Estimation of tree structure parameters from video frames with removal of blurred images using machine learning

Kenta Itakura and Fumiki Hosoi†
(Graduate School of Agricultural and Life Sciences, The University of Tokyo, Yayoi 1-1-1, Bunkyo-ku, Tokyo 113-8657, Japan)

Abstract

Recently, structure from motion (SfM), which converts multiple images to a detailed three-dimensional (3D) model, has been used to extract 3D structural information about vegetation. However, multiple still images (e.g., >100 images) are necessary for the 3D reconstruction. This requires multiple shutter releases, but taking many images is time consuming and labor intensive. One possible solution is to take video recordings from which many images can be obtained by dividing the video clips into video frames. However, frames from videos are sometimes blurred owing to camera vibration, which leads to inaccurate construction of the 3D model. Furthermore, their resolution is lower than that of still images, which may lead to inaccurate 3D reconstruction and estimation error of tree trunk diameter, tree height, and the number of trees observable in the 3D images. We propose a method to record videos, remove blurred video frames using machine learning, and construct 3D images. We compare the accuracy of the 3D models reconstructed from video frames with that from the still images. The blurred video frames were classified by a convolutional neural network (CNN) with an accuracy of 97%. The classification to remove these video frames improved the accuracy of the 3D models based on video frames taken at a walking speed of more than 0.5 m/s, which included many blurred ones. There was no significant difference in the accuracy of tree trunk diameter and tree height estimation between the 3D models obtained from the video frames and the still images when using the CNN classification. At a close enough distance (e.g., 20 m), the 3D model reconstructed from video frames was as accurate as the models constructed from still images. Video recording enables effective data collection for SfM, and the present method can be applicable to the 3D reconstruction of trees in various fields.

Key words: Convolutional neural network (CNN), Structural parameter estimation, Structure from motion (SfM), Three-dimensional imaging, Video

1. Introduction

In a variety of fields such as urban forestry, arboriculture, ecological horticulture, and forestry, it is important to measure the three-dimensional (3D) structures of plants (Morgenroth and Gomez, 2014). This is because plant structure closely relates to important functions such as the cycling of materials and energy through photosynthesis and transpiration, and the maintenance of plant microclimates (Hosoi et al., 2010).

Estimating the structural parameters of trees using two-dimensional (2D) images has been widely used. For example, Mano (2017) proposed a height measurement system using a commercial time-lapse camera and Nagai et al. (2018) evaluated the spatial distribution of the forest disturbance over large areas with 2D images obtained from a drone. However, an analysis using 2D images is not accurate enough to estimate structural parameters owing to the complicated 3D structure of trees (Konishi et al., 2009).

With the exponential increase in computational power and the widespread availability of digital cameras, the use of a photogrammetric approach called structure from motion (SfM) is becoming widespread (Müller-Linow et al., 2015; Rose et al., 2015; Obanawa et al., 2014; Dandois and Ellis, 2013). SfM is regarded as a more developed version of stereo photogrammetry, and entails converting a few images to detailed 3D models of the target objects by estimating the position of the camera (Hayakawa and Obanawa, 2016; James and Robson, 2012; Verhoeven et al., 2012). With SfM, the size of a paddy field, and the height (Teng et al., 2016) and growth stage of its rice plants (Hama et al., 2016) could be estimated with high accuracy; a method for estimating biomass based on plant height derived from crop surface models has also been reported (Bendig et al., 2014).

In previous studies, SfM was often utilized with an unmanned aerial vehicle, and studies with use of images taken on the ground for SfM are limited (Obanawa et al., 2014). However, it has recently been reported that structural parameters such as leaf area and plant height can be estimated accurately with use of images taken in the laboratory and on the ground for SfM (Rose et al., 2015; Zhang et al., 2016; 2018). Therefore, it is expected that SfM will become widely used for the images taken on the ground as well in the future.

