ATTENTION REGIONS IN TERRAIN ROUGHNESS CLASSIFICATION FOR OFF-ROAD AUTONOMOUS VEHICLES

Gabriela Gresenz, Jules White, Douglas C. Schmidt

Vanderbilt University
Department of Computer Science
1025 16th Avenue South, Nashville, TN 37212

ABSTRACT

This research examines the use of attention regions in terrain roughness prediction for off-road autonomous vehicles. We approach terrain roughness prediction as an image classification problem, labeling images of upcoming drivable terrain with a measure of roughness derived from the z-axis acceleration readings taken by a vehicle's inertial measurement unit. We used deep learning models to learn two roughness labeling schemas and examined whether an attention region surrounding the upcoming drivable terrain could improve the ability of our classifiers to predict terrain roughness.

The results of our experiments demonstrate that in all cases, models utilizing images with no attention region show increased performance compared to those utilizing images with an attention region. These results indicate there is important contextual information in the non-path pixels of the images that assists models in predicting upcoming terrain roughness.

Index Terms— Autonomous vehicles, off-road terrain, terrain roughness, deep learning, attention region

1. INTRODUCTION

Enabling autonomous vehicles to drive safely in off-road settings is important to achieve increasing levels of autonomy in passenger-carrying autonomous vehicles. A key goal of the automotive industry is to achieve "Level 5" autonomy, where vehicles can operate safely in all settings without human supervision [1]. To achieve this level it is crucial that passenger-carrying autonomous vehicles can handle off-road terrain.

Moreover, an Autonomous Ground Vehicle (AGV) is a type of autonomous vehicle designed to perform specific tasks [2]. Applications of AGVs include search and rescue, mining [3, 4], and planetary exploration [5]. AGVs are often unsupervised, so they must be able to handle a wide variety of terrain they encounter (e.g., slowing down for a rough patch).

Much prior work has focused on classifying terrain into discrete groups of terrain type [6, 3, 7, 8, 9]. Prior work has also used semantic segmentation to understand terrain roughness. For example, Dahlkamp et al. [10] used visual and LiDAR imagery to classify pixels of terrain as drivable or undrivable in their robot Stanley [11], which won the DARPA Grand Challenge in 2005 [12]. Stavens and Thrun [13] used LiDAR imagery to classify terrain pixels as either "smooth" or "rugged." Suryamurthy et al. [14] created a unified method for terrain segmentation and roughness prediction.

Many studies [3, 9, 13, 15] involving terrain roughness use z-axis acceleration, which is a measure of a vehicle's vertical acceleration. Labeling upcoming terrain with a measure of z-axis acceleration helps gauge terrain roughness. It also shows how the upcoming terrain will directly affect the vehicle's movement.

For example, Stavens et al. [13] standardized z-axis acceleration with respect to speed. Our initial approach was to account for speed in a similar way, though we did not find a linear relationship between z-axis acceleration and speed in our data. This finding likely occurred because the speeds of the vehicle used by Stavens et al. (i.e., a car) were much faster than the speeds of our vehicle (i.e., a mountain bicycle).

A key question not yet answered by other researchers is whether or not terrain roughness classification for monocular images can be aided by an *attention region*. An attention region is a way to specify the region of the image where the focus of the network should be directed. Prior research has examined classification of qualitative terrain composition and semantic segmentation to understand terrain roughness.

In contrast, our approach examines terrain roughness prediction as a classification problem. In particular, we label monocular images of off-road terrain with a measure of roughness based on the future outputs of the vehicle's inertial measurement unit (IMU). We then examine whether an attention region around the drivable terrain in the images can assist deep learning classifiers in predicting roughness labels.

2. RESEARCH CHALLENGES AND EXPERIMENTATION APPROACH

This section first describes the following challenges involved in training deep learning models to predict terrain roughness:

- **Creating a measure of roughness representing the entirety of the upcoming terrain is hard** because the exact length of terrain visible may be unknown. As a result, it is hard to determine the z-axis acceleration readings corresponding to the visible terrain in the image.
- **A quantitative roughness metric measuring forces experienced by the vehicle may not be intuitive**. While certain objects (such as sticks and rocks) may generally correspond to increased terrain roughness, the z-axis acceleration of the vehicle as a result of traveling over these objects may be unknown. Labels are a measure of acceleration experienced by the vehicle, so it may be hard for humans and/or deep learning classifiers to derive an image's label.
- **It may be hard to learn the roughness label**. Images show a number of other intricate objects (such as large trees and leaves) in addition to the upcoming terrain. A roughness classifier thus may not be able to "focus" on the upcoming drivable terrain.

