A Self-learning Framework for Statistical Ground Classification using Radar and Monocular Vision

Annalisa Milella
Institute of Intelligent Systems for Automation, National Research Council, via Amendola 122/D-O, 70126, Bari, Italy
e-mail: milella@ba.issia.cnr.it
Giulio Reina
Department of Engineering for Innovation, University of Salento, via Arnesano, 73100, Lecce, Italy
e-mail: giulio.reina@unisalento.it
James Underwood
Australian Centre for Field Robotics (ACFR), School of Aerospace, Mechanical and Mechatronic Engineering, The University of Sydney, NSW 2006, Australia
e-mail: j.underwood@acfr.usyd.edu.au

Received 24 May 2013; accepted 24 January 2014

Reliable terrain analysis is a key requirement for a mobile robot to operate safely in challenging environments, such as in natural outdoor settings. In these contexts, conventional navigation systems that assume *a priori* knowledge of the terrain geometric properties, appearance properties, or both, would most likely fail, due to the high variability of the terrain characteristics and environmental conditions. In this paper, a self-learning framework for ground detection and classification is introduced, where the terrain model is automatically initialized at the beginning of the vehicle's operation and progressively updated online. The proposed approach is of general applicability for a robot's perception purposes, and it can be implemented using a single sensor or combining different sensor modalities. In the context of this paper, two ground classification modules are presented: one based on radar data, and one based on monocular vision and supervised by the radar classifier. Both of them rely on online learning strategies to build a statistical feature-based model of the ground, and both implement a Mahalanobis distance classification approach for ground segmentation in their respective fields of view. In detail, the radar classifier analyzes radar observations to obtain an estimate of the ground surface location based on a set of radar features. The output of the radar classifier serves as well to provide training labels to the visual classification module. Once trained, the vision-based classifier is able to discriminate between ground and nonground regions in the entire field of view of the camera. It can also detect multiple terrain components within the broad ground class. Experimental results, obtained with an unmanned ground vehicle operating in a rural environment, are presented to validate the system. It is shown that the proposed approach is effective in detecting drivable surface, reaching an average classification accuracy of about 80% on the entire video frame with the additional advantage of not requiring human intervention for training or *a priori* assumption on the ground appearance.

1. INTRODUCTION

Future off-road mobile robots will have to explore larger and larger areas, performing difficult tasks with limited human supervision, while preserving, at the same time, their safety. In this respect, the ability to detect a drivable surface is a critical issue. If robotic vehicles could reliably and robustly identify traversable ground in unstructured and unknown environments, the implications would be of great importance for many applications, including exploration and reconnaissance, search and rescue operations, and driving safety. In these contexts, conventional navigation systems that assume *a priori* knowledge of the geometric terrain properties, appearance properties, or both, can exhibit failure cases during periods of highly varying terrain or environmental conditions, such as changing illumination or weather phenomena such as fog and snow. To address these issues, perception systems that use online learning strategies may be beneficial for reliable long-term ground detection.

In this paper, a novel approach for terrain analysis is presented that combines radar sensor with monocular vision in a self-learning scheme. Specifically, a self-taught radar classifier is used to estimate the ground (i.e., the drivable surface) location and to automatically supervise the training of a second classifier based on visual features. Once trained, the visual classifier can segment the entire

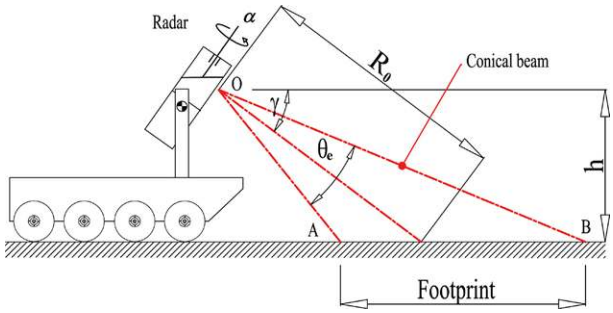


Figure 1. A millimeter-wave radar mounted with a fixed nodding angle can be used to scan for drivable ground in the vicinity of a robot.

video frame into ground and nonground, identifying also ground regions located at a significant distance from the camera. In addition, it can further subdivide ground into subclasses corresponding to different terrain components. By fusing range data provided by the radar and color information produced by the camera, it is finally, possible to obtain a "rich" three-dimensional (3D) map of the environment. Both the radar and the visual classification module rely on a mixture of Gaussians (MOG) model of the ground with K components estimated online via expectation maximization (EM) and a Bayesian information criterion (BIC) -based approach, and they adopt a Mahalanobis distance outlier rejection scheme to estimate the membership likelihood of a given observation to the ground class. The system also features an online updating strategy that allows the ground model to be continuously adapted to changes in the ground characteristics. In previous research by the authors (Reina, Underwood, Brooker, & Durrant-Whyte, 2011b), it was shown that millimeter-wave (MMW) radar technology can be effectively used to improve robot perception by directing a mechanically scanned radar at the front of the vehicle with a constant nodding angle to scan for traversable ground in the vicinity of the vehicle, as shown in the explanatory scheme of Figure 1. In such a configuration with a single sweep of 360 deg, the system is able to survey a finite but relatively large region of the environment, i.e., with a grazing angle γ of about 11 deg, a height h of about 2 m, and an elevation beam width θ_e of about 3 deg, the conical radar beam intersects the ground at a distance of approximately 11.4 m with a footprint varying between 5 and 9 m according to the specific scan angle α. In this work, by combining short-range radar sensing with monocular vision, the ground can be detected at far greater ranges, which is a prerequisite for reliable long-range navigation. Some preliminary parts of this research were presented in Reina, Milella, & Underwood (2012a).

In summary, the following main advantages can be drawn from the proposed system: (a) a unified self-learning framework for ground detection that can be applied using a single sensor or combining different sensor modalities; (b) a radar-vision combination to extend the region of inference from the narrow field of view of the radar to the wide field of view of the camera, and to provide rich 3D environment mapping by fusing range (radar) and color (vision) information; (c) use of MOG for detection of multiple terrain components within the broad class of ground; (d) an adaptive online ground modeling approach that makes the system feasible in long-range and long-duration applications.

The remainder of the paper proceeds as follows. Section 2 describes related research. Section 3 provides an overview of the self-learning framework. In Section 4, the ground modeling and classification strategy is described. Sections 5 and 6 detail the radar and the visual classifier, respectively. In Section 7, the system is validated in field tests performed with an unmanned vehicle. Section 8 concludes this paper.

2. RELATED WORK

Traversable ground detection under all visibility conditions is critical for a mobile robot to navigate safely in unstructured environments. Due to the importance of terrain surface detection and classification for autonomous driving, several approaches have been proposed in the recent literature, using different ground models and sensor combinations.

In general, ground detection methods can be classified into the following categories: deterministic (no learning), supervised, and self-supervised. In deterministic solutions, such as in Huertas, Matthies, & Rankin (2005), Pagnot & Grandjean (1995), and Singh et al. (2000), some features of the terrain including slope, roughness, or discontinuities are analyzed to segment the traversable regions from the obstacles. Some visual cues such as color, shape, and height above the ground have also been employed for segmentation in DeSouza & Kak (2002) and Jocherm, Pomerleau, & Thorpe (1995). However, these techniques assume that the characteristics of obstacles and traversable regions are fixed, and therefore they cannot easily adapt to changing environments. Without learning, such systems are constrained to a limited range of predefined settings.

A number of systems that incorporate supervised learning methods have been proposed, many of them in the automotive field and for structured environments (e.g., in road-following applications). These include ALVINN (Autonomous Land Vehicle in a Neural Network) by Pomerleau (1989), MANIAC (Multiple ALVINN Network in Autonomous Control) by Jocherm et al. (1995), and the system proposed by LeCun, Huang, and Bottou (2004). ALVINN trained a neural network to follow roads and was successfully tested at highway speed in light traffic. MANIAC was also a neural net-based road-following navigation system. LeCun et al. used end-to-end learning to map visual input to steering angles, producing a system that could avoid

obstacles in off-road settings, but it did not have the capability to navigate to a goal or map its surroundings. Many other systems have been recently proposed that include supervised classification (Hong, Chang, Rasmussen, & Shneier, 2002; Manduchi, Castano, Talukder, & Matthies, 2003; Rasmussen, 2002; Reina, Ishigami, Nagatani, & Yoshida, 2010). For instance, in Rasmussen (2002), features from a laser range-finder and color and texture image cues are used to segment ill-structured dirt, gravel, and asphalt roads by training separate neural networks on labeled feature vectors clustered by road type. These systems were trained offline using hand-labeled data, thus limiting the scope of their expertise to environments seen during training. Dima, Vandapel, & Hebert (2004) recognized this problem and proposed using active learning to limit the amount of labeled data in a mobile robot navigation system. Only recently, self-supervised systems have been developed that reduce or eliminate the need for hand-labeled training data, thus gaining flexibility in unknown environments.

