Treetop Detection using Convolution Neural Networks Trained through Automatically Generated Pseudo Labels

Changlin Xiao[a], Rongjun Qin [a,b,]*, and Xu Huang[a]

[a] *Department of Civil, Environmental and Geodetic Engineering, The Ohio State University, Columbus, OH 43210, USA.*

[b] *Department of Electrical and Computer Engineering, The Ohio State University, Columbus, OH 43210, USA. (qin.324@osu.edu).*

Abstract: Using remote sensing techniques to detect trees at the individual level is crucial for forest management while finding the treetop is an initial and important first step. However, due to the large variations of tree size and shapes, traditional unsupervised treetop detectors need to be carefully designed with heuristics knowledge making an efficient and versatile treetop detection still challenging. Currently, the deep convolutional neural networks (CNNs) have shown powerful capabilities to classify and segment images, but the required volume of labelled data for the training impedes their applications. Considering the strengths and limitations of the unsupervised and deep learning methods, we propose a framework using the automatically generated pseudo labels from unsupervised treetop detectors to trains the CNNs, which saves the manual labelling efforts. In this study, we use multi-view satellite imagery derived digital surface model (DSM) and multispectral orthophoto as research data and train the fully convolutional networks (FCN) with pseudo labels separately generated from two unsupervised treetop detectors: top-hat by reconstruction (THR) operation and local maxima filter with a fixed window (FFW). The experiments show the FCN detectors trained by pseudo labels, have much better detection accuracies than the unsupervised detectors (6.5% for THR and 11.1% for FFW), especially in the densely forested area (more than 20% of improvement). In addition, our comparative experiments when using manually labelled samples show the proposed treetop detection framework has the potential to significantly reduce the need for training samples while keeping a comparable performance.

Keywords: Treetop Detection, Neural Network, Pseudo Label, Deep Learning

1. Introduction

Forest is one of the most important ecological components and plays an important role in the global ecosystem. Detailed tree-level attributes such as tree counts, tree heights, and canopy sizes are essential for monitoring forest regeneration, quantitative analysis of forest structures and dynamics (Mohan et al., 2017; Weng et al., 2015; Zhao et al., 2014). Treetop detection often as a first step of the process, has drawn substantial attention and many methods have been proposed to extract treetops from different data sources (Franceschi et al., 2018; Hosoi et al., 2012; Khosravipour et al., 2015). The identified treetops can be directly used for forest inventory assessment (Pearse et al., 2018; Pont et al., 2015), and offering height information for category, growth, volume estimation and crown-level segmentation (Hill et al., 2017; Kathuria et al., 2016; Latifi et al., 2015; Wang et al., 2016; Xiao et al., 2019).

With the spectral images, an often-used assumption is that trees reflect the light shedding on them in a decreasing manner from top to bottom (Culvenor, 2002; Özcan et al., 2017). Hence, treetops can be detected as brightest spots in the satellite or aerial images, and based on this, window-based local maxima filters are proposed to find the brightest points as treetops (Pouliot and King, 2005; Wulder et al., 2000). On the other hand, 3D point cloud data is widely available and considered as important sources for individual treetop detection (Ferraz et al., 2016; Saarinen et al., 2017; Str îmbu and Str îmbu, 2015). Most of the 3D based methods use the canopy height model (CHM), which naturally extract the treetops as local maxima to identify the trees at the individual level. Hence, for either optical image-based or 3D point-based tree detection methods, local maxima detection is one of the most popular and effective means to find potential treetops. However, the optimal window size of the local maxima detector may vary with the regions (Monnet et al., 2010). To address, a few methods proposed filters that can adaptively adjust their window sizes by considering the slope change or using the allometric equation that describes the relationship between tree crown size and height (Özcan et al., 2017; Song et al., 2010; Wulder et al., 2000). However, the slope is sensitive to the CHM errors and surface relief, while the use of the allometric equation requires knowledge of tree species which is often unknown (Liu et al., 2015). Another class of methods use template matching for treetop detection (Quackenbush et al., 2000; Tarp-Johansen, 2002). Representative and complete templates normally lead to good matching results, but such methods may need a large number of training samples and their transferability to different datasets is usually low (Mallinis et al., 2008). The morphological operation based regional maxima (Khosravipour et al., 2015) applies the morphologically reconstructed image to subtract the original image thus highlight the peak areas as top-hat (Qin and Fang, 2014; Vincent, 1993). As described in (Xiao et al., 2018; Xiao et al., 2019), the top-hat by reconstruction (THR) morphological operation on DSM (digital surface model) can efficiently detect treetops while being robust to the crown size of trees. However, this method requires multispectral information as an input and may produce repeated detections.

Manually crafted treetop detectors are often sensitive to data noise, while it is known that deep convolutional neural networks (CNN) trained through a large number of samples appear to be robust, and has been reported in various benchmark tasks (e.g. image classification) to have outstanding performances. Different from the low-level hand-crafted features, the convolutional neural networks can learn high-level semantic information from the training samples. In the remote sensing community, many CNNs have been applied to scene classification (Wang et al., 2018a), change detection (Wang et al., 2018b), and the patch-level (Mubin et al., 2019) and pixel-level (Paoletti et al., 2018; Sherrah, 2016; Sun et al., 2017) image segmentation and object detection. Considering that trees have distinct while rather complex geometrical and spectral characteristics, the attempt of using CNN as a more capable model for individual tree detection studies have been reported (Csillik et al., 2018; Freudenberg et al., 2019; Li et al., 2016; Mubin et al., 2019). However, these works demonstrated impressive performance in regions where well-annotated data is available, while equivalent applications in wild environments where such annotated data is not available are not yet well studied.