To help address these challenges, we examined the addition of an attention region surrounding the upcoming drivable terrain in the images. Our goal was to determine whether an attention region can assist a deep learning network in focusing on the information relevant for terrain roughness prediction. We hypothesized that coloring all non-path pixels of the image either black or white would cause the network to focus only on the visible terrain, ignoring potentially distracting features (such as large trees), in learning terrain roughness.

The remainder of this paper is organized as follows: Section 2.1 describes the data collection process and the labeling schemas used in this experiment; Section 3 examines our process for adding an attention region to the images; Section 4 analyzes the results of experiments we performed to determine whether an attention region can assist the network in predicting terrain roughness; and Section 5 presents concluding remarks.

2.1. Overview of Our Data Collection Method

To create the roughness labels in our data set, we needed to collect images and label them with a measure of the z-axis acceleration that was experienced by the vehicle. To collect this data, we used a mountain bike as the vehicle. The mountain bike was instrumented with: (1) Garmin 830 dual GPS receivers, (2) Garmin Virb Ultra dual-high resolution IMUs, (3) a Garmin Virb Ultra 4k 30fps camera time synchronized to both accelerometers, and (4) a Garmin Bike Speed Sensor 2 wheel rotation speed sensor.

Data was collected from mountain biking trails in Percy Warner Park in Nashville, Tennessee, USA during daytime in the late summer and beginning of fall 2020. Each data collection session obtained video data from the camera attached to the handlebar of the mountain bike and sensor data from the bike's various sensors, such as the IMU.

We extracted video frames roughly 1 second apart to minimize overlap. A total of 12,982 images were collected covering 43.9 miles of terrain. The terrain ranged from relatively smooth packed earth to 1–2 foot rocks to large exposed tree roots to muddy creek crossings.

The original images and roughness labels are available at `kaggle.com/magnumresearchgroup/offroad-terrain-dataset-for-autonomous-vehicles`. Likewise, the attention region images are available at `kaggle.com/magnumresearchgroup/offroad-terrain-attention-region-images`.

2.2. Labeling Images

Images were labeled with each of two roughness metrics:

- **Label 1**, which measures the standard deviation of a 1 second sampling of z-axis acceleration readings taken from directly ahead of the image's timestamp, discretized using k-means clustering with k = 2, and
- **Label 2**, which measures the standard deviation of a 1 second sampling of z-axis acceleration readings taken from directly ahead of the image's timestamp, discretized using k-means clustering with k = 4.

The selection of these labeling schemas and k values is described below.

The standard deviation of a 1 second window of z-axis acceleration readings describes the entirety of the terrain in the sample and accounts for cases when the sample's mean is nonzero. A 1 second sample could encompass from around 1–7 meters, depending on the speed of the vehicle (which was typically travelling between 1–7 m/s). However, we decided that it was best to draw from a standard sample size so no one sample was more susceptible to outliers than others.

After images and the associated z-axis acceleration data were collected and the standard deviation of the 1 second window of z-axis acceleration readings ahead of each timestamp was calculated, we discretized this roughness metric using the following two different approaches: (1) the *Label 1 approach* used k-means clustering on the standard deviation data with k = 2 to assign classes to images, and (2) the *Label 2 approach* used k-means clustering with k = 4 to derive the class values for the images.

To select the two methods for discretizing the roughness metric, we evaluated how effectively image labels discretized using data visualization, k-means clustering with k = 2, with k = 3, and with k = 4 could be learned by deep learning classifiers using an intermediary, held out selection set. Based on selection set performance, we determined that k-means clustering with k = 2 and k = 4 were the optimal methods for discretizing the roughness metric.

The labels discretized with the k = 2 groups were learned most effectively, likely because only two classes had to be learned. The labels discretized with the k = 4 groups showed a smaller jump in performance from the k = 3 groups than from the k = 2 groups to the k = 3 groups and provided more specificity than the k = 3 groups. Moreover, the k = 4 groups were learned more effectively than the data visualization groups.

2.3. Filtering Images

The dataset consisted of 12,982 images. However, we performed both sensor and visual validation to filter the images in our dataset. There were 7,061 images that met both sets of criteria described below.

For each image, we confirmed that there was sufficient data to calculate the image's label. If so, we confirmed that in the time window relevant to labeling the image (both 5 meters ahead of the bike and 3 seconds ahead of the bike), the data met certain criteria. For example, we checked whether the relevant sensor readings were continuous and that the bike's speed was nonzero.

An important consideration was ensuring that the images we selected showed a forward-facing image that was unobstructed. To verify this property for each image, we manually checked each image for obstructions and removed any images that did not clearly show a visible path in front of the vehicle.

3. CREATING ATTENTION REGIONS AROUND UPCOMING DRIVABLE TERRAIN

The key question that we investigated in our research was whether or not an attention region would improve terrain roughness classification performance. To create an attention region around the drivable terrain, we first trained a semantic segmentation network to identify the pixels in the image corresponding to the path. We then colored all non-path pixels outside the attention region either black or white.