With self-supervision, a reliable module that determines traversability can provide labels for input to another classifier. Typically, the self-supervising module consists of a classifier producing reliable results in the short range that are then used to train a second classifier operating on distant scenes. Bootstrapping of the supervising module is performed based on manual training or using some constraint on the ground geometry. For instance, in Brooks & Iagnemma (2012), two proprioceptive terrain classifiers, one based on wheel vibration and one based on estimated traction force, operating in the short range, are used to train an exteroceptive vision-based classifier that identifies instances of terrain classes in the long range. Both supervising modules rely on *a priori* knowledge of the terrain classes in the environment and use either hand-labeled training data or predefined thresholds, thus solving only in part the self-supervision problem. In Vernaza, Taskar, & Lee (2008), data from a stereo camera are used to train a monocular image classifier that segments the scene into obstacles and ground patches, in the submodular Markov random field (MRF) framework. Specifically, first, the largest planar region in the stereo disparity image is sought using a robust least-squares procedure in order to determine ground points in the short range. Then, short-range classification is used as input to the learning algorithm for MRF-based classification in the long range.

LIDAR sensors have been proven to be effective for supervision in several works. A notable example can be found in Dahlkamp, Kaehler, Thrun, & Bradski (2006) using a laser scanner to supervise a monocular camera. Specifically, the laser is employed to scan for flat surface area in the vicinity of the vehicle. This is achieved by looking for height differences within and across map cells and modeling point uncertainties in a temporal Markov chain. The detected area is then projected in the camera image and is used as training data for a computer vision algorithm to learn online a visual model of the road. LIDAR and vision are also combined in a self-supervised framework in Zhou et al. (2012) to detect terrain surfaces in forested environments. The supervising module consists of a LIDAR-based manually trained SVM classifier.

Although vision and LIDAR generally provide useful features for scene classification (Reina, Milella, Halft, & Worst, 2013), they are both affected by weather phenomena or other environmental factors, such as dust. Some LIDARs offer a degree of mitigation such as sensing the "last echo" return, but this fails once the obscurant reaches a sufficient density. Cameras are also highly affected by lighting conditions and are ineffective in the presence of airborne obscurants. Millimeter-wave radar operates at a wavelength that penetrates dust and other visual obscurants. Furthermore, radar can provide information of distributed and multiple targets that appear in a single observation, and the wide beam width allows information to be extracted from a greater footprint of the environment. By contrast, LIDAR systems are generally limited to one target return per emission, although multipeak and last peak-based lasers solve this problem to some extent and are becoming more common. The ability of radar to perceive the environment in low visibility conditions was demonstrated in numerous papers, for example in Peynot, Underwood, & Scheding (2009) and Reina, Underwood, Brooker, & Durrant-Whyte (2011b). Nevertheless, radar has shortcomings as well, since, although the large footprint is advantageous, specularity and multipath effects result in ambiguities in the data, which create challenges for accurate mapping or for feature extraction for classification and scene interpretation tasks. Consequently, to expand the range of possible applications, radar should be combined with other sensors. Video sensors lend themselves very well to this purpose, since, in good visibility conditions, they generally supply high resolution in a suitable range of distances and provide several useful features for classification of different objects present in the scene (Mateus, Avina, & Devy, 2005). Due to the complementary characteristics of radar and vision, it is reasonable to combine them in order to get improved performance.

The fusion of radar and vision has been discussed mostly in the context of driver assistance systems featuring object detection and classification modules (Alessandretti, Broggi, & Cerri, 2007; Ji, Luciw, Weng, & Zeng, 2011; Sole et al., 2004; Wu, Decker, Chang, Camus, and Eledath, 2009). For instance, in Sole et al. (2004), radar and vision independently detect targets of interest, and then a high-level fusion approach is adopted to validate radar targets based on visual data. A radar-vision fusion method for object classification into the category of vehicle or nonvehicle is developed in Ji et al. (2011). It uses radar data to select visual attention windows, which are then assigned a label and processed to extract features to train a multilayer in-place learning network (MILN). In Alessandretti et al. (2007), a vehicle detection system fusing radar and vision

data is proposed. First, radar data are used to locate areas of interest on images. Then, a vehicle search is performed in these areas mainly based on vertical symmetry. A guard rail detection approach and a method to manage overlapping areas are also developed to speed up and improve the performance of the system. The combination of a fixed radar sensor with vision through sensor fusion techniques has also been successfully demonstrated at the DARPA Urban Challenge (Atreya et al., 2007). However, in this research, a forward-facing RADAR system was employed with a narrow horizontal field of view of only 15 deg and specifically tuned for vehicle detection. The sensor was mainly used for position and velocity estimation of vehicles and obstacles directly ahead rather than for general scene interpretation.

Research on radar-vision combination has been developed by the authors in previous work. A first approach was introduced in Milella, Reina, Underwood, & Douillard (2011), which used an expert rule-based radar ground detection approach with manually tuned thresholds to supervise a visual classifier. The latter employed a one-class classification strategy based on a single Gaussian model of the ground to segment each incoming video frame into ground and nonground regions. However, the use of a unimodal Gaussian model poses limitations when multiple terrain types are simultaneously present in the scene. The adoption of machine learning to improve radar classification was proposed in Reina, Milella, & Underwood (2012b). Specifically, a self-trained radar classifier was developed, where the ground model was automatically learned during a bootstrapping stage and continuously updated based on the most recent ground labels to predict ground instances in successive scans. In this paper, the same radar classifier serves as the supervising module for a visual classifier. Both classifiers adopt a self-learning framework, which can be considered to be generally applicable independent of the type of sensors used. To account for multimodality in the feature data set distribution, a MOG is used to model the ground appearance, thus also allowing for detection of terrain subclasses in addition to ground segmentation. In summary, with respect to previous research by the authors concerning radar-vision combination, a novelty of the present work lies in the adoption of MOG for ground modeling, using either radar or visual features. MOGs have been previously adopted in the literature for visual ground modeling [see, for instance, Dahlkamp et al. (2006)], however a different approach is proposed here in which not only the modes but also the number of components of the MOG are estimated online using an EM-BIC algorithm, and this is applied to two sensor modalities. An additional novel contribution is the development of a unified self-learning framework for ground detection that can be applied to a single sensor or combining different sensor modalities. Differently from most approaches in the literature, the proposed system does not require manual training of the supervising module nor *a priori* assumptions on the ground geometry. The framework is, therefore, of general applicability; for example, an embodiment using rich 3D data (i.e., range and color information) obtained by stereovision only was demonstrated in Reina & Milella (2012) for an autonomous agricultural vehicle. Here, radar data and monocular vision are used instead.

3. SELF-LEARNING FRAMEWORK

Hereafter, by "self-learning" we will denote the automatic training of a classification system. The training set can be obtained either via a self-teaching approach, whereby the classifier uses its own predictions to teach itself (i.e., *self-taught learning*), or using the output of another classification module (i.e., *self-supervised learning*) (Zhu, 2005). Self-learning systems eliminate the need for hand-labeled training data, thus gaining flexibility in unknown environments. The burden of hand-labeling data is relieved and the system can robustly adapt to changing environments on-the-fly.

In this paper, first a radar-based classifier using a self-teaching strategy is recalled that labels radar observations into ground and nonground. The training instances for the radar-based classifier are automatically produced using a rolling training set, which is initialized at the beginning of the robot's operation via a bootstrapping approach and progressively updated. Initially, the robot has no knowledge of the relationship between ground characteristics and the ground class. The only underlying assumption to initialize the training set is that the vehicle starts its operation from an area free of obstacles in the radar field of view, so that the radar system initially looks at ground only. Then, features are extracted from the radar data and associated with the ground class. When sufficient data are accumulated, the radar-based ground classifier is trained. This allows the system to predict the presence of ground in successive scans based on past observations. New ground-labeled instances are used to replace old ones, so that the model is progressively updated. The radar classifier serves as well to train a visual classification module that segments each incoming video frame into ground and nonground regions. Specifically, the radar will scan for flat, drivable surface area in the vicinity of the vehicle. Once identified, these radar-labeled ground points are projected in the visual image and they are used to set interest windows from which visual features incorporating the appearance of ground are extracted and employed to build a visual model of the ground. In addition to the ground segmentation task, the visual classifier can also discriminate different terrain components within the broad class of ground. Thus, the radar-based and vision-based modules work in cascade as shown in Figure 2, featuring a self-supervised learning scheme that can perform image segmentation and detect the different local components of the ground.

