Autonomous Terrain Classification with Co- and Self-Training Approach

Kyohei Otsu\textsuperscript{1}, Masahiro Ono\textsuperscript{2}, Thomas J. Fuchs\textsuperscript{3}, Ian Baldwin\textsuperscript{2}, and Takashi Kubota\textsuperscript{4}

Abstract—Identifying terrain type is crucial to safely operating planetary exploration rovers. Vision-based terrain classifiers, which are typically trained by thousands of labeled images using machine learning methods, have proven to be particularly successful. However, since planetary rovers are to boldly go where no one has gone before, training data are usually not available \textit{a priori}; instead, rovers have to quickly learn from their own experiences in an early phase of surface operation. This research addresses the challenge by combining two key ideas. The first idea is to use both onboard imagery and vibration data, and let rovers learn from physical experiences through self-supervised learning. The underlying fact is that visually similar terrain may be disambiguated by mechanical vibrations. The second idea is to employ the co- and self-training approaches. The idea of co-training is to train two classifiers separately for vision and vibration data, and re-train them iteratively on each other’s output. Meanwhile, the self-training approach, applied only to the vision-based classifier, re-trains the classifier on its own output. Both approaches essentially increase the amount of labels, hence enable the terrain classifiers to operate from a sparse training dataset. The proposed approach was validated with a four-wheeled test rover in Mars-analogous terrain, including bedrock, soil, and sand. The co-training setup based on Support Vector Machines with color and Wavelet-based features successfully estimated terrain types with 82\% accuracy with only three labeled images.

Index Terms—Visual Learning, Semantic Scene Understanding, Space Robotics

I. INTRODUCTION

T
HE importance of terrain classification for planetary surface missions is highlighted by the experience of NASA’s Mars rover \textit{Curiosity} in the Hidden Valley. As shown in Fig. 1, the narrow valley has a relatively steep slope on both sides and a floor that is constituted of rippled sand. Initially, these ripples appeared to be safe and non-hazardous terrain. However, it turned out that the rover’s wheels got more and more embedded in deep sand. As a result, the operations team backed up and chose an alternate path over a harder substrate to continue the traverse toward Mount Sharp. In addition, Fig. 2 reminds us of two past events: the \textit{Spirit} rover ended its mission when the wheels got stuck in soft soil, and the wheels of the \textit{Curiosity} rover are significantly damaged when traversing over a terrain with angular embedded rocks. As is evident from these examples, on-board terrain classification is important both for efficiency and safety. However, the current Mars rovers lack the classification capability. As a result, a labor intensive, ground-in-the-loop process is required, where human operators manually identify terrain hazards on rover images and plan a path to avoid potentially hazardous terrain.

Fig. 1: \textit{Curiosity}’s experience in Hidden Valley highlights the need for an on-board terrain classification capability. (Image by NASA/JPL-Caltech)

(a) Wheel stuck in soft soil on MER \textit{Spirit} rover. (b) Punctured wheels on MSL \textit{Curiosity} rover.

Fig. 2: Terrain risks on the Martian surface (Images by NASA/JPL-Caltech)

Manuscript received: September, 1, 2015; Revised December, 21, 2015; Accepted January, 14, 2016.

This paper was recommended for publication by the Associate Editor and Reviewers’ comments. *This research was carried out in part at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration, and in part at the Institute of Space and Aeronautical Science, Japan Aerospace Exploration Agency, under a funding by Japan Society for the Promotion of Science. U.S. Government sponsorship acknowledged.

\textsuperscript{1}K. Otsu is with Department of Electrical Engineering and Information Systems, the University of Tokyo, Bunkyo, Tokyo 113-8656, Japan kyon@ac.jaxa.jp

\textsuperscript{2}M. Ono and I. Baldwin are with the Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA 91109, USA \{Masahiro.Ono, Ian.Baldwin\}@jpl.nasa.gov

\textsuperscript{3}T. Fuchs is with Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA fuchs@caltech.edu

\textsuperscript{4}T. Kubota is with Institute of Space and Aeronautical Science, Japan Aerospace Exploration Agency, Sagamihara, Kanagawa 252-5210, Japan kubota@isas.jaxa.jp

