Robotic forest vegetation quantification for environmental monitoring using hand-held cameras

4M25 Advanced Robotics Final Technical Report, April 2022
Andrew Wang, Department of Engineering, University of Cambridge

Abstract

Automatic vegetation mapping in dense forest areas is desirable for a number of environmental reasons, allowing non-technical users and robots to efficiently monitor local natural environments for better forest management. We propose a computer vision model which quantifies videos taken from walking around a forest to produce a vegetation score at each point. We show that this can be done by efficiently training neural networks for depth estimation and semantic segmentation to understand the forest image scenes.

We train and evaluate on a forest image dataset from the literature. The model achieves high correlation with human-labelled vegetation scores on a sample test video and low error on the image networks on the validation set.

All code and experiments can be found on

1 Introduction

Vegetation mapping is an important task in many fields of practical local environmental monitoring. Non-technical researchers and robots may wish to automatically quantify vegetation in given hand or robot-recorded video scenes. This can be used to compare vegetation levels from time to time in forest and other natural environments, or by combining GPS data to produce vegetation heatmaps. Example applications are for quantifying (de)forestation [1], natural damage such as by pests [2], and for environmental regeneration [3]. We focus on face-forward head-height scenes of local forest vegetation and not solutions using birds eye aerial imagery, as these can't extract local details in dense forest cover.

This removes the time-consuming and laborious need to manually quantify vegetation. Current methods are largely field-based and inefficient. [4] use quadrat sampling at sparse locations and at the timescale of about twice a decade, to monitor forest conservation and rewilding efforts in the forest. This sort of technique involves calculating a rough estimate of percentage cover using the Domin scale [5], and spatiotemporal extrapolation to give an indicator of vegetation cover in the forest. Basic image statistics-based methods also exist for specific datasets [6]. Methods such as Normalized Difference Vegetation Index [7] rely on aerial imagery so are irrelevant here.

We develop a computer vision model to solve this robotics task. This consists of a mathematical model to quantify the "vegetation levels" in an unstructured forest scene, where object locations and sizes are not well-defined, unlike urban scenes. Then,

given videos taken by a hand-held camera as one walks through forest paths or by an autonomous robotic vehicle, our model detects and quantifies vegetation present per frame to record the total vegetation presented during the video.

We train our model with a very low amount of labelled data using transfer learning (TL), which leverages an image recognition network such as ResNet [8], pre-trained on huge datasets of everyday images, to tune the network to our custom dataset. This means that we need fewer images to train the model to a sufficient accuracy. Similarly, our approach is easily applied to other similar areas such as monitoring of agricultural crops, of small low-level shrubbery or protected species etc.

Finally, we assess the quality of our automated image-based vegetation quantification method by comparing it to manual methods.

2 Model

Given an image at a time t, we consider the vegetation present in the *vicinity* $\mathcal{V}_t$ of the current scene, denoted by the true vegetation index of the kth class $VI_{k,t}$, and predicted by our model $\widehat{VI}_{k,t}$.

Firstly we consider the vicinity to be all points in the scene that are, from a human interpretation of the scene, close to the camera, i.e.

$$\mathcal{V} = x, y : \mathcal{D}_w(x, y||x_c, y_c) \leq \gamma_w$$

where x, y are real world coordinates, x_c, y_c are the camera coordinates in the real world, $\mathcal{D}_w(\cdot||\cdot)$ is a real world distance measurement, and γ_w is some distance threshold.

Then we create manual vegetation index labels $VI_{k,t}$ by using a DOMIN-like scale [5] on each image's vicinity, since it is not possible for us to manually sample from the physical locations of our forest image datasets.

