Low-viewpoint forest depth dataset for sparse rover swarms

Chaoyue Niu, Danesh Tarapore and Klaus-Peter Zauner

Abstract—Rapid progress in embedded computing hardware increasingly enables on-board image processing on small robots. This development opens the path to replacing costly sensors with sophisticated computer vision techniques. A case in point is the prediction of scene depth information from a monocular camera for autonomous navigation. Motivated by the aim to develop a robot swarm suitable for sensing, monitoring, and search applications in forests, we have collected a set of RGB images and corresponding depth maps. Over 100k images were recorded with a custom rig from the perspective of a small ground rover moving through a forest. Taken under different weather and lighting conditions, the images include scenes with grass, bushes, standing and fallen trees, tree branches, leaves, and dirt. In addition GPS, IMU, and wheel encoder data was recorded. From the calibrated, synchronized, aligned and timestamped frames about 9700 image-depth map pairs were selected for sharpness and variety. We provide this dataset to the community to fill a need identified in our own research and hope it will accelerate progress in robots navigating the challenging forest environment. This paper describes our custom hardware and methodology to collect the data, subsequent processing and quality of the data, and how to access it.

I. INTRODUCTION

Forests are ecologically and economically important, affect the local as well as wider climate and are under pressure from changing weather patterns and deceases. They are also a formidable challenge for small all-terrain ground robots. In ongoing research we are aiming at developing a rover platform for this environment. We envisage robot swarms as a useful tool in the efforts to protect, reform, and extend forests.

Robots swarms are teams of robots that coordinate their actions in a distributed fashion to perform an assigned task. A common feature of existing swarms is the underlying assumption that the robots act in close proximity to each other [1]. For real-world outdoor applications over extended areas such a density is neither desirable nor feasible. A dense swarm would not only be very costly, but also highly intrusive to the environment. Recently available technologies in long range radio communication and efficient battery technologies, however, allow for the reconceptualisation of swarms as scalable groups of robots acting jointly over distances up to 1 km. Such robots need to be low cost and high in autonomy.

Reliable terrain analysis is a key requirement for a mobile robot to operate safely in challenging environments, such as in natural outdoor settings. To enable safe autonomous operation of a swarm of robots during exploration, the ability to accurately estimate terrain traversability is critical. The terrain traversability is a complex function of both the terrain characteristics, such as slopes, vegetation, rocks, etc and the robot mobility characteristics [2]. To support this analysis for off-road path planning we are developing a vision system capable of running on small on-board computers. To also keep the cost of sensor low, we are interested in monocular depth estimation [3] to predict local depth from single images or sequences of images from a single moving camera. Aside from optical flow and geometric techniques, machine learning has been applied to achieve this. A number of authors have trained depth estimation models by using deep neural network architectures ([4], [5], [6], [7], [8]).

Most existing outdoor depth map datasets focus on the usage of unmanned driving. The KITTI dataset [9] records street scenes in cities. The Freiburg Forest dataset [10] is mainly to record the whole forest, although it records the distant view of the forest, it lacks of close-range data such as a tree, tree branch, leaf, and bush. Since the dataset has to be manually labeled for image segmentation, it only has 366 images. That is far from enough for deep neural network training. Make-3D dataset ([11], [12]) records many outdoor scenes and some close-up depth data, but it is mainly concentrated on the buildings in the city. We have browsed most of RGB-D Datasets and found out most of them are recorded indoors [13]. Although the current indoor depth map datasets [14] record close-range depth data, it is not taken from the natural outdoor environment. Thus neither of current indoor and outdoor datasets is suitable for our research.

A common feature of these above datasets is that the image is taken at high point of view. Since the robot we are going to use is relatively small and easy to carry, it can be put into backpack. When swarms of robots walk through the forest, they mainly acquired images with low point of view. Thus we need to record depth map with a low point of view. In addition, as the maximum velocity of the swarms of robots is not fast, which is determined by the mobility and mechanical structure of the robot, we should record the depth map at lower walking speed to meet the requirement of velocity. Thus, our dataset requires low point of view and low walking speed of recording.

II. MOBILE SENSOR PLATFORM SETUP

To facilitate efficient data collection we decided to manually move the camera along the path to be recorded record, rather than to record with a robot-mounted camera. The recording rig shown in Figure I was constructed by attaching two incremental photoelectric rotary encoders to an electrical
Fig. 1. **Depth data rig.** Our recording system is equipped with a D435i depth camera, rotary encoder, microcontroller board, and GPS.