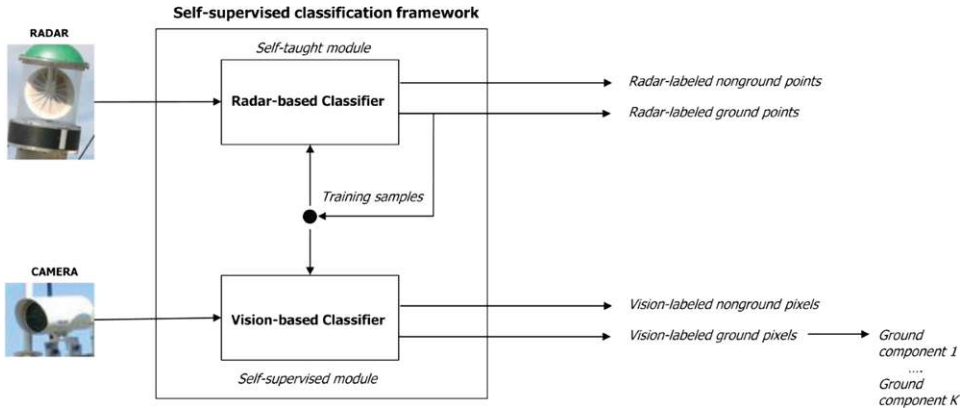


Figure 2. Architecture of the proposed self-learning scheme. The training stage of the visual classifier is supervised by the radar classifier that in turn is self-taught.

4. STATISTICAL GROUND CLASSIFICATION

The accuracy of a ground classifier depends largely on the model adopted for the ground, which is in turn tightly connected with the environmental conditions. In natural settings, conventional ground modeling strategies based on *a priori* assumptions about the geometric or appearance characteristics of the ground are not feasible. In this work, a feature-based representation is adopted for both the radar and the visual classifier. The model of the ground is then built as a MOG, with K Gaussians to be found, defined over the radar and visual feature space, respectively. Based on this model, a Mahalanobis distance classifier is used for the radar and the visual classifier to predict ground in their respective fields of view.

Since only the ground class is defined and modeled, the proposed method can be regarded as a one-class classification approach (Tax, 2001). One-class classification methods are generally useful in two-class classification problems, where one class, referred to as the target class, is relatively well-sampled, while the other class, referred to as the outlier class, is relatively undersampled or is difficult to model. Typically, the objective of a one-class classifier is to construct a decision boundary that separates the instances of the target class from all other possible objects. In our case, ground samples constitute the target class, while nonground samples (i.e., obstacles) are regarded as the outlier class. It is worth noting that, in principle, both ground and nonground samples may be obtained from the radar module to train other types of classifiers (e.g., discriminative two-class classifiers), without affecting the idea behind the self-learning framework. Nevertheless, in open rural environments, nonground samples are typically sparse; in addition, the variation of all possible nonground classes is unlimited. That makes it difficult to model the nonground class, whereas, although it changes geographically and over time, the ground class is generally less variable than random objects. Furthermore, our objective is to build a visual model of the ground. Therefore, it is reasonable to formulate the problem as a distribution modeling one, where the distribution to estimate is the visual appearance of the ground.

In the remainder of this section, the general ground modeling and classification approach is introduced, while details about the radar and the visual classifier are provided in Sections 5 and 6, respectively.

4.1. Ground Modeling

Our basic model for ground representation is a MOG, where each component describes a local ground component. MOG models have been extensively used in the literature for clustering, since each cluster can be easily represented in a compact form using three main parameters: mean vector, covariance matrix, and number of members of the cluster. Expectation maximization (EM) is a common method to estimate the parameters of a MOG, however it requires *a priori* knowledge of the number of components K of the Gaussian mixture. The choice of the optimal number of Gaussian components is a critical issue, especially for online estimation problems, such as in terrain modeling applications. On the one hand, a small number of components may be unable to correctly identify nonhomogeneous ground regions; on the other hand, a high value of K could lead to an overfitting of the training set with a loss of generalization power of the classifier. Furthermore, in autonomous exploration, *a priori* knowledge of K would entail that the number of habitats be known prior to training, which is not generally the case.

In this work, EM and the Bayesian information criterion (BIC) are used to fit the data using a MOG model, and estimate, at the same time, the optimal number of Gaussian components. The BIC (Schwarz, 1978) has been widely adopted to assess the fit of a model and to compare

competing models, based on a measure of information. The BIC statistic is computed as

$$\mathrm{BIC} = -2 \cdot \log L + f \log n, \tag{1}$$

where f is the number of free parameters (which, in turn, depends on the number of clusters K and on the number of feature variables m), L is the maximum likelihood achievable by the model, and n is the sample size. Being defined in such a way, the BIC aims to balance the increase in likelihood due to the use of a higher number of parameters, by introducing a penalty term that grows as long as the number of parameters is increased. Based on BIC, two models can be compared for model selection purposes, with the model having the smaller value of the BIC statistic being preferred.

In this investigation, the optimal MOG model is found, based on a recursive procedure that starts with a single-component assumption and iteratively applies EM with a growing number of components up to a predefined maximum value. Then, the best fitting model is determined to be the one that leads to the smallest value of the BIC statistic. Specifically, let X_t be an $n \times m$ data table representing a sample of x_j vectors with $j = 1, 2, \ldots, n$, each characterized by m traits: $X_t = \{x_1, \ldots, x_n\}$. These vectors constitute the training set at a given time t to construct the ground model as a mixture of multivariate Gaussians with K components, $G_t^K = \{g_1, g_2, \ldots, g_K\}$, where each component g_i, $i = 1, 2, \ldots, K$, is represented by $g_i = (\overline{x}_i, S_i, n_i)$, where $\overline{x}_i$ is the mean value, S_i is the covariance matrix, and n_i is the number of feature vectors belonging to component g_i. To estimate G_t^K, a single Gaussian distribution is initially fit to the data (i.e., $K = 1$ is assumed); then, the number of Gaussian components is incremented one unit at a time until a maximum number of components $K_{\max}$ is reached. It is worth noting that, in the proposed approach, the training set is built upon a rolling window reflecting the appearance of a small portion of the environment that is successively encountered by the robot along its path. It thus represents a continuously updated picture of the local properties of the ground and, as such, it is reasonable to expect K to vary in a limited range (e.g., $K_{\max} = 5$). An additional stopping criterion is also employed based on the mixing proportions of the components in the MOG: if the minimum mixing proportion of a component is less than a threshold, then iteration is stopped and only the MOGs estimated up to the previous iteration are retained. At each iteration, the BIC statistic associated with the model G_t^K is computed. Finally, the MOG with the smallest BIC $G_t^{K^*}$, i.e., the model corresponding to the highest Bayesian posterior probability, is selected as the best fitting model according to the BIC approach. To verify whether this model actually represents a significant improvement with respect to a model with a lower number of Gaussian components, the absolute difference between its BIC and the BIC associated with the model with $(K^* - 1)$ components is computed. Following Raftery (1995), if this difference is greater than 10, then there is very strong evidence in favor of the model with K^* components and $G_t^{K^*}$ is retained as the best model, otherwise the model with a number of clusters $(K^* - 1)$ is preferred to reduce complexity.

It should be noted that, in order to account for ground changes during the vehicle travel, the EM-BIC MOG fitting algorithm is applied on a frame-by-frame basis such that the ground model is recomputed with the new acquired ground-labeled observations.

4.2. Ground Classification

Given a new observation z, where z is either a radar feature vector in the radar classifier or a visual feature vector in the visual classifier (see Sections 5.1 and 6.1 for a description of the radar and visual features, respectively), the classification step is aimed at assessing whether the observation is an instance of ground or not. A Mahalanobis distance-based approach is adopted whereby the Mahalanobis distance and its distribution are employed to predict if a pattern has an extremely low probability of belonging to ground and may be suspected to be an outlier. In detail, the algorithm proceeds as follows. First, the squared Mahalanobis distance of the feature vector z with respect to each component of the current ground model G_t^K is computed as

$$d_i^2 = (z - \overline{x}_i) S_i^{-1} (z - \overline{x}_i)^t, \tag{2}$$

where $\overline{x}_i$ is the mean value and S_i is the covariance matrix of the ith component for $i = 1, 2, \ldots, K$, K being the number of available terrain subclasses. Then, the minimum squared Mahalanobis distance $d_{\min}^2 = \min\{d_1^2, d_2^2, \ldots, d_K^2\}$ (i.e., the distance of z from the closest ground subclass) is found, and is compared with a cutoff threshold for classification.