However, there are some problems with using videos taken on the ground for SfM. One issue is that numerous still images (e.g., >100 images) are necessary for 3D reconstruction. Multiple still images require multiple shutter releases, which can be time consuming and labor intensive. One possible solution for these issues is to take videos from which many images can be obtained.
by dividing video clips into video frames (in this paper, we call the images from videos “video frames”). In this way, the shutter does not need to be released at each frame, thus shortening the time needed for capturing multiple images. However, two concerns should be considered. First, some blurred frames are included within the video frames, mainly due to camera vibration. If a target frame is blurred, corresponding points in each frame pair cannot be detected, and it becomes difficult to match picture pairs. As a result, accurate 3D point cloud images cannot be built, and this leads to error in tree structure parameter estimation. Second, when constructing 3D point clouds using video frames, it is not known whether structural parameters can be estimated as accurately as when constructing 3D point clouds from still images, because the resolution of video frames is lower than that of still images. The resolution of an object in an image decreases as the distance between the camera and the object increases, which may result in the failure of 3D reconstruction and lead to estimation error in the structural parameters and the number of trees observable in the 3D images. Therefore, the estimation accuracy of tree trunk diameter and tree height, and the number of trees observable at a study site should be considered when reconstructing 3D models from still images and video frames taken while measurement conditions are changing, such as the number of video frames and still images or the distance between the camera and the target tree.

We may be able to solve the first problem if blurred video frames can be detected and removed automatically using a recent image classification technique. In the field of image classification, the latest generation of convolutional neural networks (CNN) has achieved impressive results (Sladojevic et al., 2016; Arai and Nagao, 2017). CNN employ a mathematical operation called convolution (Ian, et al., 2016), which may be able to offer a solution to the first problem. As for the second problem, by comparing the estimation accuracy of some fundamental structural parameters, such as tree trunk diameter and tree height, we can examine 3D point clouds obtained from video frames and still images when we change the distance between the object and camera.

This study aims (1) to investigate whether blurred images from videos can be detected and removed by using CNN classification, and to assess if this removal contributes to an improvement in the accuracy of 3D point clouds, and (2) compare the accuracy of tree trunk diameter, tree height, and number of trees observed at a study site using 3D point clouds obtained from video frames and still images.

2. Materials and Methods

2.1 Study site and direct measurement of tree trunk diameter and height

This experiment was conducted at the 58.3 ha Shinjuku Gyoen National Garden in Tokyo, Japan, which has more than 10,000 trees, such as cherry blossom (Cerasus Mill.), turp tree (Liriodendron tulipifera), plane trees (Platanus), Himalayan cedar (Cedrus deodara), and bald cypress (Taxodium distichum). This site includes a very wide variety of species; thus, we were able to select trees appropriate for the experiments. Five plane trees (platanus) were selected for the estimation of tree trunk diameter. We bound the tree trunks with vinyl rope, and measured the diameter using a tree caliper on the vinyl rope. The diameters ranged from 38.8 to 53.7 cm.

Two white enkianthu (Enkianthus perulatus) and three cherry blossoms (Cerasus Mill.) were selected for the tree height estimation. The height ranged from 3.33 to 6.16 m. We measured the heights using a range finder (Lasertape FG-21-HA, RIEGL, Australia) with an accuracy of ± 5 cm. Tree height was measured with the range finder 5 times, and the average was used as the actual tree height to mitigate the measurement error.

The Yayoi campus of the University of Tokyo, Tokyo, Japan was also selected as a study site. Japanese zelkava (Zelkova serrata) was selected to study the relationship between the estimation accuracy of tree trunk diameter, the distance between camera and target, and the number of images used for 3D reconstruction. The vinyl rope was placed at a total of three points, and tree trunk diameter estimation was conducted. In this study, the selected tree trunks were not hidden by shrubs and the tree canopies were not overlapped with other canopies. As a first step for 3D reconstruction from videos aiming for forest investigation, the effectiveness of the use of video should be evaluated correctly without the influence of complexity of plants. From this reason, the trees in this study were chosen.