In this work, we hypothesize that CNN is a complex and robust tool as shown by many existing works (Jindal et al., 2016; Rolnick et al., 2017; Veit et al., 2017), tolerate inaccurate samples generated from traditional unsupervised treetop detectors (i.e. pseudo labels), thus to be operated in a fully unsupervised fashion. Therefore, we test this hypothesis by developing a processing framework that takes such pseudo labels for CNN training and thus treetop prediction (as shown in Fig. 1). As mentioned, the use of CNN for treetop detection has several advantages: 1) with a large amount of training data under different scenarios, the network can learn the high-level semantic features and are applicable to various scenarios without parameter tuning; 2) without hand-crafted feature descriptors, the network can automatically learn and extract the complementary 2D spectral and 3D structural information for more accurate detections.

The proposed framework is applied to the multi-view high-resolution satellite imagery derived DSM and the orthophoto containing multispectral data. To utilize an existing CNN structure, we turn such a dataset into three bands (referred as multi-cue data hereafter) that contains respectively red channel of the multispectral data, normalized difference vegetation index (NDVI) and the DSM. As shown in Fig. 1, the pseudo labels are firstly generated by the unsupervised detector. The multi-clue data is used to train a Fully Convolutional Neural Network (FCN) and predict the treetops pixels as the foreground. The contribution of this paper is two-fold: 1) we propose an efficient and effective unsupervised framework to train the CNNs for treetop detection; 2) we demonstrated that an FCN trained through these pseudo labels offers comparatively better detection accuracy than the pseudo labels themselves.

2. Study Area and Data Processing

The study area is located in Don Torcuato, a small city on the west side of Buenos Aires, Argentina. As illustrated in Fig. 2, an urban area and a densely forested area are selected as the test sites. The dimension of the urban area which contains both forest and urban environments is 6.740 km by 6.914 km (22469 pixels × 23048 pixels). The densely forested area (0.934 km × 0.928 km) is selected to test and verify the treetop detection in the dense forest which is known as a challenging area for the task. The satellite images in this work are from the multi-view benchmark dataset provided by John's Hopkins University Applied Physics Lab (Bosch et al., 2016; Bosch et al., 2017), containing multiple worldview 2/3 images over this area across two years. Worldview 2/3 has a multispectral sensor which includes 8 bands (coastal, blue, green, yellow, red, red edge, near-IR1, near-IR2). Hence, in addition to regular RGB colour, some spectral information, for example, normalized difference vegetation index (NDVI) can be calculated and utilized to find the vegetation area.

The satellite images were taken under various conditions containing on-track and off-track stereos with the ground resolution around 0.3-meter. To derive an accurate DSM, we selected five pairs of the on-track stereo images captured in December 2015, with the maximal off-nadir angle between 7-19 degrees and the average intersection angle between 15-21 degrees. A fully automated pipeline proposed by (Qin, 2017) that consists of 1) pansharpening, 2) automatic feature matching, 3) pair-wise bundle adjustment, 4) dense matching and 5) a bilateral-filter based depth-fusion, is applied to generate the high-quality DSM and subsequently true orthophoto. Comparing to the ground truth LiDAR data, the root mean square errors (RMSE) of the DSM on this benchmark dataset varying between 2.5-4 meters. The RMSE is absolute accuracy at checking points which does not represent the relative accuracy of the object reconstruction and we believe the 0.3-meter resolution is enough for the regular treetop detection. The core method is a hierarchical semi-global matching and the readers may find more details about the method on this data in (Qin, 2017). The generated orthophoto and DSM of the urban area are shown in Fig.3 while giving an example of the detailed trees in the images and their 3D shapes.

3. Methodology

The proposed framework which uses the detections of unsupervised detectors to train a CNN for treetops detection mainly contains three parts: the unsupervised treetop detector, the generation of pseudo labels from unsupervised treetop detector, and the CNN and the training for the treetop detection. In this section, these three parts are sequentially presented and explained.

3.1 *Unsupervised Treetop Detectors*

Many tree detection methods, along with treetop detection, have been proposed with different data sources (Skurikhin et al., 2013; Wulder et al., 2000; Xiao et al., 2019). Among these detectors, the top-hat by reconstruction (THR) operation is adapted to detect treetop for the multi-clue data, because it is more efficient and less sensitive to the tree crown size. As a comparison, the regular local maxima filter with the fixed window (FFW) is used as another treetop detector to generate the pseudo labels.

3.1.1 *Top-Hat by Reconstruction (THR)*

Since most of the trees have distinguishable treetops, we naturally assume the local maxima on the DSM are the treetops. Compared to many other local maxima detectors, the grey-level morphological top-hat by reconstruction operation is an effective method for blob-like shapes detecting and less sensitive to filter size (Qin and Fang, 2014; Vincent, 1993). Morphological top-hat is defined as the peaks of an image grid computed by morphological operations, and several state-of-the-art tree detection methods have used this morphological operation to successfully isolate the treetops (Khosravipour et al., 2015; Xiao et al., 2019).

Following the THR detection work in (Xiao et al., 2019), a disk-shaped structuring element ($\mathbf{e}$) is used to do the grey-level morphology erosion on the DSM to generate a marker image $\boldsymbol{\varepsilon}(\mathrm{DSM}, \mathbf{e})$. Then, the morphological reconstruction mask $\mathbf{B}_{\boldsymbol{\varepsilon}(\mathrm{DSM},\mathbf{e})}$ is generated from the maker image with an iterative process. Finally, by subtracting the morphological reconstruction mask $\mathbf{B}_{\boldsymbol{\varepsilon}(\mathrm{DSM},\mathbf{e})}$ from the DSM, the peaks on DSM can be extracted. From the peaks on the DSM, several post-processing methods are used to further detect the treetops. Firstly, the normalized difference vegetation index (NDVI) is adopted to remove the local maxima in the non-vegetation area. Then, low trees are filtered by checking above-ground height which can be calculated by subtracting the heights of nearby terrain areas on the DSM. Finally, a non-maximum suppression is used to refine the treetops which are too close to each other. An example of the treetop detection is shown in Fig. 4 where the image (b) is the DSM of the test area while the NDVI mask (c) is used to remove the local maxima in the non-vegetated area. The final treetops are shown as blue dots in rectangles (Fig. 4, image (a)), while the red dots are the local maxima that filtered out by the height check and the green stars without blue dots are the ones that filtered out by the non-maximum suppression. More processing details can be found in (Xiao et al., 2019).

