An Off-Road Terrain Dataset Including Images Labeled With Measures of Terrain Roughness

Gabriela Gresenz, Jules White, Douglas C. Schmidt
Department of Computer Science, Vanderbilt University
{gabriela.r.gresenz, jules.white, douglas.c.schmidt}@vanderbilt.edu

***Abstract*—This paper describes the structure and functionality of a dataset designed to enable autonomous vehicles to learn about off-road terrain using a single monocular image. This dataset includes over 12,000 images of off-road terrain and the corresponding sensor data from a global positioning system (GPS), inertial measurement units (IMUs), and a wheel rotation speed sensor. The paper also describes eight roughness metrics derived from IMU z-axis acceleration for labeling the images in our dataset and evaluating which can be learned most effectively.**

The results of our experiments showed that the two metrics learned most effectively were Label 6 (the standard deviation of a 1-second sampling of z-axis acceleration readings directly ahead of the image's timestamp, discretized using k-means clustering with k = 2) and Label 8 (the standard deviation of a 1-second sampling of z-axis acceleration readings directly ahead of the image's timestamp, discretized using k-means clustering with k = 4).

***Index Terms*—Autonomous driving, off-road terrain, terrain roughness, deep learning, dataset**

I. Introduction

Research on autonomous vehicles has been a key focus in recent years, leading to the rise of increasingly autonomous vehicles operating on roadways. For example, in 2019 it was reported that over 1,400 roadway autonomous vehicles across over 80 companies were in the testing stages [1] in the U.S. alone. However, it is also crucial that researchers explore autonomous vehicles in off-road environments. The passenger-carrying autonomous vehicle industry eventually seeks to achieve "Level 5" autonomy, which enables entirely autonomous operation in all conditions [2]. Therefore, if a vehicle ends up in an unexpected situation or on a route containing unmarked or unpaved terrain, the vehicle should be equipped to traverse it safely.

In addition, Autonomous Ground Vehicles (AGVs) are autonomous vehicles designed to complete specific tasks without human supervision [3]. AGVs have applications in search and rescue, mining [4], [5], and planetary exploration [6]. AGV's encounter a wide range of off-road terrain that they must handle autonomously to complete their tasks safely.

Our research provides an extensive off-road terrain dataset including over 12,000 images from a monocular camera and sensor readings from a GPS, IMUs, and a wheel rotation speed sensor. We also derive eight potential measures of terrain roughness from the IMU z-axis acceleration readings that can be used to label images and evaluate their effectiveness.

Past research has performed semantic segmentation for terrain processing [7]–[9] and classification of qualitative terrain type [4], [10]–[13]. To the best of our knowledge, however, past research has not examined classifying terrain roughness as a measure of the vehicle's future kinetics from a single, monocular image, as is provided by the labels in our dataset.

II. Research Challenges and Experimentation Approach

This section first describes the following three challenges involved in preparing an off-road terrain dataset:

- **Lack of relevant off-road terrain data**. Data collection at scale for roadway autonomous vehicles is relatively feasible due to the vast network of roads on which humans driving vehicles equipped with sensors can travel to collect data. In contrast, there is a much smaller network of relevant off-road drivable terrain.
- **Traversing rough off-road terrain can cause an unsteady camera**, which yields images where any drivable terrain ahead is not clearly visible. Moreover, trees surround most of the drivable portions of off-road terrain. Images are therefore susceptible to poor lighting and uneven sunlight that may obstruct the image view.
- **Labeling images of upcoming drivable terrain with a single quantative roughness metric derived from IMU z-axis acceleration readings is hard** because the length of terrain visible in an image may be unknown. As a result, determining the z-axis acceleration readings corresponding to the upcoming drivable terrain in the image is challenging. Validating that a given roughness metric effectively labels the images in the dataset is also hard because a human may not have intuitive insight as to how the vehicle's motion will be affected by traversing this terrain, even though certain visual cues may be indicative of terrain roughness.

To address these research challenges and assist the autonomous vehicle community in making progress for off-road environments, we collected and evaluated the off-road terrain dataset described in this paper. This dataset includes eight potential roughness labeling schemas for images we collected. The remainder of this paper is organized as follows: Section III describes the dataset; Sections IV to VI then address (1) What roughness metric should be used to label images, (2) how do we filter and select the images in our dataset, and (3) which of these potential roughness labeling schemas can be learned most effectively; and Section VII presents concluding remarks and outlines future work.