2.2 Acquisition of videos and still images

To acquire the training datasets for the CNN, a wide variety of video frames of vegetation were collected at different walking speeds to obtain many blurred and non-blurred video frames, both in Shinjuku Gyoen National Garden and on the campus of the University of Tokyo. We then obtained video frames of trees to reconstruct the 3D models and evaluate the effectiveness of blurred frame removal using CNN classification. For this purpose, we varied walking speed to 0.33, 0.50, and 1.00 m/s, which correspond to slow, normal, and fast walking speeds, respectively, while taking videos alongside the trees selected in Shinjuku Gyoen National Garden. When taking video, the camera vibrated slightly in the vertical direction, and this was the main factor that caused the blurred frames. The amplitude of the vibration was approximately 1.5 cm, and the frequency was 1.00, 2.00, and 3.00 Hz when walking speed was 0.33, 0.50, and 1.00 m/s, respectively. Because the vibration frequency depended on the walking speed, the percentage of blurred frames varied according to the walking speed. The distance between the camera and the tree was 3 m, and the camera was handheld. Still images of the target trees were also taken to compare with the 3D point clouds obtained from the videos. We used a Canon EOS M2 (Japan) to take videos and still images with resolutions of $1088 \times 1920$ pixels and $3456 \times 5184$ pixels, respectively. One video and 120 still images of a study site in Shinjuku Gyoen National Park were also taken to compare the 3D models obtained from video frames with those obtained from still images. At the study site, we captured 120 images for the 3D models obtained from video frames with those obtained from still images. It took less than 2 min to capture the video, and approximately 5 min to capture the 120 still images.

To relate the accuracy of tree trunk diameter estimation to the distance between camera and object and the number of
video frames without blur, still images and video were taken at the University of Tokyo. The distance between the camera and the tree was 10, 20, and 30 m. The number of images was 10, 20, and 30, and the walking speed used while recording videos was about 0.2 m/s, at which speed there were no blurred video frames. This walking speed was chosen to compare the 3D models from still images with models constructed from video frames that were not blurred.

2.3 Removal of blurred video frames using CNN classification, and its influence on reconstruction of 3D point cloud images

A CNN, which is one type of deep learning algorithm, was used for the blurred video frame classification. A CNN is a flexible model that can produce outputs after training in accordance with our aims. Thus, we can design a neural network by letting it learn what kind of output it should produce from the input (Ota et al., 2017). In this study, one of the most general types of CNN, containing a convolution layer, pooling layer, and fully-connected layer, was adopted (Okatani, 2013; Fujita and Takahara, 2016; Arai and Nagao, 2017; Nomura et al., 2018).

The workflow of the study, from capturing images to the tree trunk and height estimation, is represented in Fig. 1. The videos of the plants taken at different walking speeds for training the CNN were classified into blurred video frames and non-blurred video frames by visual inspection, and the dataset (2000 blurred video frames and 2000 non-blurred video frames) was prepared from these video frames. We selected 1500 and 200 video frames randomly to be used as training data and test data, respectively, from each class. Input video frames were converted to Fourier-transformed images. We did not use raw video frames (the frames before the Fourier transform) for training the CNN, because the blur in the frames increases the low frequency component in the spatial frequency domain, and this increase can be observed clearly in their Fourier transform, which represent the distribution of frequencies in the video frames. It is expected that the accuracy of the classification increases by using Fourier-transformed images for feature extraction. After obtaining a classifier with high classification accuracy, video frames for SfM obtained with variable walking speed were classified as blurred or non-blurred by the classifier.

In the next step, 3D point clouds were built using the image files obtained from videos and still images with and without blurred frames. For 3D point cloud image construction using SfM (Hayakawa et al., 2014; Dandois et al., 2015), about 25 video frames and still images were used. The software used for the construction was Agisoft Photoscan Professional (Agisoft LCC, Russia). The PC used in this study was equipped with an Intel® Core™ i7-6700 processor with a clock frequency of 3.40 GHz, 8.0 GB of RAM, and an Intel® HD Graphics 530 GPU. The tree trunk diameter and tree height estimation were then conducted with the constructed 3D models. In the 3D point clouds, the vinyl rope on the tree trunk could be observed, and tree trunk diameter was estimated by calculating the distance.