3.1.2 *Local Maxima Filter with a Fixed Window (FFW)*

As a comparison of using different unsupervised treetop detector, a local maxima treetop filter with a fixed window is also implemented on the DSM as described in (Wulder et al., 2000). After repetitive tests, the window size with 7 pixels which corresponding to 2.1 meters is selected for the filter because it has the best average performance in the test scenarios.

Similar to the THR detector, NDVI and height information are used to remove the noise in the final treetop detections.

3.2 *Generation of Training Samples*

To be compatible with the FCN pre-trained weights, we convert the multi-band orthophoto and DSM into a three-channel fused multi-clue image which includes the red band, NDVI and DSM values that normalized into 0-1. In the urban area, we randomly selected seven patches with 1000 × 1000 pixels (300 × 300 meters^2) as test sites. Tree detection can be particularly challenging in the densely forested area, however very important. Therefore, a selected site with such a dataset is incorporated in the experiment (shown in Fig. 5, image (b)). In each test site, 3000 sub-patches are randomly selected as training samples. The size of the training patch is designed as 48 × 48 pixels corresponding to 14.4 × 14.4 meters^2 which is normally large enough to cover a tree. The training masks (0 for the background area, 1 for treetop area) are generated by finding treetops through the local maxima detectors as introduced in sections 3.1. To allow redundancies and be robust to noise, we make a 3 × 3 window around the treetop as the ground truth, and Fig. 5 gives an illustration of the generation of training samples.

As mentioned, the treetops that detected by unsupervised local maxima detectors may have many errors and this makes the training samples noisy. Conceptually problematic, actually the training with noisy datasets for CNNs has been widely studied (Jindal et al., 2016; Rolnick et al., 2017; Veit et al., 2017) and the possibility of robust training with noisy labels was shown. For example, in (Rolnick et al., 2017), they claimed the deep neural networks are robust to massively noisy labels (accuracy is less than 50%) and demonstrated on several popular public datasets with different types of noise. In our case, the worst detection accuracy of the labels that from unsupervised local maxima detector is around 60%, thus we believe a network trained with such a dataset could be used to generalize a treetop detector. Additionally, as demonstrated in (Li et al., 2019), the early stopping is provably robust to label noise even with a large fraction of corrupted labels. Hence, in this study, we aim to use these non-manually labelled training samples with early stopping to train a treetop detection network and test if it is better than the unsupervised detectors used to offer the training data.

3.3 *Treetop Detection Network*

Many networks have been studied for tree detection and segmentation, among which the fully convolutional network (FCN) (Long et al., 2015) is one of the most basic networks. FCN use the fully convolutional layer instead of the fully connected layer thus can efficiently perform the pixel-level semantic segmentation at arbitrary input size which is practical for remote sensing data. In this study, we adopt a cropped residual FCN for the treetop detection as illustrated in Fig 6. Instead of using five blocks of the convolution and max-pooling layers, we only use three blocks because the input size is smaller (48 × 48) and the task is simpler (binary). After the three max-pooling layers, the size of the feature maps would be reduced to 6 × 6, and the two following fully convolutional layers will produce a prediction for the two classes (treetop and background) at a down sampled resolution (6 × 6). The last layer is an up-sampling layer which is used to resize the output image as big as the input image. Since there are only three max-pooling layers, the up-sampling size will be eight times which makes this network an adapted version of FCN-8s that without layer-connections. More details about the architecture of the residual FCN can be found in (Long et al., 2015) and Res-Net in (He et al., 2016). There are other sophisticated networks can be used for the treetop detection, such as the U-Net (Ronneberger et al., 2015) and the Mask-RCNN (He et al. 2017) which we will compare and discuss in the experiment with more details. The output of the treetop FCN is a two-channel probability distribution map. The values in the first and second channel represent the possibilities of being treetops and non-treetops, separately. Since the detected treetops are usually represented as segments, we need to further locate them at the pixel level by finding the highest point in each region.

To train the network, 24000 training samples are generated as described in section 3.2, and 80% of them are used to train the network while the other 20% is used as validation data to avoid overfitting. Since the reference data for independent testing are manually labelled and never used in the training, all the study sites are used as test data. The training epoch is set as 200 but with an early stopping scheme, and the other training parameters are 256 for the batch size and 0.0001 for the Adam optimizer with cross-entropy loss as the target function. Since the treetop only takes a small part of the image, we weight the loss as [1, 10] for the non-treetops area and the treetop area separately in the cross-entropy function.