Digital Object Identifier (DOI): 10.1109/LRA.2016.2525040

2377-3766 ©2016 IEEE
There is a solid body of existing work on autonomous terrain classification which will be reviewed in Section II. It is still worth mentioning that a previous work in [1] successfully demonstrated automated vision-based classification on the images acquired by the Engineering Navigation Cameras (NAVCAMs) on Curiosity, which are used for the day-to-day ground operations. However, a drawback of this approach is that it requires a significant volume of training data, typically consisting of thousands of labeled images. A unique challenge for terrain classification on another celestial body is sparse availability of training data. The first visitor to a new world would have no chance to obtain ground images a priori. Even if it is not the first visitor, ground images from different landing sites would not be very useful. In fact, the past Mars rover missions observed significant terrain diversity in different landing sites. Consequently, rovers have to quickly learn from a limited amount of its own experiences in an early phase of surface operation. Furthermore, the available training data are typically biased toward benign terrain since rovers usually avoid hazardous terrain identified by orbiters, resulting in even more sparse availability of images for hazardous terrain.

The contribution of this paper is to develop and demonstrate a new terrain classification method for planetary rovers that requires significantly less training data. Specifically, in the experiment that we performed, the proposed method showed competent performance with only three labeled images. The main idea is three-fold:

- Use vibration measurement, and train two separate classifiers – one for imagery and the other for vibration.
- Employ the co-training approach [2], where two classifiers train each other iteratively.
- Apply the self-training approach [3] to the vision based classifier, where the classifier is trained iteratively by feeding its own outputs as training data.

The first idea has been already examined in existing works [4], [5]. However, their training approaches are uni-directional, where vibration is used to train the vision-based classifier but not vice versa. The proposed approach is distinct from theirs in terms of its bi-directional training procedure. This co-training approach enables the drastic reduction in the amount of training data without sacrificing accuracy. The self-training approach further enhances the accuracy by removing the bias in the training dataset. For example, self-training is used to increase the volume of training data of hazardous terrain, whose data availability is usually sparse unproportionally. Each of the three ideas is not new; however, to the best of our knowledge, the combination of the three are new, and an application to terrain classification for planetary rovers is also novel. The overall technical approach is viewed as a semi-supervised learning scheme, which is summarized in Fig. 3.

An additional contribution of this paper is the hardware demonstration of the proposed approach in a Mars-analogue environment. The tests were performed with an ATRV-Jr test rover at the Mars Yard at Jet Propulsion Laboratory, which has been used by the past and ongoing Mars rover missions including Mars Science Laboratory and Mars Exploration Rovers. Through the experiment, the combined co- and self-training approach successfully demonstrated that it can accurately classify multiple terrain types using only a small quantity of manual labels.

II. RELATED WORKS

Terrain classification provides the semantic information of a given region on terrain, such as symbolic terrain classes, traversability estimates, and numerical parameters of physical properties. A number of researches have been conducted to estimate those properties based on features from exteroceptive sensor data. Specifically, several algorithms tackle challenging planetary environments, where terrain monotonicity poses significant difficulty in distinguishing varying soil composition and size.

Although the existing onboard software of Mars rovers can assess the terrain traversability from purely geometric analysis [6], learning-based approaches become more and more popular in order to be adaptive to the complex outdoor environments where pre-defined rules or heuristics would not work well [7]. Several approaches to the vision-based remote classification were successfully demonstrated in the literature [1], [8], [9]. However, vision sensors are sensitive to environmental conditions such as illumination, and thereby requiring a significant volume of training set covering all expected environmental conditions [8]. There are other techniques which characterize terrain properties based on robust mechanical signatures such as vibration [10], [11] and tactility [12], while the sensing range of proprioceptive sensors is limited to the close proximity of the rover.