2.1 Overall model design

Given this ground truth model above, our model behaves in a similar way: detect the vicinity of the scene and identify the vegetation present. Firstly, we assume that $\mathcal{D}_w(\cdot||x_c, y_c)$ is proportional to the *depth* D_w of a real world point x, y:

$$\mathcal{D}_w(x, y||x_c, y_c) \propto D_w(x, y)$$

Then, given only the pixels of an image u, v, we approximate the depth $D_w(x, y)$ with a function $\hat{D}(u, v)$ called the pixel depth map, that we learn from data. Note this is implicity performing a pixel-to-world camera calibration $(u, v) - > (x, y)$. Then the vicinity in the image scene $\hat{\mathcal{V}}$ is

$$\hat{\mathcal{V}} = u, v : \hat{D}(u, v) \leq \hat{\gamma}$$

where $\hat{\gamma}$ is set manually: further work will calculate this parameter by calibrating the exactness of the real-world vicinity to that calculated by the depth map. For now, we we calibrate the whole time-series using a factor $\frac{VI_{k,0}}{\widehat{VI}_{k,0}}$.

Next, to identify the vegetation present, we learn a function from data called the pixel classification map to assign class labels to each pixel in a scene:

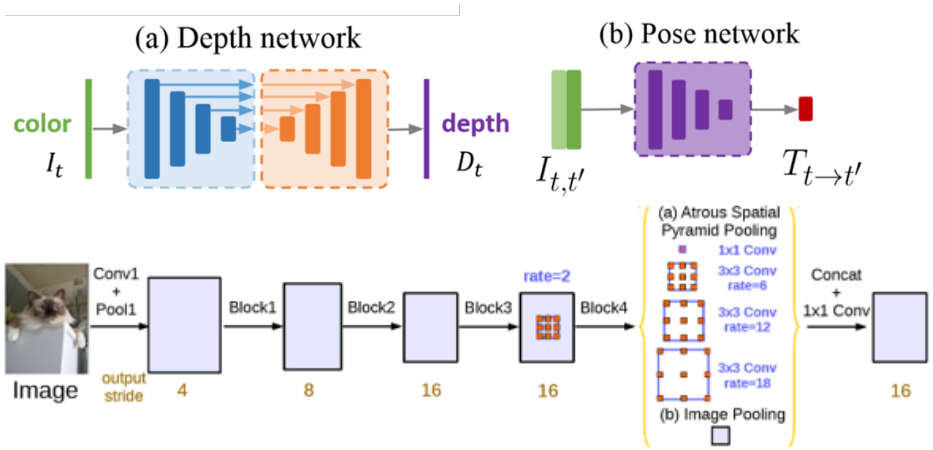


Figure 1: Above: Monodepth2 model overview from [10]. The depth network is a standard U-Net encoder-decoder network [11] with a ResNet encoder [8], and the pose network, which estimates the pairwise image perspective, is also a smaller ResNet. Note that absolute scale can never be inferred. Below: DeepLabv3 model overview from [12], which is based on a ResNet with atrous convolution steps. Future work is to explore combining these networks to share one ResNet for feature extraction efficiency.

$$\hat{Z}(u_i, v_i) -> z_i, z_i \in \mathcal{K} = 1, 2, \ldots K$$

Finally, the vegetation index is estimated by the model as:

$$\widehat{VI}_{k,t} = \sum_{(u_i,v_i)\in\hat{\mathcal{V}}} \mathcal{I}(z_i = k)\hat{D}(u_i, v_i)$$

where we multiply by the depth too to correct for pixel size at different depths in the frame. The functions $\hat{D}(u, v)$ and $\hat{Z}(u, v)$ are estimated as described below.

2.2 Depth estimation network design

The task of estimating the pixel depth map function $\hat{D}(u, v)$ from a single frame is called monocular depth estimation, for which there exists many different methods in the robotic forest navigation domain [9]. Because it is difficult to obtain ground-truth image-depth data, we choose a self-supervised model Monodepth2 [10]. This can be trained using the slight differences in perspective in sequential video frames (Figure 1).

2.3 Semantic segmentation network design

The task of estimating the pixel classification map function $\hat{Z}(u, v)$ is called semantic segmentation, and has been widely used to recognise various kinds of vegetation and natural objects in images and videos [13]. [14],[15] showed that many models are close to state-of-the-art in forest image segmentation, including their own model AdapNet++ as well as DeepLabv3 [12] and U-Net [11]. We choose the DeepLabv3 for the above reason, and as it is lightweight and fast to train using transfer learning (Figure 1).