4. Experiment

4.1 *Accuracy Assessment Metrics*

To quantitatively validate the individual treetop detection accuracy, the true positives (TP), false positives (FP) and false negatives (FN) are used to compute the correct detections, wrong detections, and the missing detections, respectively. Based on their numbers, we can calculate the detection accuracy (*A*) or recall (*r*), commission error (e_{com}) and the omission error (e_{om}):

$$A = r = \frac{n_{\mathrm{TP}}}{N}, \tag{1}$$

$$e_{\mathrm{com}} = \frac{n_{\mathrm{TP}}}{n_{\mathrm{TP}}+n_{\mathrm{FP}}}, \tag{2}$$

$$e_{\mathrm{om}} = \frac{n_{\mathrm{FN}}}{n_{\mathrm{TP}}+n_{\mathrm{FN}}}, \tag{3}$$

where $n_{\mathrm{TP}}, n_{\mathrm{FN}}$ and n_{FP} are the number of treetops in TP, FN, and FP, while *N* is the total number of the reference treetops. Other metrics like precision (*p*) and F-score (*F*) can be derived as:

$$p = \frac{n_{\mathrm{TP}}}{n_{\mathrm{TP}}+n_{\mathrm{FP}}}, \tag{4}$$

$$F = \frac{2rp}{r+p}. \tag{5}$$

In the experiments, if the detected treetop is in the reference mask, it is a correct detection, otherwise a false detection. If there are multiple detections in the same reference treetop, only one detection will be identified as correct while all the others are false positives.

4.2 *Reference Data*

In this study, we are limited to collect the field samples. To get the reference data, we labelled the individual treetops by visual inspection through 3D visualization of the orthophoto and DSM with the best human recognition efforts. In the reference data, the treetop is labelled as a small area which contains the treetop and around crowns. The size of the treetop mask varies with the trees since in some cases it is hard to find a precise treetop and Fig. 7 gives two examples of the reference data.

4.3 *Experiments and Discussions*

With the experimental datasets, we performed several different experiments. 1) Analysing the performances of training FCN with pseudo labels generated from THR and FFW, separately, as well as a comparison of using manual labels. 2) Refining the training with extra samples that are identified as incorrect detections in the FCN prediction results. 3) Examining the scalability and generality of the proposed FCN detector in the whole urban area. 4) Experimenting with the detection ability of only use RGB and height information with the FCN. 5) The analysis of using different patch sizes for FCN training. 6) The comparison of using different CNNs.

4.3.1 *Training with Pseudo Labels*

To analyse the performance of the FCN trained from pseudo labels, we separately generated two sets of training samples that came from the unsupervised local maxima filter with a fixed window (FFW) and the top-hat by reconstruction operation (THR). In the experiment, we empirically set the window size of FFW as 7 pixels, correspondingly 2.1 meters of ground sampling distance; and set the structure element size of THR as 5 pixels. As described in section 3.2, the results of the two local maxima detectors are used to generate the training samples for their FCNs. Additionally, for comparison, the manual reference labels are also used as training data.

The comparisons between unsupervised local maxima detectors and FCN-based detectors are carried out through all these test sites. For each site, the results of detection accuracy (A), commission error (e_{com}), the omission error (e_{om}), and the F-score (F), are calculated and shown in Table 1 and Table 2. As we can observe, the FCN detectors have much better A than the unsupervised detectors (11.1% for FFW and 6.5% for THR) which indicates the FCN detectors have more robust performances at various scenarios compared to the elaborately designed unsupervised local maxima detectors. This can be arguable because FCN essentially learns a continuous function and the training dataset are sufficiently large to ensure a good functional estimation. Thus, as a result, the FCN tends to predict treetops with wider spectrums (higher completeness) and potentially less omission error (e_{om}) than the originally used training data from unsupervised detectors. On the other hand, both the two FCNs have a higher commission error (e_{com}), which may lower the detection precision and increase the F-score. This error is mainly related to the redundant detections in one tree, but with a non-maximal suppression, most of the incorrect detections can be filtered.

Comparing the two unsupervised local maxima detectors, the THR has a much better (10.5%)

average detection accuracy (A) than the FFW. However, their FCN detectors do not have such large performance gaps (the A of THR-based FCN is 6% better than the FFW-based FCN) which means the FCN has a strong learning ability that can capture similar valuable features from the relatively poor training samples. Of course, better training samples still result in better performance as we can observe that the THR trained FCN has much lower omission error (e_{om}) than FFW trained FCN (23.3% vs. 29.3%).

However, the performances of the FCN detectors at two places are not as good as the unsupervised detectors. As compared to the THR detector, the THR-based FCN detector performs better in sites 1, 2, 4, 5, 6 and 8, but have slightly worse (3.24%-5.24% lower) detection accuracy (A) in site 3 and 7. As illustrated in Fig. 8, the two test sites contain many sparsely located trees which can be easily identified by both THR and FCN detectors. However, several low trees are ignored by the FCN detector. The missing of low trees is mainly due to the lack of training samples in the sparsely forested area, and the false-negative samples of wrongly labelled low trees from the urban areas. As we can find out from Table 1 and Table 2, in the urban areas where contain many buildings, the detection accuracy (A) of the local maxima detectors is lower which may result in many incorrect samples (near one-fourth to half of the detections). Particularly, the low trees close to the man-made objects may be missed by the unsupervised detectors and act as negative samples, thus subsequently affected the performance of the FCN. Nevertheless, the FCN detector has a great (21.7%) improvement in the densely forested area (site 8). The 3D surface in the densely forested area is relatively smoother as shown in Fig. 9, such a 3D surface can present clearer geometric shapes of the treetops that can be utilized by FCN to detect more treetops.

Since most deep learning based tree detection methods are using manual labels (Freudenberg et al., 2019; Mubin et al., 2019), to compare the performance of using automatically generated training samples and manual labels, we train another FCN with the reference data. In the experiment, the treetop reference data is processed as manual labels to generate the training samples as described in Section 3.2. To validate their performance, 20% of the training samples are randomly selected as validation data that have not been included in the training. After the same training processing, the final detection results can be found in Table 3.

The quality of the training samples is critical for deep learning networks. As we can observe from Table 3, well and truly, the elaborately labelled training samples have better performance than the pseudo labels. However, without laborious manual works, the automatically generated pseudo labels still gain close scores in A and e_{com}, which also indicate the great potential of the proposed treetop detection framework in practical applications, such as for the natural forest, where the manual labels are hard to collect.