Self-supervised learning has been used to extend the sensing range by training exteroceptive sensors (e.g., camera) using proprioceptive measurements (e.g., vibration). Several physical parameters including wheel slips [7], terramechanics soil parameters [4], terrain deformation [13], and power consumption [14] have been estimated with this framework. One of the challenges for this approach is that the model training of the remote classifier depends on the output from the local classifier, which may be in practice contaminated by misclassification. In general, the performance of learning algorithms is degraded if the training data contain wrong labels. To make things worse, the training samples obtained from experiences can be unbalanced to specific classes. For example, rovers tend to follow safe routes rather than challenging routes, resulting in a large training set for “traversable” terrain and a smaller
set for “untraversable” terrain [15]. This is apparently not a pleasant situation as the bias in training data would result in poor estimation accuracy for untraversable dangerous terrain.

In terms of learning algorithms, the previous self-supervised approaches follow the two-step learning framework: The initial step is the training of proprioceptive classifier, which is performed either by supervised (e.g., [4], [14]) or unsupervised (e.g., [5], [7]) manners. The next step is the training of second classifier, based on the prediction result from the first classifier. This automatic supervision omits the process of human labeling, while the challenges discussed above pose difficulty to achieve classification accuracy. This research focuses on semi-supervised learning algorithms which effectively learn both from labeled and unlabeled data. In particular, the co-training approach [2] is applied in place of the conventional two-step learning framework, in order to learn from erroneous predictions by weak supervision between multiple classifiers. Moreover, the self-training approach [3] is applied to the vision-based classifier to utilize samples which lack proprioceptive measurements. These weakly supervised approaches have many successful applications ranging from natural language processing [16] to computer vision [17]. More sophisticated methods are proposed such as [18]. Both approaches work in the same pipeline without modifying existing supervised algorithms.

III. CO-TRAINING AND SELF-TRAINING

The proposed terrain classification method is built upon co-training and self-training approaches, which are categorized as semi-supervised learning algorithms. As shown in Fig. 3, the co-training approach lets the two classifiers supervise each other by iteratively providing a classifier output as training data for the other classifier. The self-training approach is used to perform bootstrap training within the vision-based classifier.

The co-training approach [2] requires multiple independent feature sets called views. Let \( w \) and \( z \) denote the vision and vibration feature vectors respectively. The initial, small, labeled feature set \( L = \{ (w_i, z_i, l_i) \} \) is given for \( l \) being the terrain label. Two separate classifiers \( h_w(\cdot) \) and \( h_z(\cdot) \) build the initial classification models based on the data in \( L \). Later, a set \( U = \{ (w_i, z_i) \} \) of unlabeled examples is obtained from sensor data, and the initial models are applied to the subsampled dataset \( U' \) which contains \( u \) samples randomly selected from \( U \). According to the confidence levels of predictions, \( n_w + n_z \) samples are moved to the labeled dataset \( L \) so that the vision-based classifier increases the samples by up to \( n_w \) and the vibration-based classifier by up to \( n_z \). The classifiers are retrained on this updated dataset \( L \). This iteration is repeated until it reaches the maximum number of iteration \( I \) or loses confident samples.

An intuitive rationale behind co-training is that two classifiers may have complementary knowledge that is informative to the other. For example, a certain terrain that is visually ambiguous may exhibit obvious characteristics in the vibration domain, and vice versa. For this reason, each view should be conditionally independent for co-training to be effective [2].

On the other hand, the self-training approach [3] performs a similar iterative procedure but within a single view. The vision-based classifier is applied to the unlabeled dataset \( V' \) which consists of \( v \) random samples from \( V = \{ (w_j) \} \). Out of predictions, \( m_w \) confident samples are added as pseudo labels to \( L \). It should be noted that this approach relies on an assumption that adjacent samples in the feature domain belong to the same class; otherwise the performance will be drastically degraded since the label mistakes will be reinforced during iterations.

The overall algorithm can be summarized as follows:

1. Given a labeled dataset \( L \) and unlabeled datasets \( U \) and \( V \) (where \( U \cap V = \emptyset \)).
2. Train separate classifiers \( h_w(\cdot) \) and \( h_z(\cdot) \) using the labeled dataset \( L \).
3. Apply \( h_w(\cdot) \) and \( h_z(\cdot) \) to the subsampled dataset \( U' \), and move \( n_w \) and \( n_z \) confident samples to \( L \). If two classifiers disagree on a specific sample, skip it.
4. Apply \( h_w(\cdot) \) to the subsampled dataset \( V' \), and move \( m_w \) confidence samples to \( L \).
5. Repeat from (2).

