

SEMANTIC SEGMENTATION AND UNREGISTERED BUILDING DETECTION FROM UAV IMAGES USING A DECONVOLUTIONAL NETWORK

Sangwoo Ham¹, Youngon Oh¹, Kyoungah Choi¹, Impyeong Lee^{1,*}

¹ Dept. of Geoinformatics, University of Seoul, Seoul, Republic of Korea – (f100pilot, ohrgon, shale, iplee)@uos.ac.kr

Commission II, WG II/6

KEY WORDS: Deep learning, UAV image, segmentation, building detection, illegal buildings

ABSTRACT:

Detecting unregistered buildings from aerial images is an important task for urban management such as inspection of illegal buildings in green belt or update of GIS database. Moreover, the data acquisition platform of photogrammetry is evolving from manned aircraft to UAVs (Unmanned Aerial Vehicles). However, it is very costly and time-consuming to detect unregistered buildings from UAV images since the interpretation of aerial images still relies on manual efforts. To overcome this problem, we propose a system which automatically detects unregistered buildings from UAV images based on deep learning methods. Specifically, we train a deconvolutional network with publicly opened geospatial data, semantically segment a given UAV image into a building probability map and compare the building map with existing GIS data. Through this procedure, we could detect unregistered buildings from UAV images automatically and efficiently. We expect that the proposed system can be applied for various urban management tasks such as monitoring illegal buildings or illegal land-use change.

1. INTRODUCTION

As many low-cost commercial UAVs have been released, demands and applications of UAV photogrammetry for urban management is keep increasing. Moreover, many local governments are trying to adopt UAVs for their public administration processes. Among the urban management tasks, local governments have traditionally used airborne aerial-images for detecting illegal buildings in green belt area (Chai *et al.*, 2015), and they started to adopt UAVs for detecting illegal buildings to take advantages of UAV photogrammetry: UAVs can acquire high-resolution aerial images of small areas more frequently and efficiently. However, it is very costly and time-consuming to detect unregistered buildings from UAV images since the interpretation of UAV images still relies on human vision. To overcome this problem, we aim to apply deep learning algorithms to detect unregistered buildings from UAV images. Specifically, we propose a system which detects unregistered buildings from UAV images through the following steps: 1) Train a deconvolutional network with publicly opened national geospatial data, 2) Semantically segment a given UAV image into a building prediction map, 3) Extract unregistered buildings through comparison of existing GIS data and the prediction map.

2. RELATED WORKS

2.1 Deep learning for semantic segmentation

Deep learning methods, especially CNNs (Convolutional Neural Networks), have demonstrated outstanding performance on computer vision due to the availability of large-scale training data and the advancement of GPGPU (General Purpose Computing on Graphics Processing Units) technology (Zeiler *et al.*, 2013). For the sake of the evolution of deep learning, there were many efforts that applied CNNs for semantic segmentation of aerial or satellite images. Saito *et al.* (2015) trained a CNN by the patch-based method to semantically segment aerial images into building maps and road maps. As the patch-based method has some limitations that it yields boundaries on patch borders,

Maggiori *et al.* (2017) adopted FCNs (Fully Convolutional Networks) to overcome the limitation of the patch-based method. However, FCNs also have some limitations that they fail to segment when objects in given images are too large or smaller than the receptive field (Noh *et al.*, 2015). Noh *et al.* (2015) introduced a semantic segmentation algorithm using a deconvolutional network which is composed of unpooling and deconvolution layers, and they could overcome the limitations of FCNs. We decided to use a deconvolutional network for semantic segmentation of UAV images since buildings in aerial images have various sizes and shapes.

2.2 Using publicly available dataset for training

Deep learning algorithms need large-scale and precisely annotated data, but the preparation of high-quality dataset is very difficult. Kaiser *et al.* (2017) trained a deep neural network with different datasets and compared its performance. They revealed that a deep neural network trained with publicly opened data such as OpenStreetMap performed reasonably since the publicly opened data are good in terms of quantity. We decided to use publicly available aerial ortho-images and corresponding digital maps provided by Korean National Geographic Information Institute rather than manually annotate label data.