Fig. 1. The study workflow, from capturing images to tree trunk and height estimation. In addition to videos (a), still images (b) were taken to compare the accuracy of 3D point clouds derived from video frames and still images.
between the points on the vinyl rope that was used for measuring the diameter by the tree caliper method in the field. For tree height estimation, the distance between the top part of the tree canopies and the ground under the top part was calculated from the 3D point clouds. Tree trunk diameters and tree heights in the 3D point clouds were measured three times, and their average values were represented as estimated values (Pan et al., 2017). In the 3D point cloud images, some points corresponding to the point on the vinyl rope existed. To reduce the error from the selection, the diameter was calculated three times and its average value was used as the estimated tree trunk diameter. Finally, estimation errors were compared between the 3D point clouds based on video frames with and without blur at each walking speed, and the still images.

3. Results and Discussion

3.1 Blurred video frame classification using CNN

The accuracy of the classification between blurred video frames and non-blurred video frames was 97.0%, and the training time was 12 min. Video frames that were incorrectly classified were only slightly blurred, which shows that blurred video frames were classified with high accuracy. As shown in Fig. 2, classification accuracy with raw video frame input instead of Fourier-transformed images was about 90% at most. For any number of input images, the accuracy when using Fourier-transformed images was significantly higher ($p < 0.01$).

One of the advantages of the CNN compared to other classification methods is that the CNN classifier can automatically find features necessary for image classification (e.g., edges, color, and spatial pattern) and use them for the task, whereas in the other classification methods, the operator must manually find and input the features, and then classify images in accordance with them. However, as shown in the results, if the features that contribute to classification are known beforehand, converting the input images to other images that emphasize those features (e.g., Fourier-transformed images in this study) will increase the classification accuracy.

The classification procedure (i.e., video input, video frame division, conversion to Fourier-transformed image, classification by CNN, and rearrangement of the video frames for SfM) can be conducted automatically; thus, the classification of blurred video frames could potentially be processed with high efficiency.

3.2 Comparison of the estimation of tree trunk diameter and tree height with and without blurred video frame removal by CNN classification

Figure 3 and Fig. 4 represent absolute tree trunk diameter and tree height estimation errors when using input video frames with and without blurred video frame removal by CNN classification, respectively. In the tree trunk diameter estimation, no significant difference is found in the walking-slowly condition, however, in the walking-normally and walking-fast conditions, the absolute errors without CNN classification were higher than those with CNN classification by 1 cm and 3 cm, respectively. The difference was significant in the walking-fast condition ($p < 0.01$). On the other hand, the accuracy of tree trunk diameter using video frame sets with CNN classification showed big error bars in Fig. 3 and no difference between the walking speed conditions ($p > 0.05$). In the tree trunk diameter estimation, the ratio of video frames classified as blurred video frames by the CNN classification under the slow, normal, and fast walking conditions was about 28, 35, and 55%, respectively. In tree height estimation, when input video frames were classified by CNN classification and blurred video frames were removed, the errors were smaller at each walking speed than when input video frames were not classified by the CNN. In Figs. 3 and 4, there was no significant difference between the results of tree trunk diameter and height estimation when using video frames that had been classified by the CNN and still images.

It is difficult to find corresponding points between image pairs when images are blurred, but finding the corresponding points is necessary for image matching in SfM. The blur in the images leads to an inaccurate 3D point cloud construction and increases tree diameter and height estimation errors. The estimation errors between the points on the vinyl rope that was used for measuring the diameter by the tree caliper method in the field. For tree height estimation, the distance between the top part of the tree canopies and the ground under the top part was calculated from the 3D point clouds. Tree trunk diameters and tree heights in the 3D point clouds were measured three times, and their average values were represented as estimated values (Pan et al., 2017). In the 3D point cloud images, some points corresponding to the point on the vinyl rope existed. To reduce the error from the selection, the diameter was calculated three times and its average value was used as the estimated tree trunk diameter. Finally, estimation errors were compared between the 3D point clouds based on video frames with and without blur at each walking speed, and the still images.