We experimented with three variations of the images in our dataset, as shown in Figure 1: (1) the original images, (2) images where the attention region was specified by coloring all non-path pixels black (which will be referred to as the "dark attention region"), and (3) images where the attention region was specified by coloring all non-path pixels white (which will be referred to as the "light attention region").

Fig. 1. Example Image, Dark Attention Region, Light Attention Region

To determine if the attention region aided roughness classification, we began by training a semantic segmentation network to classify the pixels in our images as either "path" or "other." We randomly selected 187 images for training and 63 images for testing. We labeled each pixel as either "path" or "other." The segmentation model was evaluated on two metrics: pixel-wise accuracy and Intersection over Union (IoU).

We trained a segmentation network in fastai using the UNet architecture and the default transformations, which apply common image transformations to random images in the training set. We excluded the horizontal flip transform in case our network did not achieve sufficient performance and we needed to augment all images in the training set with a horizontal flip. The network achieved an accuracy of 96.96% and an IoU of 0.9687 on the test set.

The next iteration of the network introduced two methods for improving performance. The training set was expanded to 376 images and included the horizontal flip transformation, which was applied randomly to some training set images. We achieved an accuracy of 97.78% and IoU of 0.9772.

When examining the test set, we noticed that the network struggled to make the correct predictions on images with dark lighting. We therefore trained another version of this network that added more images with dark lighting to the training set. This network produced much more desirable results on images with dark lighting. Moreover, this network was inclined to predict wider—rather than narrower—paths, which we felt was a desirable characteristic for our purposes.

We would prefer that images with an attention region have some non-path regions visible, as opposed to not enough visible path. This network achieved an accuracy of 97.15% and an IoU of 0.9704 on the test set. It is important to note, however, that the results of the previous versions of this network on this same test set influenced the data we added to the training set. These metrics are thus not an indicator of how well the network will generalize.

Subsequent analysis showed that these metrics—along with our visual confirmation in the network's ability to detect the path in images with dark lighting—indicated that this version of the network was ideal for our purposes. We moved forward with this version of the network to apply the attention region to our images.

3.1. Preparing the Dataset

We used the resulting network to save a version of each of the 7,061 images in our dataset where the pixels classified as "path" were tinted dark blue and the pixels classified as "other" were tinted light blue, referred to as the "overlay." Figure 2 shows an image with its correct label and overlay.

Fig. 2. Example Image, Correct Label, Overlay

We then manually validated each of these 7,061 images with overlays and assigned them a score from -2 to 2 based on the following criteria. A score of -2 indicated that many "path" pixels were classified as "other", a score of -1 indicated that a moderate amount of "path" pixels were classified as "other," a score of 0 indicated that most pixels were correctly classified, a score of 1 indicated that a moderate amount of "other" pixels were classified as "path," and a score of 2 indicated that many "other" pixels were classified as "path."

We included all images with a score of 0 or 1 in our dataset for a total of 5,375 images. Images scoring -2 or -1 may not contain enough of the path for terrain learning, while images scoring 2 may include too many non-path pixels, thereby defeating the purpose of an attention region. Images scoring 0 were almost perfectly segmented. Images scoring 1 contained a little extra information, but not enough to detract from the attention region.

4. ATTENTION REGION EXPERIMENTS AND RESULTS

This section describes the results of experiments conducted to assess whether adding an attention region around the upcoming drivable terrain assisted deep learning classifiers in predicting terrain roughness. We trained a roughness classifier for each combination of the two labeling schemas (Labels 1 and 2) and the three attention region variations (original, dark, and light), resulting in six models discussed below.

4.1. Models

The roughness classifiers were implemented with fastai [16] using transfer learning with the ResNet50 [17]. The composition of our dataset with respect to each labeling schema was skewed. We balanced the training and validation sets by undersampling the majority classes.

All images were resized from their original size of 2,160 by 3,840 pixels to 270 by 480 pixels to decrease training time. We used the default transformations in fastai [18], which applies commonly used image transformations to images in the training set selected at random. We excluded the horizontal flip transform because it is a topic for future work, e.g., balancing classes by oversampling non-majority classes with a horizontal flip.

4.2. Analysis of Results

The results are summarized in Table 1. In all cases, the models trained with the original images had greater overall accuracy and average accuracy by class. However, the model utilizing Label 1 and the dark attention region was able to best identify Class 1, despite having the lowest overall accuracy between all models utilizing Label 1. The model utilizing Label 2 and the dark attention region did not predict Class 3 for a single image in the test set.