Under the assumption of normality of the feature vector distributions, it can be shown that the squared Mahalanobis distance is asymptotically distributed as the m degrees of freedom chi-square distribution χ_m^2 (Mardia, Kent, & Bibby, 1979), with m being the number of feature variables. Let α denote a constant probability level: $0 < \alpha < 1$. Let $\chi_{m;\alpha}^2$ denote the appropriate quantile of the distribution. Then, it holds that

$$p\left(d_{\min}^2 \geq \chi_{m;\alpha}^2\right) = 1 - \alpha, \tag{3}$$

which means that values of $d_{\min}^2$ greater than (or equal to) $\chi_{m;\alpha}^2$ appear with a probability equal to $(1 - \alpha)$. Hence, any patch with minimum Mahalanobis distance $d_{\min}^2$ satisfying the inequality

$$d_{\min}^2 \geq \chi_{m;\alpha}^2 \tag{4}$$

may be suspected to be an outlier at significance level $(1 - \alpha)$. Otherwise it will be labeled as a ground. It should be noted that, once the significance level, i.e., the admitted probability of classifying a patch as nonground when it is actually a ground, has been fixed, the classification threshold is set accordingly.

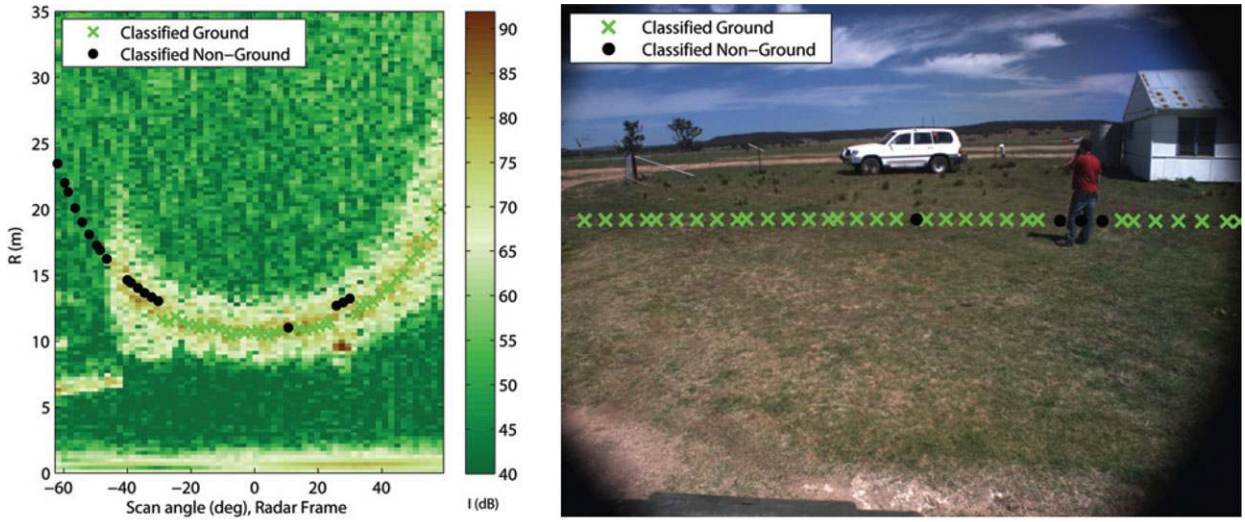


Figure 3. (a) Radar image with overlaid results obtained from the radar-based classifier. Green cross: ground-labeled return. Black dot: nonground labeled return. (b) Classification results projected over the colocated camera image.

5. RADAR CLASSIFIER

A self-trained classifier using radar features was previously presented by the authors. Here, it is shown how the same radar-based classifier can be considered as a part of a more general self-learning framework for reliable ground detection. In this section, the classifier is briefly reviewed; the reader is referred to Reina et al. (2012b) for more details.

5.1. Radar Features

The radar is assumed to be mounted on a frame attached to the vehicle's body and tilted forward (see Figure 1). In such a configuration, a single sensor sweep generates a bidimensional intensity graph (i.e., radar image or B-scope), as a result of the convolution of the scene with the radar beam. A sample radar output is shown in Figure 3(a), referring to the scenario of Figure 3(b). The abscissas in Figure 3(a) represent the scan angle, whereas the ordinates represent the range measured by the sensor. The radar image can be thought of as composed of a foreground and a background. The background is produced by the ground echo, i.e., the intensity return scattered back from the portion of terrain that is illuminated by the sensor beam. Radar observations belonging to the background show a wide pulse produced by the high incident angle to the surface. Conversely, obstacles present in the foreground appear as high-intensity narrow pulses. It was shown that the power return of the ground echo for a single scan angle can be expressed as a function of the range R (Reina et al., 2011b),

$$P_r(R) = k\frac{G(R, R_0)^2}{\cos\gamma}, \tag{5}$$

where k is a constant quantity, R_0 is the slant range, G is the antenna gain (usually modeled as Gaussian), and γ is the grazing angle, as previously explained in Figure 1. It should be noted that the radar ground echo refers to the intensity return scattered back from the portion of terrain that is illuminated by the conical beam of the sensor, usually referred to as the footprint. For our system, the footprint length varies as a function of the scan angle between 5 and 9 m, thus limiting the radar resolution for segmentation purposes.

By extracting and processing the portion of the radar signal pertaining to the background, a set of radar intensity and shape features representative of the ground class can be obtained. To this aim, the theoretical ground model (5) can be fitted to radar data under the assumption that a good match between the model and the experimental data attests to a high likelihood of ground. Conversely, a poor goodness of fit suggests low likelihood due, for example, to the presence of an obstacle, or highly irregular or occluded terrain. One should note that $P_r(R)$ is a parametric function defined by the parameters R_0 and k. k can be interpreted as the power return at the slant range R_0, and it is chosen as the first feature defining the ground appearance. Both parameters R_0 and k can be estimated by data fitting for the given scan angle. Output from the fitting process includes the updated parameters R_0 and k as well as an estimate of the goodness of fit. The coefficient of efficiency was found

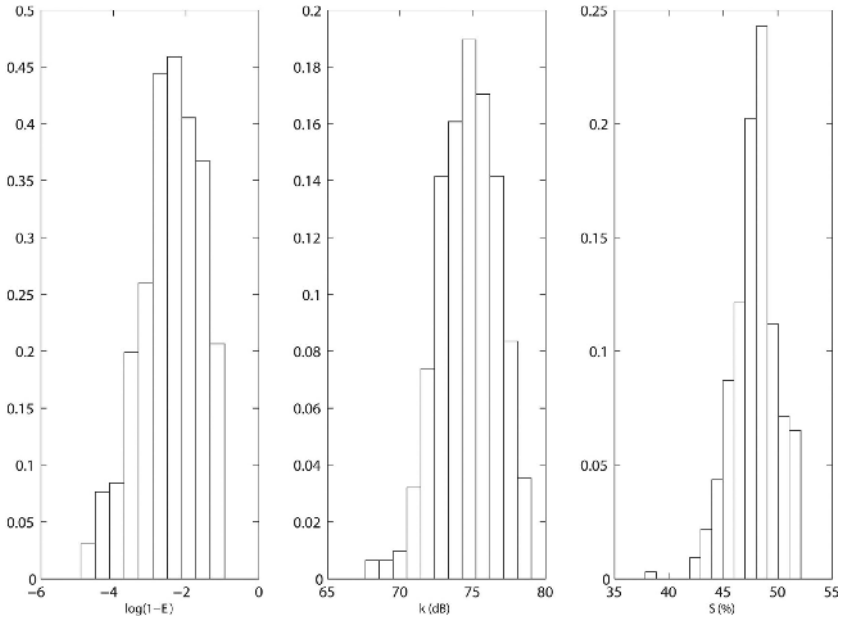


Figure 4. Normalized histograms of the distribution of the radar features for a training set referring to mixed terrain (sandy and grass). All three histograms exhibit an approximately unimodal distribution, which can be reasonably modeled with a single Gaussian.

to be well suited for this application, and it is chosen as the second feature of our model,

$$E = 1 - \frac{\sum_{i=1}^{N_s}(t_i - y_i)^2}{\sum_{i=1}^{N_s}(t_i - \bar{t})^2}, \tag{6}$$

t_i being the data point, $\bar{t}$ the mean of the observations, y_i the output from the regression model, and N_s the number of points. In addition to k and E, a third feature is also used, i.e., the shape factor S defined as

$$S = \left| \frac{I_0 - I_{\text{end}}}{I_0} \right|, \tag{7}$$

where I_0 and I_{end} are the initial and final intensity value of the ground echo. Our hypothesis is that a normal ground echo should have similar initial and final intensities due to the physical interaction between the radar emission and the ground. A high value of S indicates a discrepancy and suggests low confidence that the signal is an actual ground echo.

In summary, three main features define the ground model: the intensity k associated with the slant range, the goodness of fit E, and the shape factor S. This set of features is used to model the ground and classify radar returns as ground or nonground, according to the classification scheme described in Section 4. The ground class corresponds to returns from the terrain, whereas the nonground class corresponds to all other returns, including sensor misreading and reflections from above-ground objects (i.e., obstacles) or from occluded areas. As an example, the results obtained from the radar-based classifier are overlaid over the radar and visual image of Figure 3. Ground labels are denoted by green crosses, whereas black dots mark nonground labels.