4.3.2 *Re-learning from Prediction*

Since the treetops have characteristic spectral and geometric properties, e.g., they must be in the vegetated areas (the red rectangle in Fig. 8, image (b) shows some obvious wrong detections), the treetop detection results can be further refined by certain constraints. According to this observation, we can re-train the network by taking these incorrectly detected results as negative examples to get better performances.

Based on this idea, we predict the treetops of all patches and find the incorrect ones by two constraints: 1) the treetop should be in the vegetated area (through the NDVI); 2) the treetop should be a local maximum in the non-maximum suppression window as described in section 3.1.1. Hence, with these constraints, the incorrectly detected treetops can be identified and corrected from the initial detections. Then, for each incorrectly detected treetop, four training samples are randomly generated around it and incorporated into the training data as negative training samples for a re-learning. In this experiment, we only use the THR treetops as training samples and perform the re-learning for two successive times (FCN_r1 and FCN_r2). In the first FCN predictions, we found 709 incorrect treetops which have been reduced by 9.7% in the re-learned predilections (640 incorrect treetops). The other metrics are shown in Table. 4, at where the FCN_r1 represents the first re-learning, while the FCN_r2 is the second re-learning.

As we can observe from Table 4, with the first re-learning, all measurements are improved, especially the detection accuracy and omission error (both around 1%). These improvements demonstrate that the FCN can be further improved with better training samples which can be refined by the prior-knowledge. However, after the second re-learning, the detection performance decreases slightly. By analysing the newly added training samples, we found many of them are from the areas that contain larger detection errors, which means many of the newly added treetop labels are incorrect. Hence, if the majority of the newly added training samples are from the areas with poor detections, the quality of the training samples will be degraded and subsequently hinder the refinement.

4.3.3 *Scalability at Large Area*

Since we use the fully convolutional architecture, it is possible to apply it to large format images. To test the scalability of the trained FCN, we apply the detection in the entire urban area. In this experiment, the FCN detector is trained with the THR pseudo labels from only a small part (1.35%) of the whole urban area.

Through the visual inspection of Fig. 10 which shows the detection results of the whole urban area, we find the most trees are successfully identified and most of them are correctly located. Even the FCN detector is trained with very limited automatically generated noisy labels, the detections are still quite satisfying without any post-processing. Certain detection errors might still exist, such as cars on the bridge identified as treetops and several omitted trees. These errors are mainly caused by the lack of representative training samples on the relatively rare and small objects in the urban scenes, for example, the traffic bridges. It is expected that with more correct training samples, the performance of FCN detector can be further improved.

4.3.4 *Treetop Detection with RGB and Height Information*

Besides the satellite data, unmanned aerial vehicle (UAV) and aerial images are becoming more popular in the remote sensing community. With these UAV or aerial images, the orthophoto and DSM can be generated with the photogrammetric techniques as mentioned in section 2. However, compared to the multi-spectral satellite images, most of these UAV or aerial images only contain the regular RGB information. Hence, to analyse the possibility of using this data for treetop detection, we carried out another experiment in which only the regular bands and height (RGB-H) are fed to train the FCN detector.

Without multi-spectral information, the RGB-H four-channel training dataset is generated similarly as described in section 3.2. Accordingly, the input size of the network is changed to four channels, and without pre-trained weights, the network is trained from scratch. After the same training procedures, the detection performance of the FCN detectors based on the THR (FCN-THR) and FFW labels (FCN-FFW) is shown in Table 5.

It can be seen that the performance of the FCN treetop detectors with only RGB-H features are not comparable to the one of using multispectral information (the average *A* has been decreased around 9.36% (THR) and 9.38% (FFW). These differences demonstrate that the multispectral information is important for tree detection especially in the densely forested area (site 8) where the performance degraded the most. Also, since this network is trained from scratch, the performance of the training is usually not as good as using pre-trained weights. Nevertheless, compared to the unsupervised detectors using the multispectral information, the FCN detections without such information show similar performances which indicate the reliability of using FCN for treetop detection with various data sources.

4.3.5 *Comparison of Using Different Sample Sizes*

Generally, a larger image patch that contains more global information is better for target detection in the networks, like what is demonstrated in the pyramid scene parsing network (Zhao et al., 2017). However, since our training samples are completely based on a set of automatically generated treetops, a larger patch size implies the need of more correctly detected treetop in the patch, which can be uncertain since we have observed the pseudo labels often have large commission and omission errors. To analyse the impact of using different training sizes, we carry out an experiment comparing the performance of using different sizes including 48 × 48, 64 × 64, 96 × 96 and 128 × 128. In this experiment, we use an identical FCN with THR labels for all patch sizes and keep the training parameters the same, and the final results can be found in the following Table 6.

As we can observe from Table 6, the smaller size (48 × 48) assumed to cover one tree has the best *A* and smaller e_{om}, while the larger sample sizes give better e_{com} and *F*. As we mentioned above, a larger patch with potentially more missed treetops could lead to a lower *A* and larger e_{om}. On the other hand, the correctly labelled non-treetops in a larger training patch also can reduce the negative impact of wrongly identified treetops in the pseudo labels. These results indicate that with the uncertainty of the pseudo labels, a smaller training size covering necessary context would be better for the treetop detection.

