusion of depth where redundant depth estimates from overlapping stereo models are merged [2]. Due to the image overlap the number of available images for an area varies from two to six. Stereo models are calculated from each image pair and depths are fused. The resulting 2.5D point cloud, which is called Digital Surface Model from Aerial Photos[1], is sampled on a regular grid of 0.5×0.5 m. Each point that is matched contains both spectral and height information.

2. BUILDING FOOTPRINT DETECTION

Semantic segmentation is the task of assigning class membership to each pixel in an image, where class may refer to object or material category. It is often useful to estimate the probability of each class of interest. The task can be performed by a fully convolutional neural network [3, 4, 5]. Networks for semantic segmentation generally have an hourglass-like shape consisting of two stages. First a feature extractor (or encoder) creates a set of representations of the image in the form of a sequence of network layers, where the spatial resolution is gradually reduced using average or max-pooling and subsampling, while simultaneously the number of channels in each layer is increased. The encoder is followed by a decoder network that gradually increases the spatial resolution and reduces the number of channels per layer. The up-sampling can

[1] In Swedish: "Ytmodell från flygbilder", see: www.lantmateriet.se

be achieved using standard (nearest neighbour or bilinear) interpolation, or by transposed convolution, where the network learns interpolation kernels from data. The number of channels in the output layer equals the number of classes. The class probabilities are usually modelled by a softmax function, and the network is trained to minimise the cross-entropy loss. Designs as those cited above mainly differ in how layers of equal spatial resolution in the encoder and decoder stages are connected (sometimes referred to as skip connections). In the present study the refinement module of Pinheiro et al. [6] was used for combining encoder and decoder layers. While FCN [3] and SegNet [5] are more simple designs, the principal differences compared to U-Net [4] are that the latter combines encoder and decoder layer outputs using concatenation, and uses transposed convolution for up-sampling, whereas Pinheiro et al. use bilinear interpolation and a more computationally efficient addition of encoder and decoder layers preceded by dimension reduction of the encoder layer.

In view of the relatively small number of training examples available, and the fact that the network was trained from scratch, the encoder network was designed with a small number of layers in comparison with many popular deep convolutional networks. The network design is shown in Fig. 1. The network was trained on imagery covering approximately 100 square kilometres of the city of Linköping and surrounding countryside containing 19,700 annotated buildings. The annotations are based on the polygons from the building footprint. The data was divided into 1,600 512×512-pixel images at 0.5 m resolution. The dataset was augmented by a factor of eight using horizontal mirroring and rotation by multiples of 90 degrees. Adam [7] was used for minimising the cross-entropy loss. The network was implemented in TensorFlow and trained on a workstation with four NVIDIA Geforce GTX 1080Ti graphics cards. The performance of the resulting network was evaluated on 1,600 images covering 100 square kilometres of the city of Norrköping and surrounding countryside. This data contained approximately 20,500 annotated buildings. The network outputs class probability estimates for each pixel, so a precision-recall curve can be produced, see Fig. 2. It is seen that a precision of 0.89 is obtained for a recall of 89% of all true building pixels. Most of the misclassifications occur at the building perimeters, but it should be noted that many small buildings are fully or partially occluded by tree crowns, and therefore cannot be detected from above. See Fig. 3 for a qualitative example of the network input, ground truth, and prediction.

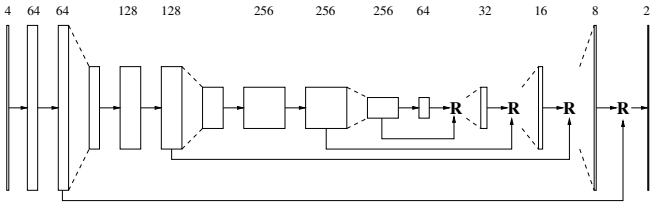

Fig. 1. Convolutional neural network design for building detection based on semantic segmentation. Horizontal arrows indicate 3×3 convolution followed by ReLU (rectified linear unit) and batch normalization, except to the output layer where a linear 1×1 projection is used. Dashed lines indicate 2×2 max-pooling for sub-sampling, and bilinear interpolation for up-sampling. R refers to refinement module, see [6]. The number of channels in each layer is shown at the top.

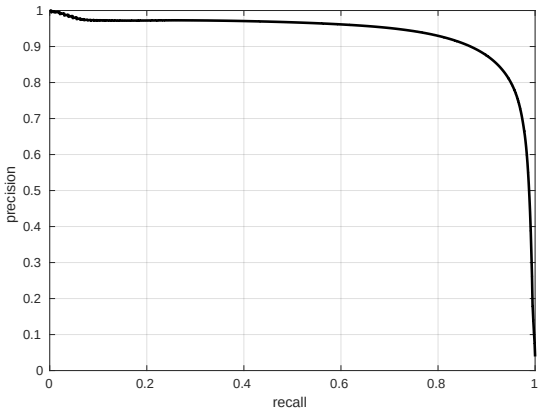

Fig. 2. Precision-recall curve for pixel classification of validation data.