For the same scene, it is interesting to look at the distribution of the radar features used in the current ground model, shown in Figure 4. Although multiple terrain types (sandy+grass) are simultaneously present in the training set, all three histograms exhibit an approximately unimodal distribution, which can be reasonably modeled with a single multivariate Gaussian (i.e., $K_{\max} = 1$, as explained in Section 4.1).

5.2. Self-training Approach

As explained in Section 3, an adaptive self-taught method is proposed that allows the ground model to be constructed and updated online following a multiframe approach without any *a priori* information. Specifically, at the beginning of the robot's operation, the training set for the radar classifier is initialized under the assumption that the vehicle starts from an area free of obstacles in the radar field of view, so that the radar "looks" at ground only. Then, the ground model is continuously updated as the vehicle moves: new ground feature vectors labeled in the most recent

acquisitions are incorporated, replacing an equal number of the oldest ground instances. The size of the rolling window is kept constant. Let $Z_{t+1} = \{z_1, z_2, \ldots, z_l\}$ denote the set of l ground-labeled cells classified at time $t+1$. Then the training set for the next acquisition scan is obtained as

$$X_{t+1} = \{(x_{l+1}, \ldots, x_n), Z_{t+1}\}. \quad (8)$$

Once the radar classifier has been trained, it can predict ground points in the subsequent scan. Radar-ground labeled instances are then used within the self-supervised scheme of Figure 2 to provide training instances to the visual classifier.

6. VISUAL CLASSIFIER

A visual classifier using images produced by a monocular camera is described. It relies on texture and color features to describe the appearance of ground. Ground can be detected in the entire video frame, and also at a significant distance from the camera, thus providing long-range information. In addition, ground subclasses can be identified for terrain typing. The learning phase for the visual classifier is supervised by the radar, which provides ground labels. In this section, first the adopted visual features are presented, and then the radar-camera integration is described.

6.1. Visual Features

A vast body of literature exists on investigations of the use of visual features for the task of terrain classification, using either color, texture, or a combination of both (Permuter, Francos, & Jermyn, 2006; Sung, Kwak, Kim, & Lyou, 2008). Approaches using interest point descriptors (e.g., SURF) have also been proposed in recent works (Filitchkin & Byl, 2012; Khan, Masselli, & Zell, 2012). In this research, the visual appearance of the ground is represented in terms of color and textural information. Color data are available as red, green, and blue (RGB) intensities. Previous research in the literature has shown that raw RGB space can be inadequate for classification purposes in outdoor navigation contexts due to its sensitivity to lighting variations and nonuniform illumination (Sofman, Lin, Bagnell, Vandapel, & Stentz, 2006). Therefore, it is generally useful to map the colors from the RGB space to a more suitable one. Here, we adopt the rg chromaticity space. It consists of a two-dimensional color space, with no intensity information (Balkenius & Johansson, 2007). In this space, each pixel is represented by the contribution of the red (r) and green (g) components, which are derived from the RGB color space as

$$r = \frac{R}{R+G+B}, \quad (9)$$

$$g = \frac{G}{R+G+B}. \quad (10)$$

Since all the components are normalized, it is also possible to compute the blue contribution, if necessary, as $b = 1 - (r+g)$. Although the rg chromaticity space contains less information than RGB or HSV color spaces, it has several useful properties for computer vision applications, and it has been demonstrated to perform similarly to other more complex normalized color representations (e.g., *c1c2c3*) (Gevers, Weijer, & Stokman, 2006; Reina & Milella, 2012). The main advantage of this space is that changes in light intensity will not change the basic color of the objects in the scene.

Textural features account for the local spatial variation in intensity in the image. Several texture descriptors have been used in the literature, including Gabor filters, wavelets, and local energy methods (Brooks & Iagnemma, 2008; Reed & du Buf, 1993). In this work, we use an approach based on the gray-level co-occurrence matrix (GLCM), a second-order texture measure. Haralick, Shanmugam, and Dinstein (1973) proposed 14 statistical features that can be extracted from a GLCM to estimate the similarity between different gray-level co-occurrence matrices. Among these features, two of the most relevant are energy and contrast (Cossu, 1988). Energy measures the textural uniformity of an image and reaches its highest value when the gray-level distribution has either a constant or a periodic form. Contrast describes the amount of local variations in an image. These two parameters have been recognized as being highly significant to discriminate between different textural patterns (Baraldi & Parmiggiani, 1995). Energy and contrast were used in our implementation for ground characterization.

In conclusion, a four-dimensional feature vector resulting from the concatenation of two scalar color descriptors and two scalar textural descriptors was adopted. One should note that more complex visual descriptors can also be used without altering the rest of the algorithm.

6.2. Radar-based Training

The radar-based classifier detects and ranges a set of points in radar-centered coordinates, which we regard as good estimates of ground and we use to automatically train the vision-based classifier. With reference to the running example of Figure 3, radar-labeled ground points are first projected over the camera image. Then, for each projected point, an attention window is set. Specifically, based on the available calibration data, interest windows are built as follows. For each labeled radar point, the corners of a squared ground portion of 0.30 m $\times$ 0.30 m centered on that point are projected on the visual image using the perspective transformation; then, the window is defined as the bounding box of the projected corners. Due to the perspective effect, the bounding boxes result in rectangular windows of varying size of about 35 $\times$ 7 pixels (see Figure 5). Successively, the image patches associated with the windows are processed

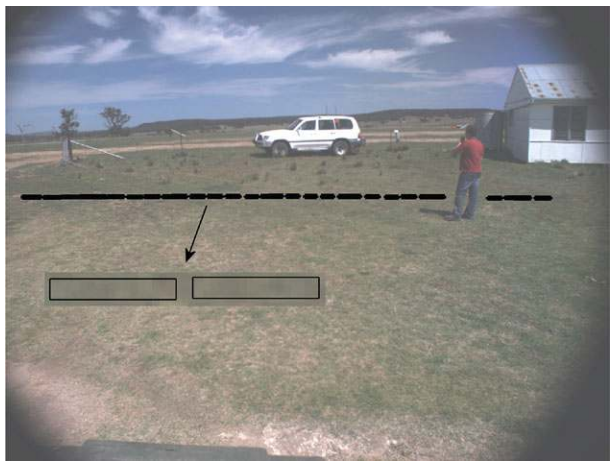


Figure 5. Projections of radar-labeled ground returns in the colocated camera image with a close-up of some attention windows. Visual features extracted from these windows are included in the training set to build a visual model of the ground.

to extract visual features and build a training set for the concept of ground. It is worth noting that, in order to update the ground class for visual scene classification during the vehicle motion, the visual ground model is continuously updated, always using the ground feature vectors extracted by the most recent radar predictions.

Once the classifier has been trained, the vision algorithm can be extended to the entire field of view of the camera. A block-based segmentation method is used to reduce the segmentation processing time at the cost of a lower resolution. Specifically, the image is divided in small patches, and for each subimage the feature vector is computed and compared with the current ground model for classification, as explained in Section 4.2.

7. RESULTS

The proposed approach for ground segmentation comprises two main steps: ground detection from radar data and self-supervised visual classification based on radar labeling. The performance of the radar-based classifier in detecting ground was previously evaluated in Reina et al. (2012b) through extensive field testing. The combined radar-vision system is demonstrated in the field in this section. First, the experimental setup is described in Section 7.1. Then, the influence of the system's parameters is evaluated through a sensitivity analysis in Section 7.2. The proposed adaptive strategy is compared to a static approach and a batch training approach in Sections 7.3 and 7.4, respectively. Finally, the overall performance of the system in detecting and mapping ground is evaluated in Section 7.5.

Figure 6. The CORD UGV used in this research along with its sensor suite.

Table I. Specifications of the custom-built radar system.

	Max. Range	Raw range resolution	Horizontal FOV	Instantaneous FOV	Scan angle Rate
Radar	120 m	0.25 m	360 deg	3 × 3 deg	3.0 Hz

Table II. Specifications of the camera.

	Image Pixel Dimensions	Resolution	Frame rate
Camera	1360 × 1024	72 × 72 ppi	10 fps

7.1. Experimental Setup

Experimental validation was performed using the CAS Outdoor Research Demonstrator (CORD), shown in Figure 6, operating in a rural environment at the University of Sydney's test facility near Marulan, NSW, Australia. The CORD test bed is an eight-wheel skid-steering all-terrain unmanned vehicle, and its onboard sensor suite includes a 95 GHz frequency-modulated continuous-wave (FMCW) radar, custom-built at the Australian Center for Field Robotics (ACFR) for environment imaging (Brooker et al., 2006), and a Prosilica Gigabit Ethernet camera, pointing down (a few degrees of pitch). The main technical properties of the two sensors are illustrated in Tables I and II for the radar and the camera, respectively. The robot was also equipped with other sensors, including four 2D SICK laser range scanners, a thermal infrared camera, and a RTK DGPS/INS unit providing accurate position and tilt estimation of the vehicle.