III. Overview of the Off-Road Terrain Dataset

The dataset is available at kaggle.com/magnumresearchgroup/offroad-terrain-dataset-for-autonomous-vehicles. The data was collected in Percy Warner Park in Nashville, Tennessee, USA via a mountain bike equipped with the following sensors: (1) dual GPS receivers (Garmin 830), (2) dual-high resolution IMU's (Garmin Virb Ultra), (3) a 4k 30fps camera time synchronized to both accelerometers (Garmin Virb Ultra), and (4) a wheel rotation speed sensor (Garmin Bike Speed Sensor 2). Data was collected on five different dates between late July and early October 2020.

The dataset contains two forms of data: sensor data and image frames extracted from video data. The videos were taken by a single monocular camera attached to the bike's handlebars. Image frames were extracted at one second intervals to minimize overlap between frames. The frame rate of our camera was $\sim$29.97 frames per second. It therefore was not possible to extract image frames at exactly one second intervals, so instead we found the image frame most closely corresponding to each second interval.

Image frames are named by their UTC timestamps in seconds and milliseconds (e.g., "1000s100ms"). We generated 12,982 images over nearly 44 miles of off-road terrain. We then excluded all images containing sensitive information (e.g., pictures of other bikers and license plates) and images taken before or after the bike traveled the off-road trail, resulting in 12,730 images included in our public dataset.

These images are not filtered by whether they contain a path, optimal lighting, or sufficient visible terrain for terrain learning so that researchers have access to the entire range of images collected by the vehicle. The size of each image is 3,840 x 2,160 pixels.

Sensor data is store in a format called a "fit file." We converted fit files to comma-separated-value (CSV) files using tools provided by Garmin [14], [15]. We then formatted each CSV to a state-based representation where each row contains the readings at a single timestamp and add UTC timestamps to each reading so the data can be used alongside the image frames. This dataset contains the information described below.

1. Formatted sensor data. There is a folder for each data collection session with the following CSVs.

1) accelerometer_calibrated_split.csv which contains the calibrated and uncalibrated acceleration readings from the accelerometer, taken $\sim$10 ms apart.
2) gyroscope_calibrated_split.csv which contains the calibrated and uncalibrated readings from the gyroscope, taken $\sim$10 ms apart.
3) magnetometer_split.csv, which contains the uncalibrated magnetometer readings.
4) gps.csv, which contains the vehicle's latitute, longitude, altitude, speed, heading, and velocity, taken $\sim$100 ms apart.
5) record.csv, which contains the vehicle's latitude, longitude, distance traveled, speed, and altitude, taken 1 second apart.

2. Roughness labels for images. CSVs of the eight potential roughness labels for the subset of images valid for these labeling schemas are included, as described in Section V.

1) labels_tsm1.csv contains Labels 1–4, as described in Section IV.
2) labels_tsm2.csv contians Labels 5–8, as described in Section IV.

The accelerometer, gyroscope, magnetometer, and GPS CSV files contain system timestamps that represent the relative time since the start of the data collection session, and calculated UTC timestamps. The GPS CSV file also contains a UTC timestamp recorded by the sensor, which may not always align with the calculated UTC timestamp due to sensor lags at certain parts in the forest. Calibrated readings correspond directly to the x-, y-, and z-axes and are in the conventionally understood units. Our data did not contain the calibration factor necessary to calibrate the magnetometer CSV, meaning that the x-, y-, and z-axis readings may not correspond directly to these axes. The speed and velocity readings in the GPS CSV file are GPS estimates and are significantly less accurate than the speed readings in the record CSV file, which are recorded from the wheel rotation speed sensor.

IV. Research Question 1: What Roughness Metric Should be Used to Label Images?

This section explores the derivation of our eight roughness labeling schemas based on the IMU z-axis acceleration readings for labeling images of off-road terrain.

A. Roughness Metric

Many studies have used z-axis acceleration to examine terrain roughness [4], [8], [13], [16]. This measure provides insight about how the vehicle's motion will be affected by traversing the upcoming terrain. Although Stavens et al. [8] standardized their measure of roughness by speed, we used a different approach with our roughness metrics since our data did not exhibit a linear relationship between z-axis acceleration and speed, likely because the speeds of our vehicle (i.e., a bicycle) were significantly slower than the vehicle (i.e., a car) used by Stavens et al.

Our roughness metric takes the standard deviation of a 1 second window of z-axis acceleration readings. This metric is a comprehensive measure of the terrain in the sample and is stable when the sample's mean is nonzero (such as traveling down a hill with increasing acceleration). While our samples could reflect between 1–7 meters since the vehicle's speed was typically between 1–7 m/s, a standard sample size was important to avoid certain samples being more susceptible to outliers than others.