3. ROOF TYPE CLASSIFICATION

In our framework for roof type classification we classify patches of buildings into the two most common roof types, ridge roofs and flat roofs. Ridge roofs include different types of ridge roofs such as gable, half-hip, hip, and mansard roofs. As only limited annotated data for building classification is available the building roof type classification is performed using transfer learning of a pre-trained CNN on RGB data.

The network architecture is illustrated in Fig. 4. It is a CNN where the input is a three band image of size $32 \times 32 \times 3$. The input size is well suited to the problem of building classification since many of the buildings fit well into this size without much interpolation. The network consists of three iterations of convolutional layers followed by ReLU and max poolings and two fully connected layers, where the first is followed by ReLU, and in the end a softmax layer followed by a classifier. The classifier outputs two classes, one for each roof type. The image input use zero-center normalization of the data. We initialize the network using weights from a pre-trained network for object classification using CIFAR10 data [8] which is common RGB data. In our data the three spectral channels contain near infrared (NIR), red, and green additional to the height information from the point cloud. Combinations of these four input channels in the training are evaluated in our experiments.

We base the preprocessing of the data on the 2D polygon associated with each building. A patch for each building

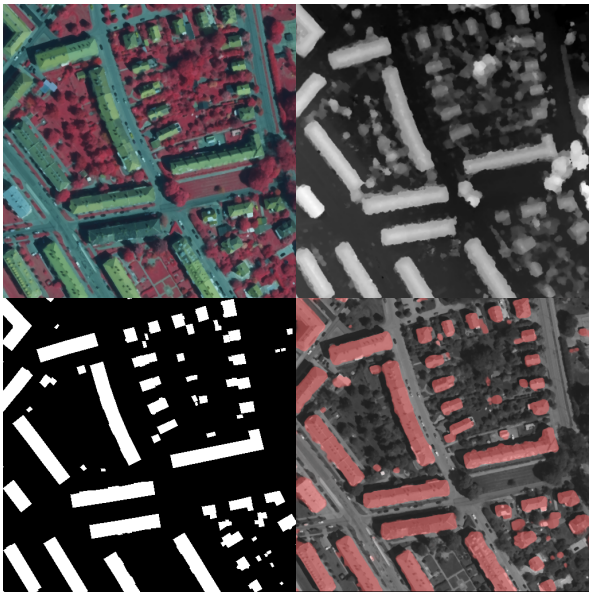

Fig. 3. Building footprint detection example. Top left: Near infrared (NIR), red, and green channels visualised as RGB. Top right: Height information. Bottom left: Ground truth annotations. Bottom right: Prediction in red. Note that several buildings are occluded by trees and therefore not visible from above.

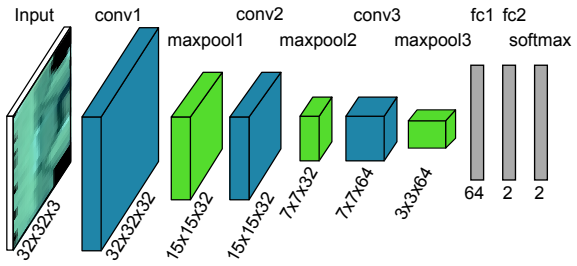

Fig. 4. Convolutional neural network for building roof type classification. The input is a three band image of size $32 \times 32 \times 3$. The network consists of three iterations of convolutional layers and two fully connected layers and a softmax layer followed by a classifier.

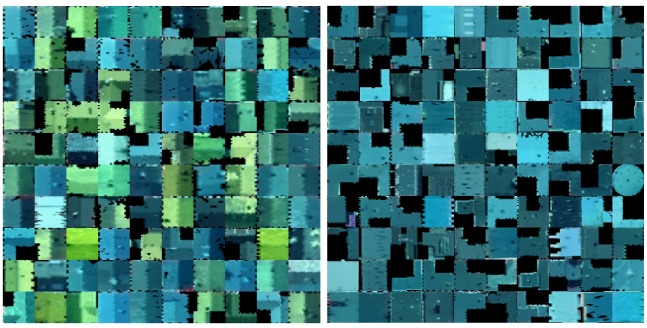

Fig. 5. Examples of training data patches for ridge roofs and flat roofs using the spectral bands in near infrared, red, and green as the three input layers.

is cropped from the point cloud using the building polygon and the background is set to zero. The 2D building polygon outlines the projection of the building on the ground. The building patches are also rotated to align the main axis with the image coordinates using the lengths of the segments in the building polygon. Depending on the building shape the main axis can be aligned both horizontally or vertically in the image. After rotation the building patches are resampled to $32 \times 32 \times 3$ pixels to fit the input layer. Examples of training patches with the bands NIR, red, and green are shown in Fig. 5. Before training the patches are also augmented using rotation and flipping to create more training data using the annotated data. This also makes the two main directions equal and removes any differences in the alignment after rotating the patches.