3. METHODOLOGY

To implement the unregistered building detection system, we built a training dataset and trained the deconvolutional network. Then, we acquired UAV images and detected unregistered buildings by comparing existing GIS data and the building prediction map derived from the deconvolutional network. Our detailed methodology for the implementation of the system is driven by following steps.

3.1 Building a training dataset

3.1.1 Using national geospatial data as a training dataset

We built a training dataset which consists of pairs of aerial ortho-images and digital maps (building layer and road layer). As we mentioned in 2.1, manually annotating a large-scale dataset is extremely difficult and time-consuming meanwhile it is more efficient to use a large-scale publicly available dataset. For that reason, we used national geospatial data which are publicly opened by Korean National Geographic Information Institute (NGII) as shown in figure 1. The national geospatial data, especially aerial ortho-images and corresponding digital maps, can be considered as weakly labeled data for following reasons. First, since the update period of aerial ortho-images and digital maps does not coincide, some demolished buildings which do not appear in an aerial ortho-image still exist in digital maps. We consider that this inconsistency of the update makes the training dataset weakly labeled. Second, the aerial ortho-images provide by NGII are not true ortho-images as shown in figure 2. Some high-rise buildings such as apartments contain relief displacements while the building layers of corresponding digital maps depict bottom of the buildings rather than the top of the buildings. Hence, aerial ortho-images and corresponding digital maps would not match each other in case of high-rise buildings.

3.1.2 Downloading, splitting and augmentation of the dataset

We downloaded the aerial ortho-images and corresponding digital maps from NGII website. As the digital maps are provided in vector data type, we converted the building layers and road layers into raster type and merged them to create label maps. Also, as input layer of a deconvolutional network has a size of $224 \times 224 \times 3$ pixels, we split the aerial ortho-images and corresponding label maps into $224 \times 224 \times 3$ pixel-sized patches as shown in figure 3. Finally, we flipped and rotated the patches for the purpose of data augmentation as shown in figure 4. We built not only the training dataset but also validation and test dataset. The details of each dataset is mentioned in table 1.



Figure 3. Example of $224 \times 224 \times 3$ pixel-sized patches

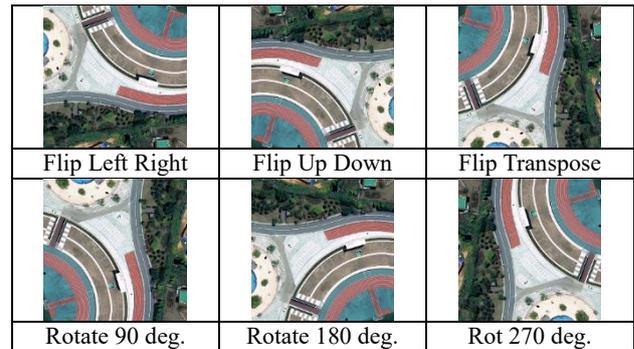


Figure 4. Example of data augmentation

Dataset	# of pixels	Area (m ²)
Training	1,348,816,191 (62.63%)	687,896,257.41
Validation	447,142,381 (20.76%)	228,042,614.31
Test	357,679,196 (16.61%)	182,416,389.96

Table 1. Number of pixels and area of training, validation and test dataset



Figure 1. Example of aerial ortho-images and digital maps



Figure 2. Relief displacements of high-rise buildings in the aerial images

3.2 Architecture of a deconvolutional network

Previous methods using CNNs such as patch-based method and FCNs have some limitations as we mentioned in 2.1. We use a deconvolutional network (Noh *et al.*, 2015) as a semantic segmentation algorithm. Deconvolutional networks can overcome some limitations of FCNs fail to segment substantially larger or smaller objects, which means they are free from scale problems. Also, since deconvolutional networks are composed of unpooling and deconvolution operations, they can naturally densify coarse label maps. Compared to other CNN based methods, we considered that using a deconvolutional network is more appropriate to segment high-resolution UAV images. As Noh *et al.* (2015) implemented a mirrored version of VGG-16 (Simonyan *et al.*, 2014) using unpooling and deconvolution operations, we also constructed the deconvolutional network using VGG-16.