We then determined which 1-second window of z-axis acceleration readings should be used to label each image. The bike traveled along particularly rough terrain, causing the angle and position of the camera to vary. The amount of upcoming terrain and its distance from the vehicle was therefore not constant across all images. For this reason, we decided to examine two terrain sampling approaches:

1) **Terrain Sampling Method 1 (TSM 1)**, which used a 1 second sampling of z-axis acceleration readings centered around the timestamp corresponding to 5 meters ahead of the image.
2) **Terrain Sampling Method 2 (TSM 2)**, which used a 1 second sampling of z-axis acceleration readings directly ahead of the image's timestamp.

We discretized the continuous roughness metric using each of four methods: (1) data visualization (examining the data distribution and z-axis acceleration readings alongside the continuous roughness metric), (2) k-means clustering with k = 2, (3) k-means clustering with k = 3, and (4) k-means clustering with k = 4. These methods will be referred to as original groups, k = 2 groups, k = 3 groups, and k = 4 groups, respectively. In calculating the 1 second sample for TSM 1, only 0.99 seconds of readings were included.

B. Labeling Images

Each image was assigned eight labels, one for each possible combination of the two methods of sampling the terrain and the four methods of discretizing the roughness metric:

1) Label 1: TSM 1, original groups.
2) Label 2: TSM 1, k = 2 groups.
3) Label 3: TSM 1, k = 3 groups.
4) Label 4: TSM 1, k = 4 groups.
5) Label 5: TSM 2, original groups.
6) Label 6: TSM 2, k = 2 groups.
7) Label 7: TSM 2, k = 3 groups.
8) Label 8: TSM 2, k = 4 groups.

V. Research Question 2: How Do We Select and Filter Images in Our Dataset?

We filtered the 12,982 images in our dataset based on sensor and visual criteria, which resulted in 7,070 images valid for Labels 1–4. To compare Labels 1–4 and Labels 5–8, we filtered the images valid for Labels 5–8 to include only images also valid for Labels 1–4, resulting in 7,061 images valid for Labels 5–8. The labeling CSVs included in our dataset do not contain two of the images used in this experiment since these images included other bikers.

We performed sensor validation to confirm that the sensor readings either 5 meters or 3 seconds ahead of each image met the following criteria: (1) the vehicle should not be stopped, (2) sensor readings should be continuous, and (3) the calculated UTC timestamp should be within 1 second of the reported UTC timestamp. In sensor validation for Labels 1–4, we included the third criterion and GPS continuity in case other sensor readings were also affected. We did not consider these criteria for Labels 5–8 because significantly less sensor data was used to calculate these labels.

The next step of validation was visual validation. We confirmed that each image contained a clearly visible path. We trained an image classifier to determine which images met this criteria. We then performed two rounds of manual validation to confirm the classifier's predictions.

TABLE I
Selection Set Performance of Labeling Schemas

	TSM 1		TSM 2		Difference (TSM 2 - TSM 1)	
	Overall accuracy	**Avg class accuracy**	**Overall accuracy**	**Avg class accuracy**	**Overall accuracy**	**Avg class accuracy**
Original groups	34.75%	36.48%	45.48%	47.72%	10.73%	11.24%
k = 2 groups	71.19%	71.33%	73.45%	75.06%	2.26%	3.73%
k = 3 groups	55.65%	46.20%	60.17%	52.30%	4.52%	6.10%
k = 4 groups	45.76%	35.72%	50.00%	46.27%	4.24%	10.55%
				Average	5.44%	7.91%

VI. Research Question 3: Which of the Proposed Labeling Schemas Can Be Learned Most Effectively?

A. Overview

We evaluated how effectively each of labeling schemas presented in Section IV-B could be learned by deep learning models. We trained eight different roughness classifiers, where each uses one of the eight labeling schemas: Model 1 (Label 1), Model 2 (Label 2), Model 3 (Label 3), Model 4 (Label 4), Model 5 (Label 5), Model 6 (Label 6), Model 7 (Label 7), Model 8 (Label 8).

B. Method

We split the data randomly as follows. 80% of the data was set aside for training and validation, 5% of the data was reserved for a "selection set" to select the most effectively learned labeling schemas, and 15% of the data was reserved for the testing set to provide a final evaluation of the selected models. The classes under each labeling schema were skewed, so we balanced the training-validation set by undersampling the majority classes for each labeling schema.