3. Results and Discussion

3.1 Blurred video frame classification using CNN

The accuracy of the classification between blurred video frames and non-blurred video frames was 97.0%, and the training time was 12 min. Video frames that were incorrectly classified were only slightly blurred, which shows that blurred video frames were classified with high accuracy. As shown in Fig. 2, classification accuracy with raw video frame input instead of Fourier-transformed images was about 90% at most. For any number of input images, the accuracy when using Fourier-transformed images was significantly higher ($p < 0.01$).

One of the advantages of the CNN compared to other classification methods is that the CNN classifier can automatically find features necessary for image classification (e.g., edges, color, and spatial pattern) and use them for the task, whereas in the other classification methods, the operator must manually find and input the features, and then classify images in accordance with them. However, as shown in the results, if the features that contribute to classification are known beforehand, converting the input images to other images that emphasize those features (e.g., Fourier-transformed images in this study) will increase the classification accuracy.

The classification procedure (i.e., video input, video frame division, conversion to Fourier-transformed image, classification by CNN, and rearrangement of the video frames for SfM) can be conducted automatically; thus, the classification of blurred video frames could potentially be processed with high efficiency.

3.2 Comparison of the estimation of tree trunk diameter and tree height with and without blurred video frame removal by CNN classification

Figure 3 and Fig. 4 represent absolute tree trunk diameter and tree height estimation errors when using input video frames with and without blurred video frame removal by CNN classification, respectively. In the tree trunk diameter estimation, no significant difference is found in the walking-slowly condition, however, in the walking-normally and walking-fast conditions, the absolute errors without CNN classification were higher than those with CNN classification by 1 cm and 3 cm, respectively. The difference was significant in the walking-fast condition ($p < 0.01$). On the other hand, the accuracy of tree trunk diameter using video frame sets with CNN classification showed big error bars in Fig. 3 and no difference between the walking speed conditions ($p > 0.05$). In the tree trunk diameter estimation, the ratio of video frames classified as blurred video frames by the CNN classification under the slow, normal, and fast walking conditions was about 28, 35, and 55%, respectively. In tree height estimation, when input video frames were classified by CNN classification and blurred video frames were removed, the errors were smaller at each walking speed than when input video frames were not classified by the CNN. In Figs. 3 and 4, there was no significant difference between the results of tree trunk diameter and height estimation when using video frames that had been classified by the CNN and still images.

It is difficult to find corresponding points between image pairs when images are blurred, but finding the corresponding points is necessary for image matching in SfM. The blur in the images leads to an inaccurate 3D point cloud construction and increases tree diameter and height estimation errors. The estimation errors
and error bars for tree height estimation were much bigger than those for tree trunk diameter estimation. Many points constitute tree trunk diameter. On the other hand, the region of a tree top is very small, and it is difficult to capture a tree top from an image. This caused larger errors in tree height estimation than in trunk diameter estimation.

The resolution of a video frame is lower than that of a still image. Because the distance between the camera and trees was comparatively short (3 m) in this study, 3D point clouds had a sufficiently high resolution even when the input images were obtained from videos. Thus, there was no significant difference between the estimation errors of video frames and still images.

By using our method to construct 3D images from video frames, we do not have to release a shutter many times to capture the target trees in a study site. As another way of effective image acquisition, interval shooting is known. However, videos can take images, for example, about 30 times as much as interval shooting per one second. Owing to the frame rate, even if the

![Graph showing tree height estimation errors with the input video frames before and after removal of blurred video frames through CNN classification. The estimation error of video frame sets without CNN classification is represented by “raw”, meaning raw video frames. The estimation error from 3D models reconstructed from still images is represented by “Still Images”.](image1)

**Fig. 3.** Tree trunk diameter estimation errors with the input video frames before and after removal of blurred video frames through CNN classification. The estimation error of video frame sets without CNN classification is represented by “raw”, meaning raw video frames. The estimation error from 3D models reconstructed from still images is represented by “Still Images”.