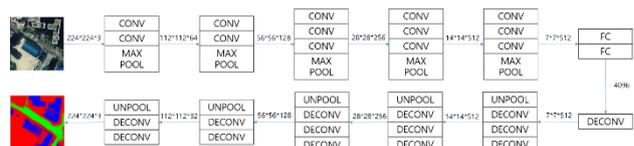


Figure 5. Architecture of the deconvolutional network

To semantically segment a given UAV image, main role of the deconvolutional network is the generation of a prediction map which consists probabilities of each category (P_{ij}) for each pixel at i -th row and j -th column. Let $224 \times 224 \times 3$ be a form of final output \hat{y} of the deconvolutional network. For each pixel of \hat{y} , the

probability of each category ($P_{ijk}, 1 \leq k \leq 3$) is calculated by softmax function.

$$P_{ij} = [P_{building} \quad P_{road} \quad P_{otherwise}]$$

$$P_{ijk} = \frac{e^{\bar{y}_{ijk}}}{\sum_{k=1}^3 e^{\bar{y}_{ijk}}}$$

In order to train the deconvolutional network, we adjusted parameters of the deconvolutional network to reduce a loss function. The loss function of the deconvolutional network is defined using cross-entropy function, and we minimized the loss function using Adam optimizer (Kingma *et al.* 2014).

$$L = - \sum_{i=1}^{224} \sum_{j=1}^{224} \sum_{k=1}^3 \bar{y}_{ijk} \ln P_{ijk}$$

3.3 Detection of unregistered buildings

We semantically segmented given UAV images using the trained model to generate building prediction maps and subtracted existing digital maps from the prediction maps. We considered the remaining pixels after the subtraction as unregistered buildings. However, as the prediction result of the deconvolutional network does not perfectly coincide with actual building shapes, it leaves a small amount of noise during the subtraction operation. We applied opening operation to get rid of the noise. Opening operation is a kind of morphology operation. Morphology operations are widely used for removing noise, isolating individual elements and joining disparate elements from images in terms of computer vision (Kaehler *et al.*, 2016). Dilation and erosion are the basic morphological operations. Dilation is a convolutional operation that replaces a given pixel with the local maximum of the kernel, and erosion yields the local minimum of the kernel. Opening is a combination of the basic morphological operations that erodes an image and then dilates the eroded image. After the comparison between existing GIS data and building prediction map, subtraction operation may yield some small bright noisy pixels. Opening operation removes small bright pixels and maintains the size of remaining regions as shown in figure 7.



Figure 7. Example of opening operation (left: before opening, right: after opening)

4. EXPERIMENTAL RESULTS

4.1 Training the deconvolutional network

We calculated not only training loss but also validation loss using validation dataset to prevent overfitting. The training loss and the validation loss decreases simultaneously at the beginning of the training, but the validation loss starts to increase when the model is overfitted to the training dataset. Through the observation of the validation loss, we trained the model for 15 epochs since the validation loss started to increase after 16th epoch. We spent 2 days and 11 hours for 15 epochs of training using Nvidia GTX 1080 Ti GPU.

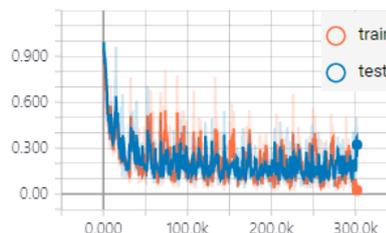


Figure 7. Decrease of the loss function

4.2 Evaluation of the trained deconvolutional network

We evaluated the deconvolutional network with quantitative metrics: recall, precision. Recall means predicted pixels among ground truth pixels, precision means ground truth pixels among predicted pixels. As the final result of the deconvolutional network is basically probability map, we need to threshold the probability map to calculate recall and precision. By varying thresholding values, we could draw recall-precision curve as shown in figure 8. The precision of building and road class at the breakeven point was 71% and 78% respectively.