The next section describes a detailed implementation of these co- and self-training approaches for the rover terrain classification problem.

IV. IMPLEMENTATION

The overview of the proposed system is shown in Fig. 4. The system takes two inputs, namely images from onboard camera and vibration data from IMU, and returns a world map where estimated terrain classes are registered.

A. Feature extraction

1) Visual features: An RGB 3-channel image from onboard camera is used as a source of vision-based terrain classification. A small patch in the image is parameterized with a \( d_w \)-element feature vector \( w = \{ w_i \}^{1:d_w} \). In the present work, a simple color-based feature is employed, but other features such as texture and geometry can be easily adopted. The original RGB image is converted to the L*a*b* color space, where L* represents lightness and a* and b* represent color dimensions.
which we call the characteristic frequencies. The vibration profile is collected at a rate of 100 Hz as the norm of 3-axis accelerations. To reduce the effect of gravity, the short-time averages are subtracted from raw signals. The characteristic frequencies are identified as $\mathcal{F} = (6.95, 8.26) \text{ Hz}$ for this profile. Note that the characteristic frequencies may vary depending on conditions such as the vehicle velocity and steering angles. Vibration data solely on the nominal conditions are used in the classification. This is a reasonable assumption as the planetary rovers tend to be operated constantly at low speeds.

### B. Feature registration

The correspondence between vision and vibration features should be resolved to perform interactive learning. Since remote measurements basically come prior to local measurements, the feature registration requires geometrical matching between stored images and newly arrived vibration data. In contrast to the previous work that finds the correspondences on the terrain patches in the 2.5D space [4], this work registers both features on superpixels in the 2D image plane. This approach is critically beneficial to improving efficiency by avoiding resource-consuming onboard DEM generation.

To find correspondences, the wheel trajectories are tracked on the image planes based on the pose estimation with a visual odometry technique [20]. We only consider local regions for feature registration in order to reduce the effect of drift error. An example of wheel track is shown in Fig. 7, where wheel positions and vibration amplitude at each time are visualized by color. Each superpixel is characterized by the average of corresponding vibration features.

### C. Classifier training

The proposed algorithm is compatible with many supervised classification algorithms. In this paper, the Support Vector Machine (SVM) [21] classifier is used both for $h_w(\cdot)$ and $h_x(\cdot)$, using the radial basis function as the kernel function. Since SVMs do not naturally provide the probability estimates, the Platt scaling [22] is applied which assigns probabilities based on logistic regression to scores.

As studied in the literature, the performance of co- and self-training depends on the selection of the parameters:

- The subsampled pool size $u, v$

---

**Fig. 5:** Visual feature extraction on bedrock and soil image (a), with superimposed superpixels (b), and the mean color of RGB (c) and $a^*b^*$ space (d) for each superpixel.

**Fig. 6:** Wavelet power spectrum at frequencies 2–12 Hz. The power increases from yellow to red.

The mean colors of $a^*$ and $b^*$ inside a patch are computed to compose a two-element feature vector, whereas the $L^*$ dimension is currently not used to reduce the effect of lighting conditions. Although this work uses a relatively simple feature, the co- and self-training approach is not specific to the type of features and hence it can be used with more sophisticated features.

**Fig. 7:** Track projection and registration with superpixels.
The proposed framework. Only three labeled images are given. We employed this simple setting to prove the applicability of the framework. Fig. 8, namely bedrock, soil and sand, and manually assigned labels superimposed on onboard images.