![Graph showing tree height estimation errors with the input video frames before and after removal of blurred video frames through CNN classification. The estimation error of video frame datasets without CNN classification is represented by “raw”, meaning raw video frames. The estimation error from 3D models reconstructed from still images is represented by “Still Images”.](image2)

**Fig. 4.** Tree height estimation errors with the input video frames before and after removal of blurred video frames through CNN classification. The estimation error of video frame datasets without CNN classification is represented by “raw”, meaning raw video frames. The estimation error from 3D models reconstructed from still images is represented by “Still Images”.
moving speed is high, a series of video frames which sufficiently covers the object can be obtained. When taking videos for SfM, further, non-blurred video frames can be selected from the lots of videos, which offers the appropriate selection of image sets for SfM. Additionally, the time to take videos of the study site is less than one fifth of the time of taking still images. From these reasons, the technique for 3D image acquisition with videos can be widely applicable for any researchers and investigators, and it will contribute to the efficient and effective forest investigations. The 3D structure of a tree offers various kinds of information such as light condition and wind condition under the canopy. Together with photosynthetic models, the amount of photosynthetic production would be estimated accurately using the 3D information. Also, the 3D tree structure influences the wind condition which is an important factor in the microclimate (Watanabe et al., 2002), and the relationship between the 3D tree structure and wind condition can be elucidated using the present 3D acquisition method clearly.

3.3 Comparison between the estimation accuracy of tree trunk diameter with still images and video frames

Table 1 shows the absolute error of tree trunk diameter estimation from still images and video frames as the distance between the target and camera is changed. As the number of images increases, and as the distance between the camera and object becomes shorter, the error decreases. When the distance is 30 m and video frames were used, the density of the 3D point cloud from video frames was too low to estimate tree trunk diameter. On the other hand, with still images, even at a distance of 30 m, tree trunk diameter could be measured with an accuracy of 2 cm. According to Itakura et al. (2018), tree trunk diameter estimation using SfM can be conducted with an accuracy of 1 to 2 cm. In this study, the error using video frames was less than 3 cm when the distance was 20 m. When the distance was less than 20 m, an accurate 3D point cloud model could be built from video frames. One of the most influential factors that enables the construction of an accurate 3D point cloud model is the actual length per side of a pixel. As the actual length per side of a pixel which is related to resolution increases, the resolution of the object in the image increases and the accuracy of the model will be increased as well. In this study, the actual length per side of a pixel in a video frame was about 0.6, 1.0, and 1.8 cm, respectively, when the distance between the camera and the tree trunk was 10, 20, and 30 m. The actual distance with still video at each distance was about 0.2, 0.3, and 0.6 cm. Higher resolution makes it easier to find corresponding points, and leads to a decrease in error in 3D point cloud models.

3.4 Comparison of 3D point clouds of several trees obtained from video frames and still images

Figure 5a and 5b show the 3D point clouds constructed with 120 still images and video frames without blur. In Fig. 5a and 5b, ten and eight trees can be observed, respectively. In Fig. 5b, two trees in an area surrounded by a white circle could not be seen, which resulted in a difference between the observable trees in Fig. 5a and Fig. 5b. The average distance to the nearest neighbor points in the 3D point clouds are 0.18 and 0.47 cm in Fig. 5a and 5b, respectively.

Because the resolution of video frames is lower than that of still images, the point density of the 3D point cloud from the video frames is lower than that from still images. As a result, trees that are far from the camera cannot be reconstructed, and the number of trees observable in the 3D point cloud decreases in the 3D point clouds based on video frames. The trees that can be observed in Fig. 5a, and cannot observed in Fig. 5b were 28 m away from the camera, and this result is in agreement with the finding that trees with a distance of about 30 m away from the camera cannot be reconstructed, as shown in Table 1. For a better 3D model reconstruction, the distance between camera and object should be as short as possible. However, if the distance is too short, the coverage area of the image will be decreased. Therefore, considering these facts, the distance should be appropriately determined to keep enough coverage and the resolution of the images (e.g. the distance between trees and cameras: 10 ~ 20 m).