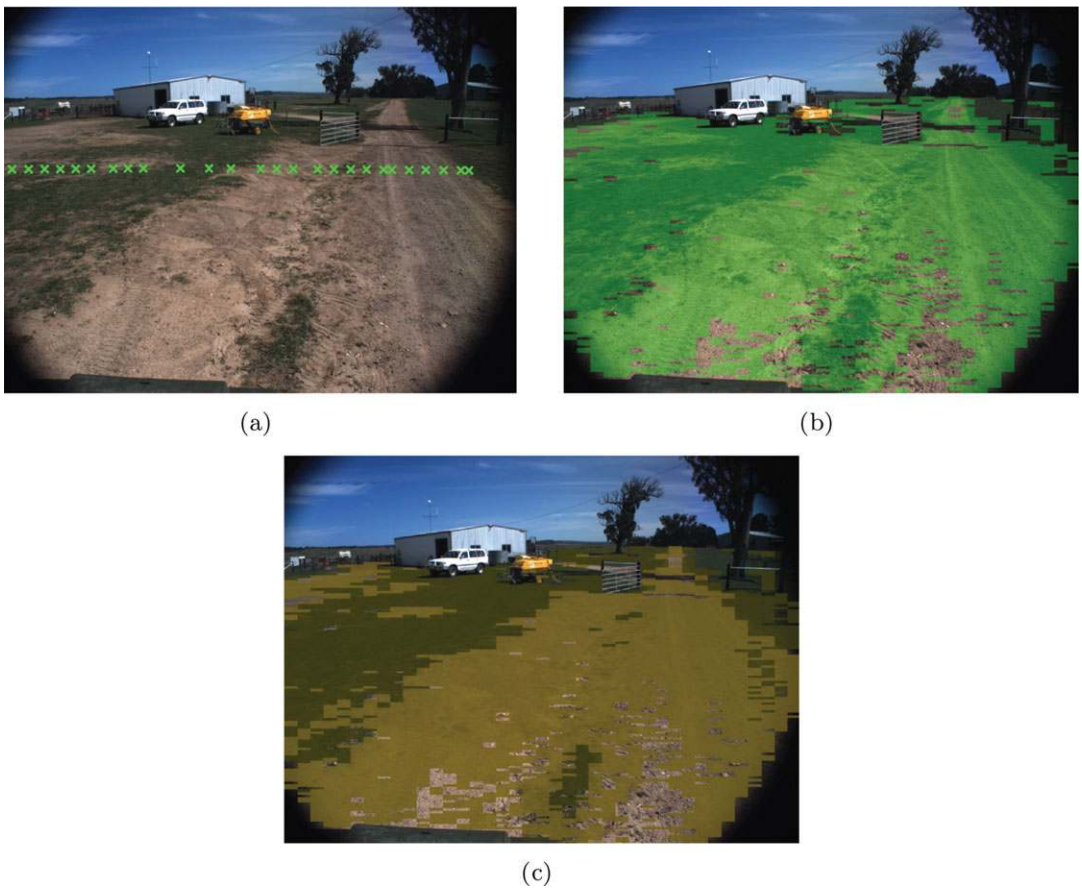


Figure 7. Sample image acquired during field validation: (a) radar-labeled ground points projected on the colocated visual image (green crosses); (b) segmented ground (green pixels); (c) identification of two ground subclasses. In (c), the two terrain types are marked using the average RGB colors of the clusterized training patterns. Note that the blue channel was removed to improve visualization.

For the system to work properly, accurate calibration and synchronization of the sensors has to be ensured, especially to guarantee appropriate reprojection of radar-labeled points on the visual image. Errors in calibration or synchronization may cause incoherency between the data acquired by the two sensors and consequent failure in any stage of the classification system. For instance, points labeled as ground by the radar may be wrongly projected on nonground regions of the image, thus causing an incorrect training of the visual ground model. For the experimental setup adopted in this work, accurate calibration information is available, including both intrinsic and extrinsic parameters of the camera. Specifically, extrinsic parameters define the relative position of the camera reference frame with respect to the radar reference frame, while intrinsic parameters are used to transform metric point coordinates into pixel coordinates. In addition, the sensors are time-synchronized. Detailed information concerning calibration and synchronization can be found in Peynot, Scheding, and Terho (2010).

During the experiments, the CORD vehicle was remotely driven to follow different paths with an average travel speed of about 0.5 m/s and a maximum speed of 1.5 m/s. Visual and radar images were collected and stored for processing offline. In each experiment, the vehicle started its operations from an area that was clear of obstacles, in order to initialize the radar ground model.

7.2. Parameter Analysis

The influence of two main parameters, i.e., the number of Gaussian components for MOG fitting and the cutoff threshold, on the outcome of the visual classifier is analyzed for a test case.

7.2.1. Influence of the Number of Gaussian Components

Due to the nature of the adopted visual features (i.e., texture and color features), a number of Gaussian components

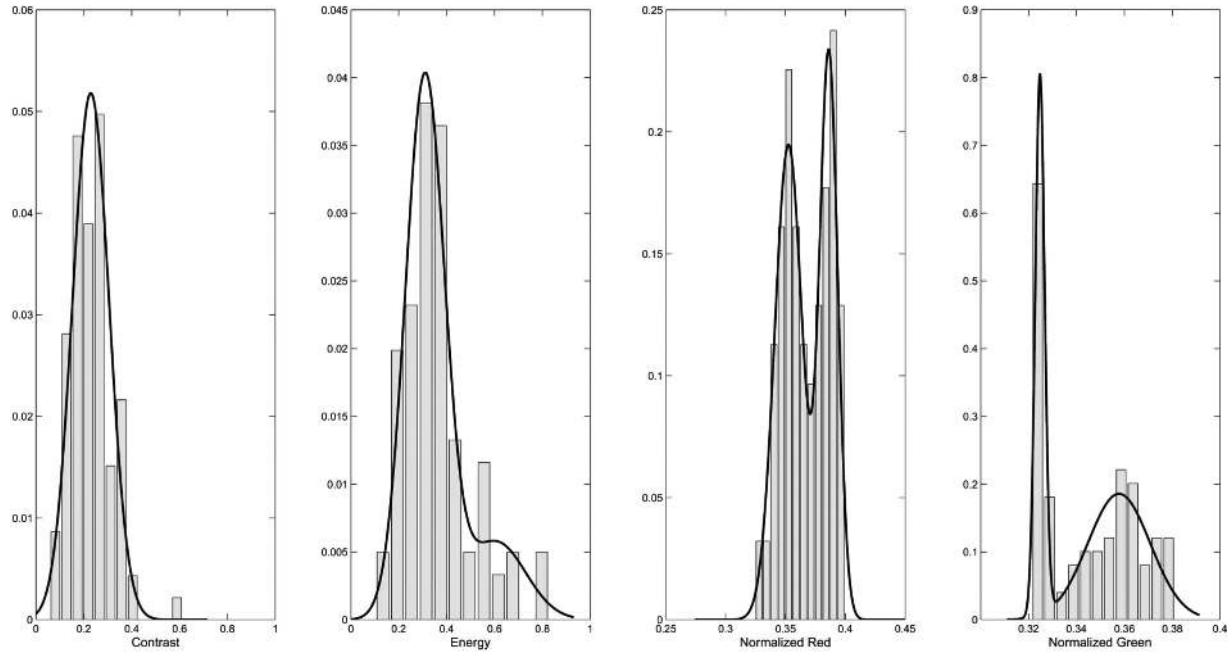


Figure 8. Normalized histograms of the distribution of the visual features for a ground training set including samples of sand and grass. The MOG fit line is displayed for each feature variable, showing that the training set can be reasonably modeled as a multivariate MOG with $K = 2$ components.

$K > 1$ is generally expected in the presence of different terrain types. Hereafter, the influence of the number of Gaussian components on the outcome of the visual classifier is evaluated for the sample scenario shown in Figure 7. Hand-labeling of the original visual image was performed to get a ground-truth for reference. In this test case, the terrain was mainly constituted by grass and sandy soil. In Figure 7(a), the radar-labeled ground points are projected over the original visual image and denoted by green crosses, providing training examples to the visual classifier. The EM-BIC algorithm returned a number of Gaussian components $K = 2$. The classification results produced by the visual classifier for the entire video frame using a cutoff threshold $\alpha = 0.999$ are shown in Figures 7(b) and 7(c). Specifically, in Figure 7(b), pixels associated with ground-labeled patches are marked in green, whereas the two terrain subclasses detected by the system are shown as dark green and brown pixels in Figure 7(c). It should be noted that these two colors have been obtained as a result of averaging RGB colors of the clusterized training samples, thus showing a coherent association between each ground class and its expected color appearance.