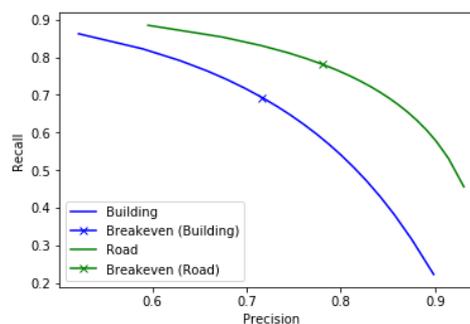


Figure 8. Recall-Precision curve of the trained deconvolutional network

We also evaluated semantic segmentation results of the deconvolutional network qualitatively. In normal cases, the deconvolutional network segmented given aerial images as shown in figure 9. However, the deconvolutional network failed to segment high-rise buildings as shown in figure 10. Since the aerial images of training dataset is not true ortho-images as we mentioned before, we consider that relief displacements made high-rise building weakly annotated. We also found that the deconvolutional network failed to segment ambiguously shaped buildings as shown in figure 11. Urban facilities such as tennis courts and parking lots were considered as 'otherwise' class in the stage of training data preparation. Some buildings which have tennis courts on the top of them appears very ambiguous, and it was one of the reasons which deteriorated the segmentation performance. Although there are some disadvantages with national geospatial data as a training data, it is an efficient way to train a deconvolutional network to segment aerial images.



Figure 9. Good cases

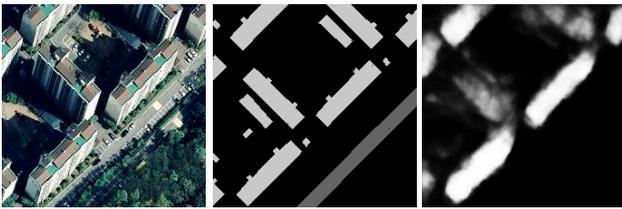


Figure 10. Segmentation result of buildings with relief displacements

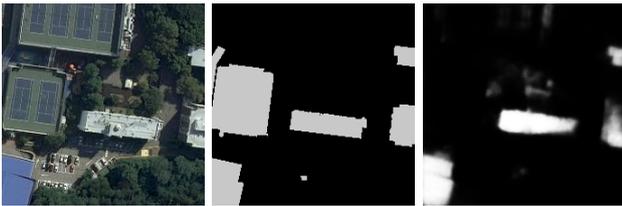


Figure 11. Segmentation result of buildings with ambiguous appearances

4.3 Acquisition of UAV images

We acquire UAV images of Gangsang-myeon, Yangpyeong-gun, Gyeonggi-do (approximately 47km away from Seoul, Korea) as shown in figure 6 due to following reasons: 1) It is suspected that there are some unregistered buildings, 2) The area is included in drone-only airspace. Using a commercial UAV (DJI Mavic Pro), we acquired 154 images of the research area (0.28 km²). The altitude from the ground level was 120m, and ground sampling distance was 5.25cm. The flight time was 17 minutes. After the flight, we built a mosaicked ortho-image of the research area using Photoscan. As our research aims to compare a building prediction map derived from UAV images and existing GIS data, we had to register the UAV images to the existing GIS data.



Figure 6. Location of the area of research

4.4 Unregistered building detection

We prepared existing GIS data of the research area to extract unregistered buildings. Specifically, we assumed that the building layer of national digital map which is updated in 2015 as existing GIS data. Before automatically detecting unregistered buildings, we manually found unregistered buildings by comparing the UAV ortho-image with the digital map of the research area. As it is depicted in figure 12, there were 13 unregistered buildings in the UAV ortho-image. Then, we segmented the UAV ortho-image into building prediction map as shown in figure 13. Finally, we subtracted the existing GIS data from thresholded probability map and applied opening operation to reduce the noise. As a result, we could detect unregistered buildings as shown in figure 14.