The image classifiers were trained in fastai [17] using transfer learning with ResNet50 [18]. We resized all images to 270 x 480 pixels to speed up training time. We used the fastai default transforms [19], which apply common image transformations to random images in the training set. We excluded the horizontal flip transform so that future work can investigate balancing classes by oversampling images in non-majority classes with a horizontal flip.

The testing and selection sets remained skewed to reflect the real-world data. We therefore evaluated the models on two metrics: overall accuracy and average accuracy by class. Average accuracy by class more heavily accounts for the model's performance on the non-majority classes than overall accuracy, whereas overall accuracy reflects the model's performance on the actual terrain.

C. Analysis of Results

We trained the models and then evaluated their performance on the selection set to determine which two labeling schemas were learned most effectively. The selection set results are shown in Table I.

We first compared TSM 1 with TSM 2 by examining them side-by-side for each method of discretizing the roughness metric. TSM 2 consistently performed better than TSM 1 in both overall accuracy and average accuracy by class for each method of discretizing the roughness metric.

We then compared methods for discretizing the roughness metric, examining only TSM 2. The model trained with k = 2 groups had both the highest accuracy and highest accuracy by class, likely because the classifier had to learn only 2 categories. The jump in performance from the model trained with the k = 2 groups to the model trained with the k = 3 groups was significantly larger than the jump in performance from the model trained with the k = 3 groups to the model trained with the k = 4 groups. Likewise, the k = 4 groups provided more specific information about the upcoming terrain than the k = 3 groups. We therefore determined that the k = 4 groups were preferable compared to the k = 3 groups.

Next, we observed that the increase in overall accuracy of the k = 4 groups compared to the original groups outweighed the much smaller increase in average accuracy by class of the original groups, making the k = 4 groups preferable. We determined that the labels learned most effectively were Labels 6 and 8, the k = 2 and k = 4 groups for discretizing the data with TSM 2.

D. Evaluation On the Test Set

We evaluated the models corresponding to these two labels on the test set. The results are included in Table II.

While we chose to extract images from the videos at 1 second intervals to minimize overlap, some images may have still contained parts of the terrain visible in chronologically consecutive images. To ensure that the roughness classifiers were learning, we further minimized potential overlap with a more intuitive chronological training and testing split: the first 70% of the images in each session were used for training, the next 15% of the images were used for validation, and the final 15% of the images were used for testing.

We trained models using Labels 6 and 8 with this chronological split: Model 9 (Label 6) and Model 10 (Label 8). The results are summarized and compared to Model 6 (Label 6, random split) and Model 8 (Label 8, random split) in Table II.

TABLE II
PERFORMANCE OF MODELS 6, 8, 9, AND 10 ON THEIR RESPECTIVE TEST SETS

Model	Labeling schema	Split	Overall accuracy	Accuracy by class
Model 6	Label 6	Random	69.91%	Class 0: 73.62% Class 1: 58.71% Average: 66.17%
Model 9	Label 6	Chronological	70.19%	Class 0: 72.76% Class 1: 62.11% Average: 67.44%
Model 8	Label 8	Random	51.32%	Class 0: 74.95% Class 1: 45.50% Class 2: 18.46% Class 3: 0.00% Average: 34.73%
Model 10	Label 8	Chronological	52.92%	Class 0: 72.41% Class 1: 45.95% Class 2: 34.22% Class 3: 7.14% Average: 39.93%

Not only did Models 9 and 10 achieve comparable accuracy to Models 6 and 8, respectively, but they surpassed the corresponding models with the random splits in both evaluation metrics. This performance increase may arise since the random split allocated 5% of the data for the selection set, while these images were used in the training and validation sets for the chronological split. The goal of this experiment was to ensure that the models were learning to predict terrain roughness, without memorizing potentially overlapping parts of terrain in chronologically consecutive images, which it succeeded in doing.

VII. CONCLUDING REMARKS

This research presents a dataset for off-road terrain collected by a mountain bike that was instrumented with various sensors. We also include eight schemas for labeling images with a measure of terrain roughness derived from the IMU z-axis acceleration readings.

Based on experiments we conducted, we identified two labeling schemas that were learned most effectively by the corresponding image classifiers: Labels 6 (TSM 2, k = 2 groups) and 8 (TSM 2, k = 4 groups). We demonstrated the performance of image classification models on these two labels, achieving 70.19% overall accuracy and 67.44% average accuracy by class for Label 6 52.92% overall accuracy and 39.93% average accuracy by class for Label 8.