**TABLE I: Nominal parameters used for the experiments.**

| Dataset                  | Labeled data $|\mathcal{L}|$ | Unlabeled data (both) $|\mathcal{U}|$ | Unlabeled data (vision) $|\mathcal{V}|$ | Test data $|T|$ |
|--------------------------|----------------------|----------------------------------|----------------------------------|--------------|
| Co-Training Growth size (vision) $n_w$ | 150                  | 3207                             | 22575                            | 2695         |
| Co-Training Growth size (vibration) $n_x$ | 3                   |                                   |                                  |              |
| Co/Self-Training Growth size (vision) $n_w$ | 3                   |                                   |                                  |              |
| Co/Self-Training Growth size (vibration) $n_x$ | 3                   |                                   |                                  |              |
| Co/Self-Training Growth size (vision) $m_w$ | 1                   |                                   |                                  |              |
| Co/Self-Training Growth size (vibration) $m_x$ | 1                   |                                   |                                  |              |
| Co/Self-Training Growth size (vision) $m_w$ | 4                   |                                   |                                  |              |

- Maximum number of iterations $I$
- The label growth size $n_w$, $n_x$, $m_w$

Note that the optimal parameter settings cannot always be identified in practical applications due to heterogeneous behaviors within different classifiers [16]. We conducted a coarse grid search to find proper parameters, as often employed in the related work. The growth size is set in proportion to the number of data samples so that the data distribution is maintained after each round. One of the findings in our case is that the growth size for self-training $m_w$ is sensitive to the overall accuracy; a large value increases the risk to assign wrong labels, while a small value reduces the adaptivity due to the lack of data for untraversed terrain.

**V. EXPERIMENT**

The experiment to validate the proposed approach has been conducted in an Mars-analogous terrain at JPL called the *Mars Yard*, which consists of multiple terrain types such as sand, soil, and several areas of bedrock simulant. An ATRV-Jr rover was equipped with a stereo camera consisting of Point Grey Flea2 cameras and Kowa 3.5mm wide-angle lenses with a baseline of 0.1m. The camera is placed 0.6m above the ground, facing downward by 20deg. The images are captured with the VGA mode (640x480 resolution) at 2 Hz. The vibration data are collected with an Crossbow IMU400CC rigidly attached to the rover body. The IMU measured three-axis linear acceleration at a rate of 100 Hz.

The traversal tests were performed on three terrain types in Fig. 8, namely bedrock (yellow), soil (blue), and sand (green). We employed this simple setting to prove the applicability of the proposed framework. Only three labeled images are given as the initial training data. The same number of superpixels from each class are randomly selected to form the initial dataset $\mathcal{L}$. The nominal parameters used in the experiments are shown in Table I as well as the size of the given datasets. The learning process is performed offline in the current implementation.

Fig. 9 shows the learning curves for the proposed co- and self-training as well as for a simple co-training without self-training. For comparison, the performances of a naive supervised algorithm using the same SVM classifier and a conventional self-supervised algorithm similar to [4] are also shown in the plot. Since these algorithms are not iterative, their “learning curves” are flat. Qualitatively, the proposed semi-supervised approaches surpassed the non-iterative methods just after a few iterations. Its performance kept improving up to ~100 iterations, while the performance of co-training-only hit the ceiling after ~10 iterations. Quantitatively, the hybrid co- and self-training approach reduced misclassification rate by 7.7% from the supervised algorithm even though the same amount of training data are given. Furthermore, as Fig. 10 presents, the advantage of the proposed approach was still maintained if more data are given for training. These results support the claim that the number of labeled images can be reduced by incorporating vibration data with the proposed co- and self-training approaches.

Next, based on data collected from a single traversal, a terrain model is reconstructed and visualized in Fig. 11. The
test rover, which was initially located on the bedrock simulator, traveled in the same straight line during the test. The generated terrain classification maps are shown in Fig. 12 in comparison with the naive supervised algorithm. The supervised algorithm could not provide correct labels for large parts of soil region. However, the proposed algorithm drastically improved accuracy after the automatic learning process. Remind that this result was achieved with only three training images. This result demonstrates that the proposed terrain classification approach works reliably even if training data is extremely sparse.

VI. CONCLUSION
A novel learning algorithm for vision-based terrain classification was proposed and validated using a testbed rover in a terrestrial Mars analogue. The proposed co- and self-training approaches successfully combine the vision- and vibration-based classifiers, which results in 7.7% error reduction with respect to the naive supervised classifier even with a extremely small number of training data. The proposed method have shown the potential for the terrain classification of a completely unknown world without an excessive amount of human intervention.

REFERENCES