Figure 12. Registered buildings (green) and unregistered buildings (purple)



Figure 13. Building probability map



Figure 14. Result of unregistered building detection

The system succeeded to detect 85% percent of unregistered buildings which we digitized manually. The system detected not only unregistered buildings but also extended buildings as shown in figure 15.

Type of detected objects	# of objects
Unregistered Building	11
Extended Building	7
Greenhouse	4
Wrong Detection	8

Table 2. The detection result of our system

Also, greenhouses and wrong objects such as parking lots or farming equipment were detected. In case of greenhouses, there were ambiguous results as shown in figure 16. It was difficult to determine whether the detected greenhouses are 'unregistered buildings' or not since the deconvolutional network detected only several greenhouses. On the other hand, the deconvolutional network detected some of farm equipment and concrete parking lots as buildings and we consider that it made the performance of the system worse.



Figure 15. Example of unregistered buildings and extended buildings that our system detected



Figure 16. Example of ambiguous results in case of greenhouses (left and center) and wrong objects (right)

5. CONCLUSION

In this paper, we proposed the system which automatically detects unauthorized buildings from given UAV ortho-images. We found that the national geospatial data (aerial ortho-photos and digital maps provided by NGII) are appropriate for deep learning training data. Although they have some limitations as a training dataset, it has shown reasonable performance without any additional manual annotation. We also found that a deconvolutional network trained with the national geospatial data can be used for semantic segmentation of UAV images. Through semantic segmentation using the trained deconvolutional network and the simple post-processing procedure (subtraction and opening operation), we could automatically detect unregistered buildings from UAV images. Our system can be applied not only for illegal building construction monitoring but also for illegal building extension monitoring since it was possible to detect extended buildings. We expect that the proposed system can be used for many kinds of urban management tasks such as illegal building monitoring in green belt areas, illegal land use change and update of digital map.

ACKNOWLEDGEMENTS

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education (NRF-2016R1A6A3A11936314).

REFERENCES

- Chai, B., Mun, J., and Lee, M., 2015. Illegal Act and Management System within Greenbelt and Improvement of Policy. *The Korea Spatial Planning Review*, 84, 145–163.
- Kaehler, A., and Bradski, G., 2016. *Learning OpenCV 3: Computer Vision in C++ with the OpenCV Library*. O'Reilly Media.
- Kaiser, P., Wegner, J. D., Lucchi, A., Jaggi, M., Hofmann, T., and Schindler, K., 2017. Learning Aerial Image Segmentation From Online Maps. *IEEE Transactions on Geoscience and Remote Sensing*, 55(11), 6054–6068.
- Kingma, D. P., and Ba, J., 2014. Adam: A Method for Stochastic Optimization. Retrieved from <http://arxiv.org/abs/1412.6980>
- Maggiori, E., Tarabalka, Y., Charpiat, G., and Alliez, P., 2017. Convolutional Neural Networks for Large-Scale Remote-Sensing Image Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(2), 645–657.
- Noh, H., Hong, S., and Han, B., 2015. Learning Deconvolution Network for Semantic Segmentation. In *2015 IEEE*

International Conference on Computer Vision (ICCV) (Vol. 11–18–Dece, pp. 1520–1528). IEEE.

Saito, S., and Aoki, Y., 2015. Building and road detection from large aerial imagery. In E. Y. Lam & K. S. Niel (Eds.) (p. 94050K).

Simonyan, K., and Zisserman, A., 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. Retrieved from <http://arxiv.org/abs/1409.1556>

Zeiler, M. D., and Fergus, R., 2013. Visualizing and Understanding Convolutional Networks. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8689 LNCS(PART 1), 818–833.