The presence of two main ground subclasses is also confirmed when considering the distribution of the feature space for the current (local) training set of the visual ground model, as shown in Figure 8. An approximately bimodal trend is visible for the color feature variables. Although the system detects "ground" as the superclass and is able to distinguish the different ground subclasses based on MOG, semantic classification of the ground subclasses (e.g. "grass" and "sandy-soil") is not explicitly addressed in this paper.

The influence of the number of MOG components on the performance of the visual classifier for this scene is highlighted in the graph of Figure 9. Results are presented in terms of false positive and true positive rates obtained using EM with a given K ranging from 1 to 5, and a fixed cutoff threshold of $\alpha = 0.999$. The tradeoff between small and large K is clearly visible. Using a number of clusters of $K = 2$ as returned by EM-BIC leads to a good tradeoff between false positives and true positives, with the additional advantage of not requiring *a priori* knowledge of the number of clusters in the training set.

7.2.2. Balancing Ground and Nonground Finding: Impact of the Cutoff Threshold

To classify a visual patch as ground or nonground, the Mahalanobis distance between its associated feature vector and the closest component of the current ground model is compared with a critical value, established as $\chi^2_{m;\alpha}$ [see Eq. (4) in Section 4.2]. Changing this threshold will result in a variation of the classification performance. In Figure 10, the impact of the cutoff on the classification performance is shown for the running example of Figure 7 using the ground model

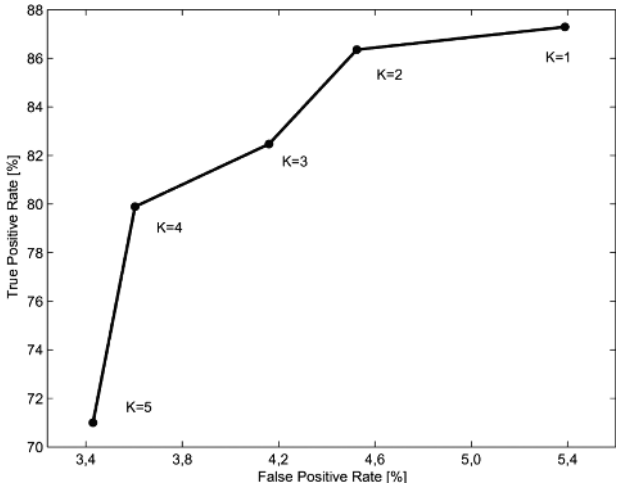


Figure 9. The impact of the number of components (K) on the classification performance for the sample case of Figure 7. Increasing K leads to a tighter fit of the ground model to the training data, which results in a reduction of both false and true positives.

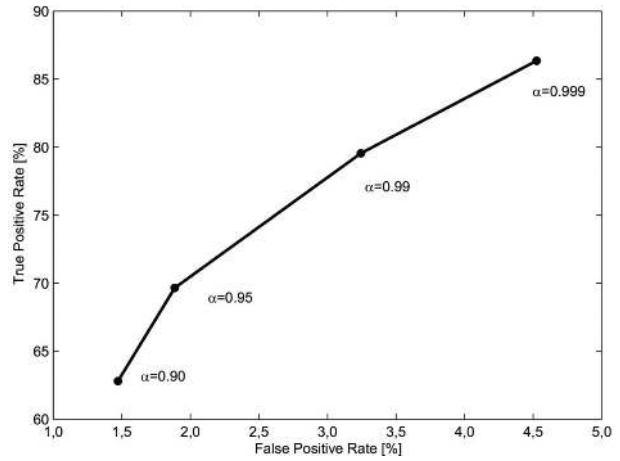


Figure 10. The impact of the cutoff on classification performance for typical significance levels of 10%, 5%, 1%, and 0.1%. Increasing the parameter α while keeping other parameters constant increases the overall number of cells identified as ground.

returned by EM-BIC, and varying the cutoff threshold. It can be observed that the choice of the optimal cutoff threshold results from the balance of ground and nonground finding, i.e., increasing the cutoff while keeping other parameters constant, increases the overall number of pixels identified as ground, thus leading to an increment of both the true positive and the false positive rate. In general, a good cut point is one that produces a large true positive rate and a low false positive rate. A possible solution for optimal threshold selection is that of building a receiver operating characteristic (ROC) curve of the classifier and choosing the threshold that maximizes the difference between the true positive and the false positive rate. A discussion about this approach can be found in Milella et al. (2011). However, a ROC-based heuristic is not a feasible option for online implementation. In the proposed framework, the sensitivity threshold is given once the significance level $(1 - \alpha)$ has been set according to the goal and specifications of the robot's mission (refer to Section 4.2).

7.3. Ground Model Update

In long-range and long-duration navigation, a static ground model will lead to failure due to terrain and environment variation. Here, an adaptive approach is proposed that allows the visual ground model to be continuously updated during robot operations, thus ensuring robustness to environmental changes. To demonstrate the advantage of online learning with respect to a system trained once at the beginning of the robot operation (offline learning), the performance of the classifier using a MOG built and updated online is compared with the results obtained using a MOG trained based only on radar data captured in the first frames and never updated.

In the following, two sequences are analyzed, one acquired in a relatively open area, at daytime, and the other one acquired in the evening, just before the sunset, in an area adjacent to a small eucalyptus forest.

7.3.1. Open Area

A sequence acquired at daytime, in an open area with relatively even ground, is considered. Some obstacles, including a fence, a metallic shed, and static cars, were present. The vehicle was driven to follow an approximately closed-loop path, where the ground changed from mostly sandy to mostly grass to mostly sandy again. Overall, 868 radar images and corresponding visual images were stored. Every tenth image was hand-labeled to build ground truth, resulting in 86 labeled frames. For these frames, Figures 11 and 12 compare the results obtained with and without online ground model update, showing, respectively, some salient frames of the sequence and the classification accuracy at each testing image, obtained using a threshold $\alpha = 0.999$. Specifically, in Figure 11 for each sample image, the first row shows the radar-labeled ground points in the current scan (green crosses) projected on the colocated visual image, while the second and the third rows display the superclass of ground (green colored pixels) as detected by the visual classifier, using, respectively, the offline learning approach and the online learning system. In Figure 12, the solid line with dots refers to the online learning algorithm, while results for the offline learning approach are shown by a solid line with crosses. From these figures, it can be observed that