4.3.6 *Comparison of Using Different Networks*

Besides the FCN, there are several semantic segmentation and object detection networks, like

U-Net (Ronneberger et al., 2015) and Mask-RCNN (He et al., 2017), could be used for the tree detection, for example, the U-Net is adopted for the palm trees detection at a large scale (Freudenberg et al., 2019). Based on the structure of the FCN, the U-Net improves the segmentation performance by involving more well-reasoned multi-layer connections and up-sampling layers. On the other hand, compared to other object detection networks, such as Faster-RCNN (Girshick, 2015), SSD (Liu et al., 2016), and YOLO (Redmon et al., 2016), the Mask-RCNN is a state-of-the-art object detection network which can simultaneously predict the bounding boxes and the masks of the objects. Hence, in this experiment, we use Mask-RCNN as a representation of the object detection networks. To analyse the abilities of different networks for treetop detection, we train the U-Net and the Mask-RCNN with the THR labels as same as the training of FCN. In the experiment, the patch size is selected as 64 × 64 to satisfy the minimal input size for the networks. The training parameters are tuned to reach their optimal performances, shown in Table 7.

It can be seen that the Mask-RCNN has the worst performance, and we found only the treetops on the trees with obvious boundaries can be identified. The Mask-RCNN is designed for object detection, and the target is usually a completed object with a clear boundary. In this case, the treetop might be a part of the tree without an easily distinguishable boundary leading to failures of object detection networks. On the other hand, the two segmentation networks (FCN and U-Net) have similar excellent performance, while the U-Net has much higher commission error (e_{com}) than the FCN (49.6% vs. 36.9%). We observed the e_{com} of the U-Net and the THR training sample are very close (49.6% vs. 48.2%), which potentially indicates that the powerful learning ability of U-Net becomes a disadvantage when the training samples have non-negligible noise. The wrongly detected treetops in the pseudo labels have been indistinguishably learned by the U-Net, leading to the similar large e_{com} in the test. This results also show that, unlike using manual labels, in the case of noise presenting in the training data, a more complex model might not necessarily outperform simpler models.

5. Conclusion

Treetop detection is critical in a wide range of forest and environmental applications. It can directly offer the count of individual trees at a large scale and greatly facilitates applications such as tree crown delineation and segmentation. In this study, we use multi-view high-resolution satellite imagery derived DSM and orthophoto as the primary data source to analyse the possibility of combining the unsupervised treetop detectors with deep convolutional neural networks for treetop detection. Considering the training of the networks needs a large number of labelled samples, instead of manual labels, we propose to generate pseudo labels using unsupervised local maxima treetop detectors. The experiments show that the convolutional neural network can learn high-level semantic features from the noisy pseudo labels and thus yield better detections than unsupervised detector.

More specifically, we adopt the residual FCN as a pixel-level classification network to segment the input image into treetops and non-treetops. The detection results of the THR (top-hat by reconstruction) and FFW (filter with fixed window) are used as the pseudo labels to train the FCN. Through the experiments, we found that compared to the unsupervised detectors, the average detection accuracies (A) have been increased by 6.5% (THR) and 11.1% (FFW) by the FCN detectors, especially for the traditionally challenging densely forested area (around 20% improvement). The FCN detectors are more robust in different scenarios and through the re-learning, the performance can be further improved in detection accuracy and commission error (both around 1%). Additionally, we demonstrated the promising performance of the proposed framework with regular RGB and height information for treetop detection, as well as the FCN's scalability at a larger urban area. In the end, the differences between different CNNs and the training samples have been compared and discussed.

There are still certain errors in the FCN detection results, such as the clustered treetops (redundant detections in one tree), and some misclassified man-made objects. More sophisticated network training mechanism may improve the detections, but we believe the major obstacle is the quality of the training data. Hence, in the future, we will consider how to improve the quality and the representativeness of the training samples generated from unsupervised detectors.

Acknowledgements

The authors would like to thank John Hopkins University Applied Physics Lab and the DigitalGlobe for providing the multi-view 3D benchmark dataset that created for the IAPRA multi-view stereo 3D mapping challenge (SpaceNet on Amazon Web Services (AWS). ‘Datasets.’ The SpaceNet Catalog. Last modified April 30, 2018. Accessed [2016]. https://spacenetchallenge.github.io/datasets/datasetHomePage.html).

References

Bosch, M., Kurtz, Z., Hagstrom, S., Brown, M., 2016. A multiple view stereo benchmark for satellite imagery*, Applied Imagery Pattern Recognition Workshop (AIPR), 2016 IEEE*. IEEE, pp. 1-9.

Bosch, M., Leichtman, A., Chilcott, D., Goldberg, H., Brown, M., 2017. Metric evaluation pipeline for 3d modeling of urban scenes. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences* 42*,* 239.

Csillik, O., Cherbini, J., Johnson, R., Lyons, A., Kelly, M., 2018. Identification of citrus trees from unmanned aerial vehicle imagery using convolutional neural networks. *Drones* 2*,* 39.

Culvenor, D.S., 2002. Tida: An algorithm for the delineation of tree crowns in high spatial resolution remotely sensed imagery. *Computers & Geosciences* 28*,* 33-44.

Ferraz, A., Saatchi, S., Mallet, C., Meyer, V., 2016. Lidar detection of individual tree size in tropical forests. *Remote sensing of environment* 183*,* 318-333.

Franceschi, S., Antonello, A., Floreancig, V., Gianelle, D., Comiti, F., Tonon, G., 2018. Identifying treetops from aerial laser scanning data with particle swarming optimization. *European Journal of Remote Sensing* 51*,* 945-964.

Freudenberg, M., Nölke, N., Agostini, A., Urban, K., Wörgötter, F., Kleinn, C., 2019. Large scale palm tree detection in high resolution satellite images using u-net. *Remote Sensing* 11*,* 312.

Girshick, R., 2015. Fast r-cnn*, Proceedings of the IEEE international conference on computer vision*, pp. 1440-1448.

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask r-cnn*, Proceedings of the IEEE international conference on computer vision*, pp. 2961-2969.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition*, Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770-778.