Fig. 2. **Sensors setup.** This figure shows the mounting positions and dimensions of each sensing device on recording system. It shows the top view of the recording platform. The black lines represent the wheels and the box. The blue lines represent the sensors we use, which are the depth camera and its base, the rotary encoder, and the micro-control board. The stippling line indicates the part obscured by other components. Red represents the name and dimension of each device. The height of camera from ground is 150mm, height of wheel axis is 50mm. Dimensions are in millimeters.

enclosure box and mounting a 100 mm diameter wheel to each encoder. The encoders were connected to a Micropython enabled ARM board (ItsyBitsy M4 Express, Adafruit) which made the time stamped rotary encoder readings available over a USB connection as a virtual com port. The enclosure is mounted at the end of a telescopic rod of the type used for paint rollers. This allows the user to roll the enclosure on its wheels along the ground by pushing it forward while walking. Inside the enclosure a RealSense D435i depth camera (Intel) was mounted 150 mm above ground with a free field of view in the direction of motion as illustrated in Figure 2.

The camera is comprised of a depth camera and a RGB colour camera with a diagonal field of view of 94° and 77°, respectively. With its global shutter, this camera is well suited to a moving platform, and it also contains an inertial measurement unit (IMU). A laptop computer is connected to the camera, to the USB connection from the rotary encoders and to a GPS receiver (BU-353-S4 SiRF Star IV, US GlobalSat). The endurance of this rig is limited by the battery of the laptop used for recording and for monitoring the camera perspective while walking with the rig.

### III. Forest environment dataset

The data for our forest environment dataset was collected at the Southampton Common, a 1.48 Km² area of woodland, rough grassland and wetlands.

#### A. Data description

Our equipped mobile sensor platform was pushed through the forest in the Common in five separate runs during different times of day and weather conditions to account for the resulting variations in lighting conditions (see Table I and select example forest scenes in Figure 3). For each run in the forest the following data was recorded from the sensor platform: (i) aligned RGB and depth images from the camera; (ii) 6 DoF IMU linear and angular acceleration of the platform (see Figure 1 for axes orientation); (iii) rotary encoder position data; and (iv) GPS location data of the platform.

All the data from the rotary encoder and IMU streams were time synchronized with the recorded images from the camera at 30 frames per second, and recorded at the same rate. The GPS location data was also synchronized with the camera feed, and recorded once per second. Recorded image data was stored lossless in 8-bit PNG file-format at 640 × 480 pixel resolution. Data from the IMU, rotary encoder and GPS sensors were stored in an easy to access CSV flat-file structure. Our full data-set comprises all our forest environment recorded data and metadata information, including over 134K RGB and depth images. A select sample of our forest data-set containing about 9700 RGB and depth images, and the corresponding time synchronized IMU, rotary encoder and GPS sensor data is available online at [https://doi.org/10.5281/zenodo.3693154](https://doi.org/10.5281/zenodo.3693154)

| Dataset index | Weather condition | Time of day | Number of images recorded | Mean luminosity |
|---------------|-------------------|-------------|----------------------------|----------------|
| Run 1         | Partly sunny      | Midday     | 27,333                     | 106            |
| Run 2         | Scattered clouds  | Midday     | 33,194                     | 107            |
| Run 3         | Cloudy, light rain| Evening    | 20,328                     | 80             |
| Run 4         | Sunny             | Afternoon  | 17,499                     | 98             |
| Run 5         | Mostly clear      | Morning    | 36,331                     | 96             |

**TABLE I**

**Forest environment recording conditions.** Our mobile sensor platform was pushed through the forest in five separate runs across different weather conditions and times of day. The data recorded illustrates variations in the lighting conditions. The luminosity or perceived brightness in an image is estimated with the Y channel in the YUV color scheme.
Fig. 3.  **Examples from the forest environment dataset.** A diverse set of scenes in RGB (left), and the aligned depth in grayscale (middle) and color (right), were recorded in the forest. In grayscale, lighter pixels indicate regions further from the camera, and white pixels are out of range. The gradients in depth are better illustrated in color, with brighter colored pixels indicating regions closer to the camera. In both color schemes, black pixels indicate the depth could not be estimated.

IV. QUALITY OF OUR FOREST ENVIRONMENT DATASET

To assess the image quality of the depth data in our forest environment dataset we consider, (i) the *fill rate*, which is the percentage of the depth image containing valid pixels (pixels with an estimated depth value), and (ii) the depth accuracy using ground truth data.

**Fill rate of depth images:** In our depth image data, the fill rate may be affected by the movement of the mobile sensor platform through the forest as well as by the luminosity of the scene, influencing exposure and consequently resulting in motion blur effects. For our analysis, the instantaneous velocity and acceleration of the mobile sensor platform was estimated using the rotatory encoder position data. The luminosity or perceived brightness was estimated from the Y luma channel of the RGB converted to YUV color scheme.

Our analysis suggest a good quality of depth data of the forest environment, with the mobile sensor platform achieving high fill rates (mean ± SD, 0.84 ± 0.11 across all depth images from all five runs). Furthermore, while the

![Velocity and acceleration of the mobile sensor platform.](image)

*Fig. 4. Velocity and acceleration of the mobile sensor platform.** The linear velocity and acceleration of the mobile sensor platform in the forward direction, while being pushed through the forest. Data for the distribution was aggregated across all five runs of the dataset. Instantaneous velocity and acceleration were estimated with the rotatory encoder position sensor.
platform was moved forward through the forest at velocities reaching 0.75 m/s, and with accelerations up to 1.75 m/s² (see Figure 4), there was no clear correlation between fill rate and platform velocity. Similarly, the luminosity of the scene (99 ± 15 across all RGB images from all five runs) was not correlated with the fill rate. This suggests that the invalid pixels in our depth dataset may be consequent to other factors, such as loose foliage encountered by the platform at sub minimum range of the camera (e.g., see leaf foliage in third panel from the top in Figure 3 also see supplementary video).