4. Conclusions

Using the Fourier-transformed images, the blurred video frames could be removed using CNN classification with an accuracy of 97%. The removal raised the accuracy of the 3D models based on video frames captured at a walking speed more than 0.5 m/s. There was no significant difference in the accuracy of tree trunk diameter and tree height estimation between the 3D models obtained from video frames and those obtained from still images when using CNN classification. Although the estimation error increased as the distance between the camera and object increased, when the distance was short enough (e.g., 20 m), the 3D model reconstructed from video frames was as accurate as one reconstructed from still images.

In this study, it was proven that the accuracy of 3D models derived from videos can be increased by removing blurred video frames through CNN classification. Video photography allows effective data collection for SfM. The tree structural parameters such as tree trunk diameter, tree height, leaf inclination angle and leaf area are very important. Recently, the 3D measurement of plants with SfM has offered easy and efficient acquisition of these information. However, it was difficult to get the 3D images that cover large area of forest with the conventional
image acquisition method (i.e., release a shutter every time). On the other hand, the video recording method for SfM shows the possibility of obtaining 3D image of a large-scale study site. With the combination of SfM and video recordings, tree biomass in the forest can be estimated from the tree trunk diameter and height. Further, by estimating the detailed parameters such as leaf area and leaf inclination angle, the mechanisms of tree structural adaptation in reaction to the ambient environment such as light condition and temperature can be investigated easily. The present method offers accurate and effective forest assessment. Moreover, it provides us with the understanding of forest microclimate and ecosystem as well. Thus, it is desirable that the present method be applied for 3D reconstruction of trees in various fields, with consideration for the distance between the camera and the target trees.

**Acknowledgements**

We are deeply grateful to Dr. Martin O’Brien for proofreading our manuscript.

**References**

Arai S, Nagao T, 2017: Intuitive visualization method for image classification using convolutional neural networks. *Information processing of Japan* 10, 1–13.

Bendig J, Bolten A, Bennertz S, Broscheit J, Eichfuss S, Bareth G, 2014: Estimating biomass of barley using crop surface models (CSMs) derived from UAV-based RGB imaging. *Remote Sensing* 6, 10395–10412.

Dandois JP, Olano M, Ellis EC, 2015: Optimal altitude, overlap, and weather conditions for computer vision UAV estimates of forest structure. *Remote Sensing* 7, 13895–13920.

Dandois JP, Ellis EC, 2013: High spatial resolution three-dimensional mapping of vegetation spectral dynamics using computer vision: *Remote Sensing of Environment* 136, 259–276.

Fujita K, Takahara A, 2016: Implementation of deep learning Ohmusha, Japan (Tokyo), pp 46.

Hama A, Hayazaki Y, Mochizuki A, Tsuruoka Y, Tanaka K, Kondoh A, 2016: Rice growth monitoring using small...
UAV and SfM-MVS technique. *Journal of Japan society of hydrology and water resources* **29**, 44–54.

Hayakawa Y, Obanawa H, Gomez C, 2014: 3D Modelling of inaccessible areas using UAV-based aerial photography and structure from Motion. *The Japanese Geomorphological Union* **35**, 283–294.

Hayakawa Y, Obanawa H, 2016: Aerial measurements of topography using small UAS and SfM-MVS photogrammetry. *Butsuri Tansa* **69**, 297–309.

Hosoi F, Nakay I, Omasa K, 2010: Estimation and error analysis of woody canopy leaf area density profiles using 3-D airborne and ground-based scanning lidar remote-sensing techniques. *IEEE Transactions on Geoscience and Remote Sensing* **48**, 2215–2223.

Ian G, Yoshua B, Aaron C, 2016: Deep learning, The MIT press, England (London), pp 321.