Hill, S., Latifi, H., Heurich, M., Müller, J., 2017. Individual-tree-and stand-based development following natural disturbance in a heterogeneously structured forest: A lidar-based approach. *Ecological Informatics* 38*,* 12-25.

Hosoi, F., Matsugami, H., Watanuki, K., Shimizu, Y., Omasa, K., 2012. Accurate detection of tree apexes in coniferous canopies from airborne scanning light detection and ranging images based on crown-extraction filtering. *Journal of Applied Remote Sensing* 6*,* 063502.

Jindal, I., Nokleby, M., Chen, X., 2016. Learning deep networks from noisy labels with dropout regularization*, Data Mining (ICDM), 2016 IEEE 16th International Conference on*. IEEE, pp. 967-972.

Kathuria, A., Turner, R., Stone, C., Duque-Lazo, J., West, R., 2016. Development of an automated individual tree detection model using point cloud lidar data for accurate tree counts in a pinus radiata plantation. *Australian Forestry* 79*,* 126-136.

Khosravipour, A., Skidmore, A.K., Wang, T., Isenburg, M., Khoshelham, K., 2015. Effect of slope on treetop detection using a lidar canopy height model. *ISPRS journal of photogrammetry and remote sensing* 104*,* 44-52.

Latifi, H., Fassnacht, F.E., Müller, J., Tharani, A., Dech, S., Heurich, M., 2015. Forest inventories by lidar data: A comparison of single tree segmentation and metric-based methods for inventories of a heterogeneous temperate forest. *International Journal of Applied Earth Observation and Geoinformation* 42*,* 162-174.

Li, M., Soltanolkotabi, M., Oymak, S., 2019. Gradient descent with early stopping is provably robust to label noise for overparameterized neural networks. *arXiv preprint arXiv:1903.11680*.

Li, W., Fu, H., Yu, L., Cracknell, A., 2016. Deep learning based oil palm tree detection and counting for high-resolution remote sensing images. *Remote Sensing* 9*,* 22.

Liu, T., Im, J., Quackenbush, L.J., 2015. A novel transferable individual tree crown delineation model based on fishing net dragging and boundary classification. *ISPRS Journal of Photogrammetry and Remote Sensing* 110*,* 34-47.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., Berg, A.C., 2016. Ssd: Single shot multibox detector*, European conference on computer vision*. Springer, pp. 21-37.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation*, Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431-3440.

Mallinis, G., Koutsias, N., Tsakiri-Strati, M., Karteris, M., 2008. Object-based classification using quickbird imagery for delineating forest vegetation polygons in a mediterranean test site. *ISPRS Journal of Photogrammetry and Remote Sensing* 63*,* 237-250.

Mohan, M., Silva, C.A., Klauberg, C., Jat, P., Catts, G., Cardil, A., Hudak, A.T., Dia, M., 2017. Individual tree detection from unmanned aerial vehicle (uav) derived canopy height model in an open canopy mixed conifer forest. *Forests* 8*,* 340.

Monnet, J.-M., Mermin, E., Chanussot, J., Berger, F., 2010. Tree top detection using local maxima filtering: A parameter sensitivity analysis*, 10th International Conference on LiDAR Applications for Assessing Forest Ecosystems (Silvilaser 2010)*, p. 9 p.

Mubin, N.A., Nadarajoo, E., Shafri, H.Z.M., Hamedianfar, A., 2019. Young and mature oil palm tree detection and counting using convolutional neural network deep learning method. *International Journal of Remote Sensing,* 1-16.

Özcan, A.H., Hisar, D., Sayar, Y., Ünsalan, C., 2017. Tree crown detection and delineation in satellite images using probabilistic voting. *Remote Sensing Letters* 8*,* 761-770.

Paoletti, M.E., Haut, J.M., Fernandez-Beltran, R., Plaza, J., Plaza, A., Li, J., Pla, F., 2018. Capsule networks for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing* 57*,* 2145-2160.

Pearse, G.D., Dash, J.P., Persson, H.J., Watt, M.S., 2018. Comparison of high-density lidar and satellite photogrammetry for forest inventory. *ISPRS Journal of Photogrammetry and Remote Sensing* 142*,* 257-267.

Pont, D., Kimberley, M.O., Brownlie, R.K., Sabatia, C.O., Watt, M.S., 2015. Calibrated tree counting on remotely sensed images of planted forests. *International Journal of Remote Sensing* 36*,* 3819-3836.

Pouliot, D., King, D., 2005. Approaches for optimal automated individual tree crown detection in regenerating coniferous forests. *Canadian Journal of Remote Sensing* 31*,* 255-267.

Qin, R., 2017. Automated 3d recovery from very high resolution multi-view satellite images*, ASPRS (IGTF) annual Conference*, Baltimore, Maryland, USA, p. 10.

Qin, R., Fang, W., 2014. A hierarchical building detection method for very high resolution remotely sensed images combined with dsm using graph cut optimization. *Photogrammetric Engineering & Remote Sensing* 80*,* 873-883.

Quackenbush, L.J., Hopkins, P.F., Kinn, G.J., 2000. Using template correlation to identify individual trees in high resolution imagery*, American Society for Photogrammetry & Remote Sensing (ASPRS) 2000 Annual Conference Proceedings, Washington DC*.

Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. You only look once: Unified, real-time object detection*, Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779-788.

Rolnick, D., Veit, A., Belongie, S., Shavit, N., 2017. Deep learning is robust to massive label noise. *arXiv preprint arXiv:1705.10694*.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation*, International Conference on Medical image computing and computer-assisted intervention*. Springer, pp. 234-241.