3 Experiments

We train and test the semantic segmentation model with the Freiburg forest dataset originally used for trail navigation in [14], consisting of videos taken by a robotic vehicle, with some pixel-level segmentation annotations $Z(u, v)$. We consider the categories: vegetation, grass, trail, sky and junk. A sample image is shown in Figure 2a. In total, 230 images are used for training, and 136 for validation.

For the experiments concerned, we do not train the depth estimation model at all, and leave it for future work. Instead, we evaluate and use the out-of-the-box model directly on the Freiburg forest validation and test sets, where the model was trained on the KITTI autonomous driving dataset [16].

We implement the models using the deep learning library PyTorch, allowing us to easily use the pre-trained ResNet weights for TL, and we train on GPU.

4 Results

All numerical results below are summarised in summarised in Table 1. Example outputs of the two networks are also shown in Figure 2. We test our overall model on an unseen sequence of video frames and compare the resulting vegetation index time series $\widehat{VI}_{k,t}$ with hand labels $VI_{k,t}$ (see Figure 3) using the correlation coefficient to surmount the calibration problem.

We evaluate the depth estimation model Monodepth2 by comparing predictions $\hat{D}_{u,v}$ on the Freiburg forest validation set images with ground truth depth masks from [14] using the correlation coefficient and the mean error per image to overcome the exact depth scale inference, averaged across the entire set.

Figure 4 shows the success of the transfer learning for the semantic segmentation model DeepLabv3. After training, we evaluate the model's predictions $\hat{Z}_{u,v}$ on the validation set with respect to the ground truth segmentation labels.

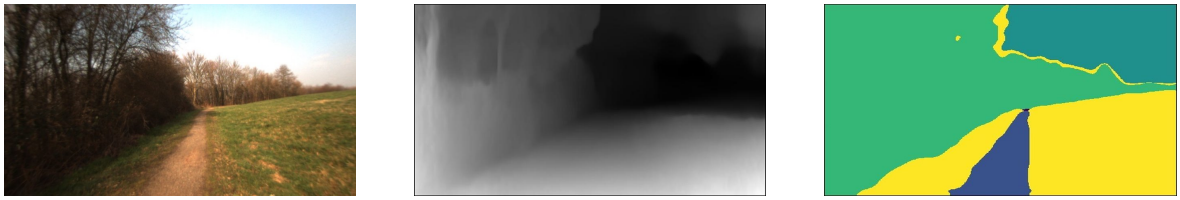
5 Discussion and conclusions

The vegetation index evaluation result is well within human interpretation errors and shows that automated image-based vegetation quantification can replace human methods. After calibration, these vegetation indices can be directly used for vegetation monitoring in forests. Further work is to be done in evaluating the model on a larger dataset with field-sampled vegetation ground truths.

Delving into the model, the depth predictions show low error compared to ground truth even when the model is untrained, and the semantic segmentation training shows high accuracy (in line with results from [14]), showing that these two models are well suited to forest scene understanding.

PMCC$(VI_{vegetation,t}, \widehat{VI}_{vegetation,t})$	90.4%
PMCC$(VI_{grass,t}, \widehat{VI}_{grass,t})$	86.0%
PMCC$(D(u,v), \hat{D}(u,v))$ mean	78.9%
MAPE$(D(u,v), \hat{D}(u,v))$ mean	30.5%
F1$(Z(u,v), \hat{Z}(u,v))$	83.8%

Table 1: All evaluation results of our model. PMCC is the correlation coefficient $\in [-1, 1]$, MAPE is mean absolute percentage error $\in [0, 1]$, and F1 $\in [0, 1]$ is the harmonic mean of the precision and recall.



(a) Original input image. (b) Mono-depth estimation (c) Semantic segmentation

Figure 2: Model outputs for an example forest image.