Itakura K, Kamakura I, Hosoi F, 2018: A Comparison study on three-dimensional measurement of vegetation using Lidar and SfM on the ground. *Eco-engineering* **30**, 15–20.

James MR, Robson S, 2012: Straightforward reconstruction of 3D surfaces and topography with a camera: Accuracy and geoscience application. *Journal of Geophysical Research* **117**, F0317.

Konishi A, Eguchi A, Hosoi F, Omasa K, 2009: 3D monitoring spatio–temporal effects of herbicide on a whole plant using combined range and chlorophyll a fluorescence imaging. *Functional Plant Biology* **36**, 874–879.

Mano M, 2017: Precise and continuous measurement of plant heights in an agricultural field using a time-lapse camera. *Journal of Agricultural Meteorology* **73**, 100–108.

Morgenroth J, Gomez C, 2014: Assessment of tree structure using a 3D image analysis technique – A proof of concept. *Urban Forestry & Urban Greening* **13**, 198–203.

Müller-Linow M, Pinto-Espinosa F, Scharr H, Rascher U, 2015: The leaf angle distribution of natural plant populations: assessing the canopy with a novel software tool. *Plant methods* **11**, 11.

Nagai S, Saitoh TM, Kajiwara K, Yoshitake S, Honda Y, 2018: Investigation of the potential of drone observations for detection of forest disturbance caused by heavy snow damage in a Japanese cedar (*Cryptomeria japonica*) forest. *Journal of Agricultural Meteorology* **74**, 123–127.

Nomura Y, Ida I, Miyaji T, Miyamoto M, Suga M, 2018: Structural integrity diagnosis for valve based on deep convolution neural network. *Journal of the Society of Materials science, Japan* **67**, 177–183.

Obanawa H, Hayakawa Y, Saito H, C Gomez, 2014: Comparison of DSMs derived from UAV-SfM method and terrestrial laser scanning. *Japan society of photogrammetry and remote sensing* **53**, 67–74.

Okatani T, 2013: Deep learning for image recognition. *Journal of Japanese Society of Artificial Intelligence* **28**, 962–974.

Ota S, Takenouchi H, Tokumaru M, 2017: Kansei retrieval of clothing using features extracted by deep neural network. *Japan society of kansei engineering* **16**, 277–283.

Pan Y, Kuo K, Hosoi F, 2017: A study on estimation of tree trunk diameters and heights from three-dimensional point cloud images obtained by SLAM. *Eco-Engineering* **29**, 17–22.

Rose JC, Paulus S, Kuhlmann H, 2015: Accuracy analysis of a multi-view stereo approach for phenotyping of tomato plants at the organ level. *Sensors* **15**, 9651–9665.

Sladojevic S, Arsenovic M, Anderla A, Culibrk D, Stefanovic D, 2016: Deep neural networks based recognition of plant diseases by leaf image classification. *Computational intelligence and neuroscience* **2016** (Article ID 3289801).

Teng P, Zhang Y, Shimizu Y, Hosoi F, Omasa K, 2016: Accuracy assessment in 3D remote sensing of rice plants in paddy field using a small UAV. *Eco-Engineering* **28**, 107–112.

Verhoeven G, Doneus M, Briese C, Vermeulen F, 2012: Mapping by matching: a computer vision-based approach to fast and accurate georeferencing of archaeological aerial photographs. *Journal of Archaeological Science* **39**, 2060–2070.

Watanabe T, Otsuka N, Saito M, 2002: Prediction of wind flow change around the development area caused by land deformation and tree felling. *Shigen to Sozai* **118**, 473–480.

Zhang Y, Teng P, Shimizu Y, Hosoi F, Omasa K, 2016: Estimating 3D Leaf and Stem Shape of Nursery Paprika Plants by a Novel Multi-Camera Photography System. *Sensors* **16**, 974.

Zhang Y, Teng P, Aono M, Shimizu Y, Hosoi F, Omasa K, 2018: 3D monitoring for plant growth parameters in field with a single camera by multi-view approach. *Journal of agricultural meteorology* **74**, 129–139.