The aerial images used in the experiments cover approximately 6.6×3.7 km on the ground with a ground sampling distance of about 0.25 m. In addition to the point cloud, 2D building polygons of the building footprint are used to crop out the relevant point cloud area for each building or building part. Also a Digital Terrain Model (the National elevation model) with resolution 1×1 m is used to recover the building height over the local terrain.

The proposed classification method using CNNs has been evaluated using buildings with manually marked roof types from two classes, houses with ridge roofs and houses with flat or very low-slope roofs. The training set contains 1,200 ridge roofs and 400 flat roofs and the test set contains 403 ridge roofs and 197 flat roofs. Multiple copies of the flat roof data were added to the training set to remove unwanted bias towards the ridge roof class. The network was trained using stochastic gradient descent with momentum using the two classes. In our experiments we have evaluated different combinations of the three spectral bands and the height information as the three input layers. The best result was obtained by combining height, red, and green as the three input layers, but almost the same result was obtained using only the spectral information from the data using NIR, red and green, see Table 1. For reference the result using only the height information in all three bands is also shown. For more information see [9].

4. DISCUSSION AND CONCLUSION

We have presented methods to support building detection and 3D reconstruction from a low resolution photogrammetric point cloud generated using multi-view stereo reconstruction of standard overlapping aerial images from nationwide data collection. This type of point cloud is very challenging

Table 1. Classification results in terms of accuracy for different input configurations. Average over ten trained networks.

Input layers (RGB)	Ridge roof	Flat roof	Total
Height, red, green	97.48%	90.80%	96.65%
NIR, red, green	97.37%	90.80%	96.55%
Height, height, height	96.35%	81.19%	94.45%

compared to point clouds from laser scanning or from high resolution aerial imagery.

The results are encouraging. We show using annotated data that building roof types can be identified with 96.65% accuracy. For the building detection a precision of 0.89 is obtained for a recall of 89% of all true building pixels. With more training data and a deeper feature extractor it should be possible to increase the performance substantially. In particular, a deeper network would be able to incorporate more high-level contextual information into the analysis, e.g., identifying a large multistory car park as a building, although it locally looks exactly like a ground level car park.

There are several other possible uses for the type of data used here. Instance segmentation (e.g. [10]) is an extension of semantic segmentation where each individual object is detected and delineated. This enables, e.g., to identify different building types in an area and determine their size. Change detection compares an image to one or several previous co-registered images of the same scene, and determines which pixels have changed in some predetermined semantic sense. In this case, one possible approach is to use a network architecture similar to the encoder-decoder design to predict a code vector (or embedding [11]) for each pixel, such that the pixel-wise Euclidian distance between the outputs for a test image and a reference is a meaningful measure of change. Other approaches to change detection are described in [12, 13]. Automatic change detection would be quite useful for the government cadastral agencies who update their maps regularly.

5. REFERENCES

[1] A. P. McClune, J. P. Mills, P. E. Miller, and D. A. Holland, "Automatic 3D building reconstruction from a dense image matching dataset," *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XLI-B3, pp. 641–648, 2016.

[2] M. Rothermel, K. Wenzel, D. Fritsch, and N Haala, "SURE: Photogrammetric surface reconstruction from imagery," in *Proceedings LC3D Workshop*, December 2012.

[3] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 3431–3440.

[4] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, Eds., Cham, 2015, pp. 234–241, Springer International Publishing.

[5] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, Dec 2017.

[6] Pedro O. Pinheiro, Tsung-Yi Lin, Ronan Collobert, and Piotr Dollár, "Learning to refine object segments," in *Computer Vision – ECCV 2016*, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, Eds., Cham, 2016, pp. 75–91, Springer International Publishing.

[7] J. Ba and D Kingma, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations*, 2015.

[8] G Hinton A Krizhevsky, "Learning multiple layers of features from tiny images," *Technical report, University of Toronto*, 2009.

[9] M. Axelsson, U. Söderman, A. Berg, and T. Lithen, "Roof type classification using deep convolutional neural networks on low resolution photogrammetric point clouds from aerial imagery," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 1293–1297.

[10] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask r-cnn," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct 2017, pp. 2980–2988.

[11] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 815–823.

[12] R. Caye Daudt, B. Le Saux, and A. Boulch, "Fully convolutional siamese networks for change detection," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, Oct 2018, pp. 4063–4067.

[13] Ashley Varghese, Jayavardhana Gubbi, Akshaya Ramaswamy, and P. Balamuralidhar, "Changenet: A deep learning architecture for visual change detection," in *Computer Vision – ECCV 2018 Workshops*, Laura Leal-Taixé and Stefan Roth, Eds., Cham, 2019, pp. 129–145, Springer International Publishing.