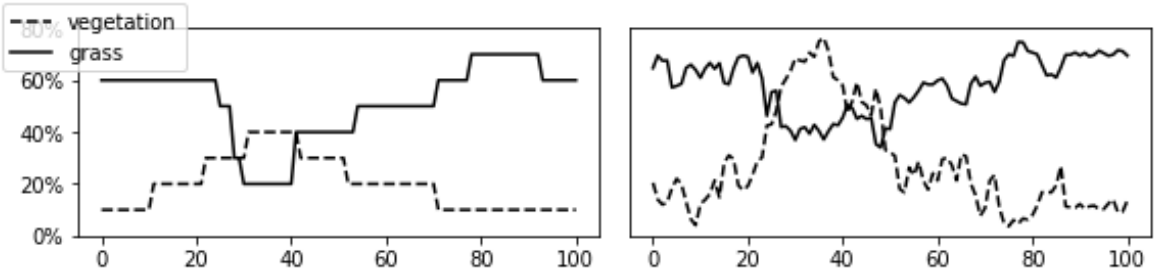


Figure 3: Vegetation index series for an example video travelling on a forest path for quantification of grass and vegetation. Left: ground truth vegetation index labels from manual examination of images. Right: predicted vegetation index time series using our model (y axis is arbitrary).

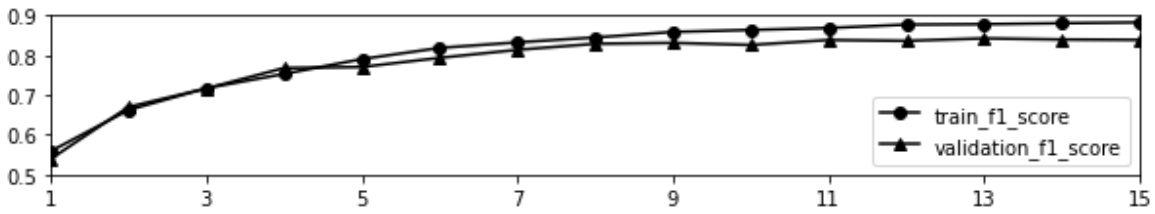


Figure 4: F1-scores vs epoch during transfer learning training of DeepLabv3 model.

References

[1] F. H. Wagner, A. Sanchez, Y. Tarabalka, and R. G. Lotte, "Using the U-net convolutional network to map forest types and disturbance in the Atlantic rainforest with very high resolution images," *Remote Sensing in Ecology and Conservation*.

[2] Iñaki García, "Aphids - Everything you need to know | CANNA UK."

[3] S. Franks, J. G. Masek, and M. G. Turner, "Monitoring forest regrowth following large scale fire using satellite data," *European Journal of Remote Sensing*.

[4] K. Milligan and D. Preston, "Wild Ennerdale Vegetation Monitoring Report 2013," Aug. 2013.

[5] J. S. Rodwell, *National vegetation classification: users' handbook*. Peterborough: Joint Nature Conservation Committee, 2006.

[6] E. M. Wood, A. M. Pidgeon, V. C. Radeloff, and N. S. Keuler, "Image texture as a remotely sensed measure of vegetation structure," *Remote Sensing of Environment*.

[7] H. Nguyen Trong, T. D. Nguyen, and M. Kappas, "Land Cover and Forest Type Classification by Values of Vegetation Indices and Forest Structure of Tropical Lowland Forests in Central Vietnam," *International Journal of Forestry Research*.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," arXiv: 1512.03385.

[9] X. Dong, M. A. Garratt, S. G. Anavatti, and H. A. Abbass, "Towards Real-Time Monocular Depth Estimation for Robotics: A Survey," arXiv: 2111.08600.

[10] C. Godard, O. Mac Aodha, M. Firman, and G. Brostow, "Digging Into Self-Supervised Monocular Depth Estimation," arXiv: 1806.01260.

[11] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," *arXiv:1505.04597 [cs]*, May 2015. arXiv: 1505.04597.

[12] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking Atrous Convolution for Semantic Image Segmentation," arXiv: 1706.05587.

[13] B. Ayhan and C. Kwan, "Tree, Shrub, and Grass Classification Using Only RGB Images," *Remote Sensing*.

[14] A. Valada, R. Mohan, and W. Burgard, "Self-Supervised Model Adaptation for Multimodal Semantic Segmentation," *IJCV*.

[15] A. Valada, G. L. Oliveira, T. Brox, and W. Burgard, "Deep Multispectral Semantic Scene Understanding of Forested Environments using Multimodal Fusion," p. 12.

[16] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *CVPR*.