**Accuracy of depth images:** To estimate the accuracy of our depth images, ground truth depth measurements with a Bosch Zamo Digital distance meter (maximum range 20 m, measuring accuracy ±2 mm) were taken. For ground truth measurements nine points at varying depths in a typical forest scene were considered. The selected points were positioned on the forest floor, fallen leaves, fallen tree branches, and on lower regions of tree trunks close to the ground (see Figure 3). Ground truth measurements were replicated thrice for each of the nine selected points. An offset of 4.2 mm was added to values returned by the ground truth sensor to account for differences in its incident position and that of our depth camera. To account for diffraction of the laser from the ground truth sensor, depth estimates with our depth camera were averaged over 7×7 pixels at the laser spot and replicated twice.

Our results indicate that depth estimated with our mobile sensor platform was close to the ground truth measurements (see Figure 6). Across all sampled points P1 to P9, the mean error registered was less than 4%. The highest error of 12% was for the point P8, which was positioned furthest from the camera.

**V. SUMMARY AND FUTURE WORK**

In this paper, we have proposed a calibrated and synchronized off-road forest depth map dataset recording different obstacles, especially for close-range depth data, such as dirt, tree, tree branch, leaf, and bush. The dataset is recorded under different weather conditions, such as partly sunny, scattered clouds, light rain, sunny and mostly clear. We also measure the quality of the depth map by fill rate and the accuracy of depth data by using a laser emitter. This dataset should be highly useful in the usage of sparse off-road swarms of ground robots. In the future, we are going to train depth estimation model on this dataset.

**REFERENCES**

[1] M. Brambilla, E. Ferrante, M. Birattari, and M. Dorigo, “Swarm robotics: a review from the swarm engineering perspective,” Swarm Intelligence, vol. 7, no. 1, pp. 1–41, 2013.

[2] H. Balta, G. De Cubber, D. Doroftei, Y. Baudoin, and H. Sahli, “Terrain traversability analysis for off-road robots using time-of-flight 3D sensing,” in 7th IARP International Workshop on Robotics for Risky Environment—Extreme Robotics, Saint-Petersburg, Russia, 2013.

[3] A. Bhoi, “Monocular depth estimation: A survey,” arXiv preprint arXiv:1901.09402, 2019.

[4] C. Godard, O. Mac Aodha, and G. J. Brostow, “Unsupervised monocular depth estimation with left-right consistency,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 270-279.
Fig. 6. **Accuracy of the depth image data.** The accuracy of depth images for nine sampled points, P1 to P9. Ground truth measurements were averaged over three replicates. Depth image data was averaged over $7 \times 7$ pixels surrounding the focal point and over two replicates. Points on the diagonal dotted line indicate depth estimates identical to ground truth measurements.

[5] D. Xu, W. Wang, H. Tang, H. Liu, N. Sebe, and E. Ricci, “Structured attention guided convolutional neural fields for monocular depth estimation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3917–3925.

[6] D. Eigen, C. Puhrsch, and R. Fergus, “Depth map prediction from a single image using a multi-scale deep network,” in *Advances in neural information processing systems*, 2014, pp. 2366–2374.

[7] I. Laina, C. Rupprech, V. Belagiannis, F. Tombari, and N. Navab, “Deeper depth prediction with fully convolutional residual networks,” in *2016 Fourth international conference on 3D vision (3DV)*. IEEE, 2016, pp. 239–248.

[8] I. Alhashim and P. Wonka, “High quality monocular depth estimation via transfer learning,” *arXiv preprint arXiv:1812.11941*, 2018.

[9] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, “Vision meets robotics: The kitti dataset,” *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[10] A. Valada, G. Oliveira, T. Brox, and W. Burgard, “Deep multispectral semantic scene understanding of forested environments using multimodal fusion,” in *International Symposium on Experimental Robotics (ISER)*, 2016.

[11] A. Saxena, M. Sun, and A. Y. Ng, “Make3d: Learning 3d scene structure from a single still image,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 5, pp. 824–840, 2008.

[12] ———, “Learning 3-d scene structure from a single still image,” in *2007 IEEE 11th International Conference on Computer Vision*. IEEE, 2007, pp. 1–8.

[13] M. Firman, “Rgbd datasets: Past, present and future,” in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2016, pp. 19–31.

[14] P. K. Nathan Silberman, Derek Hoiem and R. Fergus, “Indoor segmentation and support inference from rgbd images,” in *ECCV*, 2012.