Saarinen, N., Vastaranta, M., Näsi, R., Rosnell, T., Hakala, T., Honkavaara, E., Wulder, M., Luoma, V., Tommaselli, A., Imai, N., 2017. Uav-based photogrammetric point clouds and hyperspectral imaging for mapping biodiversity indicators in boreal forests. *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences* 42.

Sherrah, J., 2016. Fully convolutional networks for dense semantic labelling of high-resolution aerial imagery. *arXiv preprint arXiv:1606.02585*.

Skurikhin, A.N., Garrity, S.R., McDowell, N.G., Cai, D.M., 2013. Automated tree crown detection and size estimation using multi-scale analysis of high-resolution satellite imagery. *Remote sensing letters* 4*,* 465-474.

Song, C., Dickinson, M.B., Su, L., Zhang, S., Yaussey, D., 2010. Estimating average tree crown size using spatial information from ikonos and quickbird images: Across-sensor and across-site comparisons. *Remote sensing of environment* 114*,* 1099-1107.

Strîmbu, V.F., Strîmbu, B.M., 2015. A graph-based segmentation algorithm for tree crown extraction using airborne lidar data. *ISPRS Journal of Photogrammetry and Remote Sensing* 104*,* 30-43.

Sun, X., Shen, S., Lin, X., Hu, Z., 2017. Semantic labeling of high-resolution aerial images using an ensemble of fully convolutional networks. *Journal of Applied Remote Sensing* 11*,* 042617.

Tarp-Johansen, M.J., 2002. Automatic stem mapping in three dimensions by template matching from aerial photographs. *Scandinavian journal of forest research* 17*,* 359-368.

Veit, A., Alldrin, N., Chechik, G., Krasin, I., Gupta, A., Belongie, S.J., 2017. Learning from noisy large-scale datasets with minimal supervision*, CVPR*.

Vincent, L., 1993. Morphological grayscale reconstruction in image analysis: Applications and efficient algorithms. *IEEE transactions on image processing* 2*,* 176-201.

Wang, Q., Liu, S., Chanussot, J., Li, X., 2018a. Scene classification with recurrent attention of vhr remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing* 57*,* 1155-1167.

Wang, Q., Yuan, Z., Du, Q., Li, X., 2018b. Getnet: A general end-to-end 2-d cnn framework for hyperspectral image change detection. *IEEE Transactions on Geoscience and Remote Sensing* 57*,* 3-13.

Wang, Y., Hyyppä, J., Liang, X., Kaartinen, H., Yu, X., Lindberg, E., Holmgren, J., Qin, Y., Mallet, C., Ferraz, A., 2016. International benchmarking of the individual tree detection methods for modeling 3-d canopy structure for silviculture and forest ecology using airborne laser scanning. *IEEE Transactions on Geoscience and Remote Sensing* 54*,* 5011-5027.

Weng, E., Malyshev, S., Lichstein, J., Farrior, C., Dybzinski, R., Zhang, T., Shevliakova, E., Pacala, S., 2015. Scaling from individual trees to forests in an earth system modeling framework using a mathematically tractable model of height-structured competition. *Biogeosciences* 12*,* 2655-2694.

Wulder, M., Niemann, K.O., Goodenough, D.G., 2000. Local maximum filtering for the extraction of tree locations and basal area from high spatial resolution imagery. *Remote Sensing of environment* 73*,* 103-114.

Xiao, C., Qin, R., Huang, X., Li, J., 2018. Individual tree detection from multi-view satellite images*, IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, pp. 3967-3970.

Xiao, C., Qin, R., Xie, X., Huang, X., 2019. Individual tree detection and crown delineation with 3d information from multi-view satellite images. *Photogrammetric Engineering & Remote Sensing* 85*,* 55-63.

Zhao, D., Pang, Y., Li, Z., Liu, L., 2014. Isolating individual trees in a closed coniferous forest using small footprint lidar data. *International journal of remote sensing* 35*,* 7199-7218.

Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2017. Pyramid scene parsing network*, Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2881-2890.

Figure List

Figure 1. The framework of using pseudo labels from the unsupervised detector to train FCN for treetop detection.

Figure 2. The two study areas. The larger one (Area A) is a typical urban area with different kind of trees and vegetation. The smaller one (Area B) is a densely forested area where the naturally grown trees are difficult to distinguish.

Figure 3. The orthophoto (a) and DSM (b) data of the study area with an example of trees in image (c) and their 3D shapes marked by circles in image (d).

Figure 4: The illustration of the local maxima treetop detection. Based on the DSM (b) and the NDVI mask (c), the true treetops (blue dots in rectangles) are finally detected in the fused-colours (near-IR2, red edge, and yellow) orthophoto (image (a)).

Figure 5. The generation of the training samples. Image (a) and image (b) show the test sites (marked by rectangles). The image (c) shows one test site with four training samples and their three-channel fused colour images and labels (image (d)).

Figure 6. The architecture of the adopted FCN. The numbers are the image/feature sizes and the feature channels in each block.

Figure 7. The reference data of two test sites. Colours are the means to distinguish closely adjacent trees.

Figure 8. The treetop detection results of test site 3 and site 7. The dots are treetops detected by FCN while the circles are treetops detected by THR. Image (a) and (d) are the RGB colour

image of the two test sites while image (b) and (e) are the detected treetops on the fused image with their zoom-in areas (yellow dash-line rectangles) in image (c) and image (f). Examples of incorrect detections are marked by a red rectangle in image (b).

Figure 9. The treetop detection in the densely forested area. Image (a) and (b) show the floated RGB image and part of the DSM in the yellow dash-line rectangle. Image (c) and (d) show the detection results where blue dots are treetops detected by the FCN and the red circles are treetops detected by the THR.

Figure 10. The treetop detection results in the whole urban area.

Tables List