The following are the key lessons we learned from conducting this research:

- **Data for off-road autonomous vehicles can be collected at scale by small, agile, and durable vehicles operated by humans.** By equipping a sturdy mountain bike with a range of sensors, we were able to gather an extensive off-road terrain dataset.
- **We can learn about the future kinetics of the vehicle as a result of upcoming terrain roughness from a single, monocular image**. While many problems in autonomous driving are being approached with expensive vehicles and extensive sensor suites, this research has shown that we can learn about the upcoming terrain with a simple, low-cost sensor set-up. However, an open question is whether these results are sufficient to control autonomous driving algorithms or if significant advancements will be necessary.

Future work could expand our dataset by collecting data from additional sensors (such as LiDAR), in other locations, or with other vehicles (such as a car or a robot). It would also be beneficial to determine a roughness metric or a method of terrain sampling that accounts for all of the visible terrain in an image. Collecting this information may require advanced equipment at the time of data collection to store the amount of terrain visible at any given time.

ACKNOWLEDGMENT

We would like to acknowledge Jiachen Xu, Shiliang Tian, and Acary Ary, who were the team of undergraduate researchers that assisted with this project.

REFERENCES

[1] D. Etherington, "Over 1,400 self-driving vehicles are now in testing by 80+ companies across the us," Jun 2019. [Online]. Available: https://techcrunch.com/2019/06/11/over-1400-self-driving-vehicles-are-now-in-testing-by-80-companies-across-the-u-s/

[2] NHTSA, "Automated vehicles for safety," Jun 2020. [Online]. Available: https://www.nhtsa.gov/technology-innovation/automated-vehicles
[3] S. G. Fernandez, K. Vijayakumar, R. Palanisamy, K. Selvakumar, D. Karthikeyan, D. Selvabharathi, S. Vidyasagar, and V. Kalyanasundhram, "Unmanned and autonomous ground vehicle," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 9, no. 5, p. 4466, 2019.
[4] A. Kurup, S. Kysar, and J. P. Bos, "Svm based sensor fusion for improved terrain classification," *Autonomous Systems: Sensors, Processing, and Security for Vehicles and Infrastructure 2020*, 2020.
[5] M. Mei, J. Chang, Y. Li, Z. Li, X. Li, and W. Lv, "Comparative study of different methods in vibration-based terrain classification for wheeled robots with shock absorbers," *Sensors*, vol. 19, no. 5, p. 1137, 2019.
[6] "Mars 2020 perseverance rover." [Online]. Available: https://mars.nasa.gov/mars2020/
[7] H. Dahlkamp, A. Kaehler, D. Stavens, S. Thrun, and G. Bradski, "Self-supervised monocular road detection in desert terrain," *Robotics: Science and Systems II*, 2006.
[8] D. Stavens and S. Thrun, "A self-supervised terrain roughness estimator for off-road autonomous driving," *arXiv:1206.6872*, 2006.
[9] V. Suryamurthy, V. S. Raghavan, A. Laurenzi, N. G. Tsagarakis, and D. Kanoulas, "Terrain segmentation and roughness estimation using rgb data: Path planning application on the centauro robot," *2019 IEEE-RAS 19th International Conference on Humanoid Robots (Humanoids)*, 2019.
[10] Y. Iwashita, K. Nakashima, A. Stoica, and R. Kurazume, "Tu-net and tdeeplab: Deep learning-based terrain classification robust to illumination changes, combining visible and thermal imagery," *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, 2019.
[11] C. Weiss, H. Tamimi, and A. Zell, "A combination of vision- and vibration-based terrain classification," *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2008.
[12] C. Bai, J. Guo, and H. Zheng, "Three-dimensional vibration-based terrain classification for mobile robots," *IEEE Access*, vol. 7, p. 63485–63492, May 2019.
[13] C. Weiss, H. Frohlich, and A. Zell, "Vibration-based terrain classification using support vector machines," *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2006.
[14] Garmin Developers, "Fitcsvtool." [Online]. Available: https://developer.garmin.com/fit/fitcsvtool/
[15] ——, "Fit protocol." [Online]. Available: https://developer.garmin.com/fit/protocol/
[16] S. Ram, "Semantic segmentation for terrain roughness estimation using data autolabeled with a custom roughness metric," Ph.D. dissertation, Carnegie Mellon University, 2018.
[17] fastai, "fastai v1 documentation." [Online]. Available: https://fastai1.fast.ai/
[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
[19] fastai, "vision.transform." [Online]. Available: https://fastai1.fast.ai/vision.transform.html#Data-augmentation