Table 1. Test Set Performance of Attention Region Models

	Label 1		Label 2	
	Overall accuracy	**Avg class accuracy**	**Overall accuracy**	**Avg class accuracy**
Original images	68.28%	Class 0: 66.92% Class 1: 72.32% Average: 69.62%	55.19%	Class 0: 71.95% Class 1: 52.25% Class 2: 34.94% Class 3: 8.82% Average: 41.99%
Dark attention region	49.55%	Class 0: 39.73% Class 1: 78.57% Average: 59.15%	44.92%	Class 0: 67.42% Class 1: 43.54% Class 2: 9.04% Class 3: 0.00% Average: 30.00%
Light attention region	67.38%	Class 0: 72.36% Class 1: 52.68% Average: 62.52%	42.44%	Class 0: 62.89% Class 1: 15.02% Class 2: 62.05% Class 3: 2.94% Average: 35.72%

The results in Table 1 suggest that while the drivable terrain holds some indicators of upcoming terrain roughness, there are other contextual clues in the non-path regions of the image that assist the classifier in learning terrain roughness.

5. CONCLUDING REMARKS

The study presented in this paper examined whether an attention region surrounding the upcoming drivable terrain in a single monocular image could improve off-road terrain roughness prediction. The following are the key lessons we learned from conducting this research:

- **Images of upcoming drivable terrain can be used to learn about terrain roughness** as a measure of the vehicle's future kinetics. Roughness classifiers trained utilizing images both with and without an attention region showed some ability to predict terrain roughness. Our future work will investigate whether our results can be further improved via custom architectures or balancing classes by oversampling with a horizontal flip.
- **Roughness classifiers utilizing images without an attention region around the upcoming drivable terrain demonstrated increased performance**, indicating that there is important contextual information in the non-path pixels of the image. Our future work will examine other attention region variations, such cropping images around the upcoming drivable terrain or passing the attention region as an additional parameter to avoid losing background information [19]. Moreover, our models utilized images where some non-path pixels could be included in the attention region. We will replicate these experiments with a manual, precise attention region drawn for all images in the dataset.

Acknowledgments

We would like to thank Jiachen Xu, Shiliang Tian, and Acar Ary, who were the undergraduate researchers assisting us with this project.

6. REFERENCES

[1] NHTSA, "Automated vehicles for safety," Jun 2020.

[2] S. George 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, pp. 4466, 2019.

[3] Akhil Kurup, Sam Kysar, and Jeremy P. Bos, "SVM based sensor fusion for improved terrain classification," *Autonomous Systems: Sensors, Processing, and Security for Vehicles and Infrastructure 2020*, 2020.

[4] Mingliang Mei, Ji Chang, Yuling Li, Zerui Li, Xiaochuan Li, and Wenjun Lv, "Comparative study of different methods in vibration-based terrain classification for wheeled robots with shock absorbers," *Sensors*, vol. 19, no. 5, pp. 1137, 2019.

[5] NASA, "Mars 2020 Perseverance Rover," 2020.

[6] Yumi Iwashita, Kazuto Nakashima, Adrian Stoica, and Ryo 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.

[7] Christian Weiss, Hashem Tamimi, and Andreas Zell, "A combination of vision- and vibration-based terrain classification," *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2008.

[8] Chengchao Bai, Jifeng Guo, and Hongxing Zheng, "Three-dimensional vibration-based terrain classification for mobile robots," *IEEE Access*, vol. 7, pp. 63485–63492, May 2019.

[9] Christian Weiss, Holger Frohlich, and Andreas Zell, "Vibration-based terrain classification using support vector machines," *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2006.

[10] Hendrik Dahlkamp, Adrian Kaehler, David Stavens, Sebastian Thrun, and Gary Bradski, "Self-supervised monocular road detection in desert terrain," *Robotics: Science and Systems II*, 2006.

[11] Sebastian Thrun, Mike Montemerlo, Hendrik Dahlkamp, David Stavens, Andrei Aron, James Diebel, Philip Fong, John Gale, Morgan Halpenny, Gabriel Hoffmann, and et al., "Stanley: The robot that won the DARPA Grand Challenge," *Springer Tracts in Advanced Robotics*, p. 1–43, 2007.

[12] DARPA RSS, "The Grand Challenge," .

[13] David Stavens and Sebastian Thrun, "A self-supervised terrain roughness estimator for off-road autonomous driving," *arXiv:1206.6872*, 2006.

[14] Vivekanandan Suryamurthy, Vignesh Sushrutha Raghavan, Arturo Laurenzi, Nikos G. Tsagarakis, and Dimitrios 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.

[15] Shastri Ram, *Semantic Segmentation for Terrain Roughness Estimation Using Data Autolabeled with a Custom Roughness Metric*, Ph.D. thesis, Carnegie Mellon University, 2018.

[16] fastai, "fastai v1 documentation," .

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[18] fastai, "vision.transform," .

[19] Sagi Eppel, "Classifying a specific image region using convolutional nets with an ROI mask as input," *arXiv:1812.00291*, 2018.