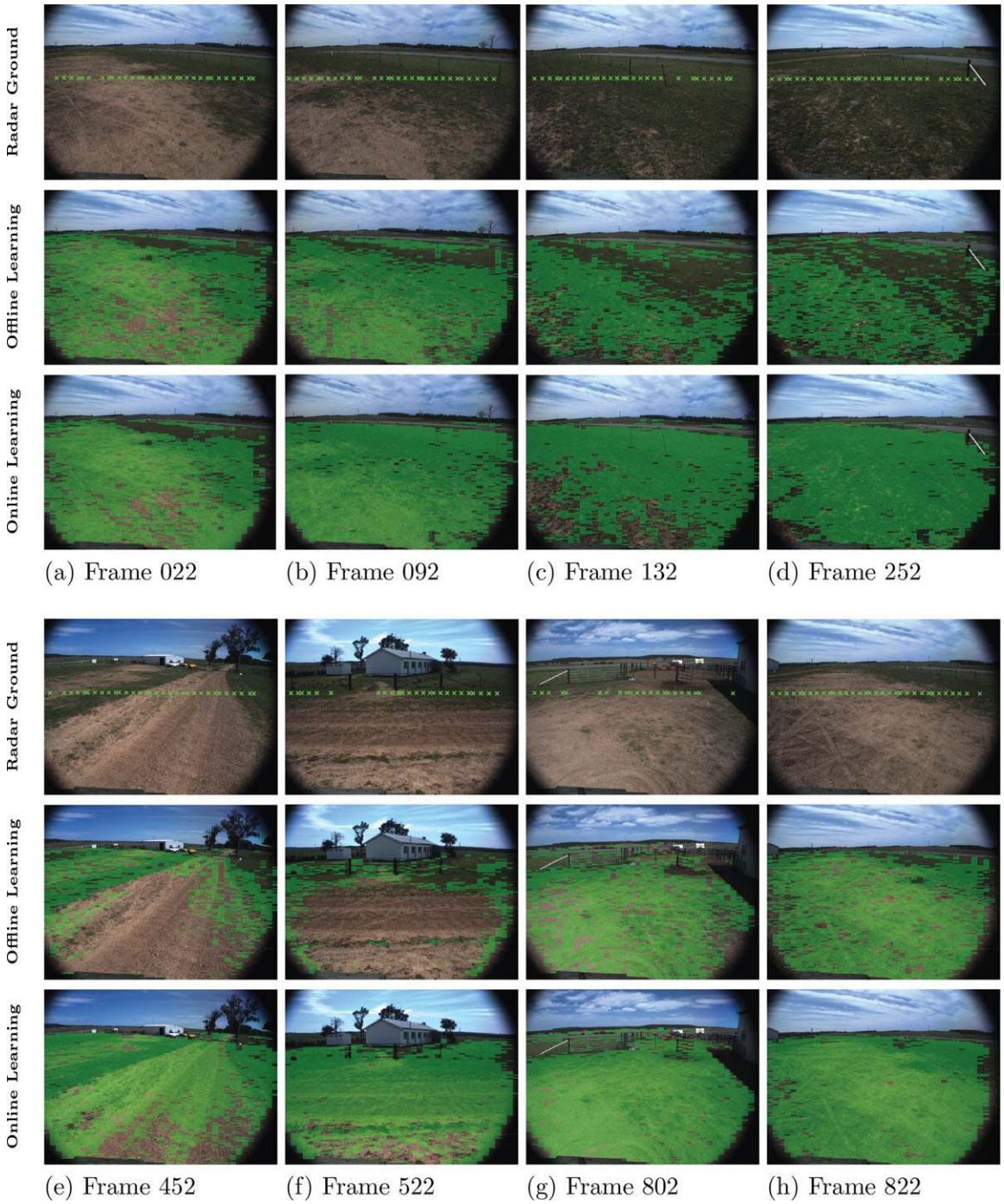


Figure 11. Sample images comparing online and offline learning. First row: radar-based ground labels (green crosses); second row: output of the offline classifier; third row: results of a classifier using the online strategy. Note that, in these images, only the superclass of ground is shown using green pixels. The scenes include objects characterized by vertical structures with different size and distance from the sensors (e.g., poles). When the distance is relatively short and the size of the object is big enough for the given camera resolution, objects are properly detected, as shown, for example, in (e)–(g). Conversely, if an object, e.g., a thin pole, appears too small in the image due to excessive distance from the sensor or to small dimensions, as in (b)–(d), it is likely to be neglected, also due to the use of a block-based segmentation approach.

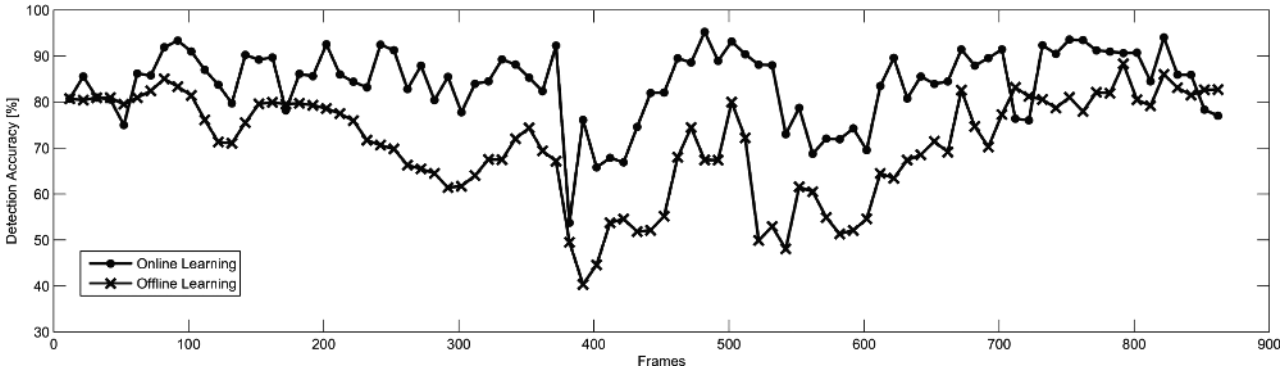


Figure 12. Classification accuracy for a sequence acquired in the field. Results obtained using the online learning strategy (solid line with dots) are compared with those obtained by training the classifier once at the beginning of the robot operation (solid line with crosses).

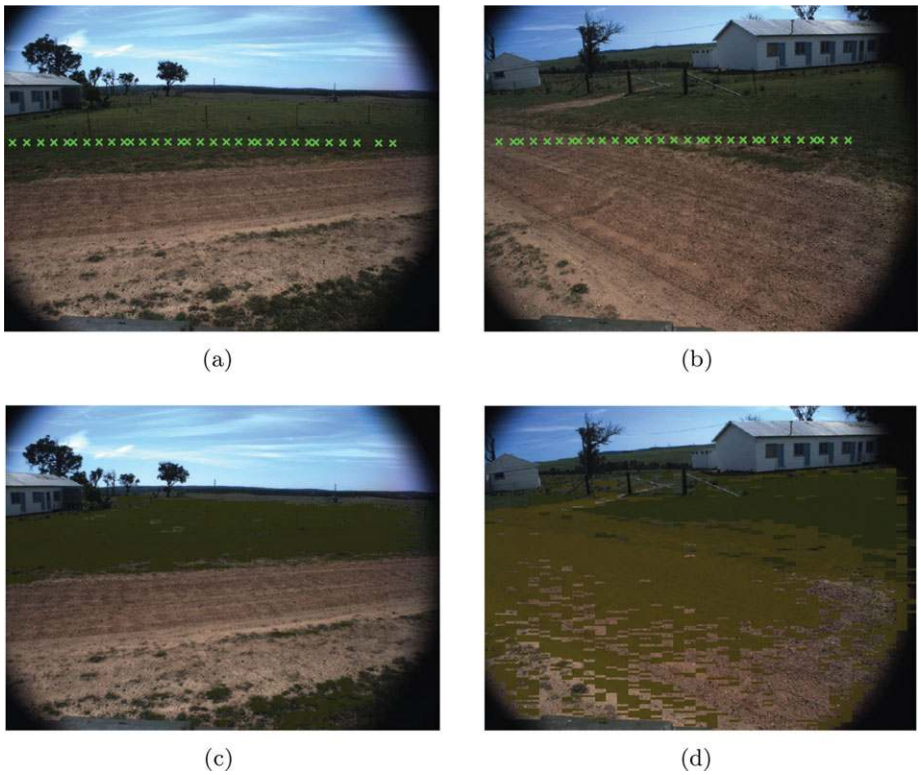


Figure 13. Adaptation to different terrain types. Initially, the ground model includes instances of grass only and sand is not recognized as ground, (a)–(c). As soon as enough instances of sand are added to the training rolling window, sand is segmented as ground, (b)–(d). In (a) and (b), green crosses denote radar-labeled ground points in the current scans. In (c) and (d), image pixels belonging to ground are marked using the average RGB color of the respective terrain type (the blue channel was removed to improve visualization).

in the first frames (approximately up to frame 90), both the online and the offline algorithm performed well, since the ground appearance did not vary significantly. Successively, the offline algorithm degraded, until the vehicle returned to its starting position (approximately at frame 800).

The ability of the online system to recover from poor classification performance can be observed in Figure 13 for the frames 382–392. Radar-based training instances of ground detected in the current scans are shown as green crosses in Figures 13(a) and 13(b) for the two frames, while

Table III. Performance of the radar-supervised visual classifier in the *Open Area*: classification accuracy using online learning (i.e., with ground model update) versus offline learning (i.e., without ground model update).

	Online learning	Offline learning
Mean accuracy (%)	84.07	70.81
St. dev. of accuracy (%)	7.80	11.34
Mean F1 score (%)	85.13	67.71
St. dev. of F1 score (%)	8.96	16.91

the results of visual classification are shown in Figures 13(c) and 13(d). In this period, the vehicle performed a left-hand turning manoeuver. Due to the narrow radar field of view, initially the ground model included instances of grass only; therefore, sand was not recognized as ground by the visual classifier [see Figures 13(a)–13(c)]. Correspondingly, a decrement in the detection accuracy can be observed in Figure 12. This highlights an intrinsic drawback of the online learning approach, as a ground portion would not be recognized as ground until samples of it are included in the ground model. On the other hand, the online learning algorithm rapidly adapted as soon as enough instances of sand were added to the training rolling window, so that also sand was segmented as ground [see Figures 13(b)–13(d)], with a consequent increment of the detection accuracy. The numerical results summarizing the classification performance for the sequence using the online approach compared to those of the offline learning system are reported in Table III in terms of accuracy and F1 score. Overall, the average accuracy resulted in 70.81% with a standard deviation of 11.34% for the offline learning system, while the online learning demonstrated better performance with an average accuracy and standard deviation of 84.07% and 7.80%, respectively.

7.3.2. Eucalyptus Forest Area

The capability of the online learning approach to deal with abrupt variations in the illumination conditions is demonstrated in Figures 14 and 15 for a sequence acquired in the evening, just before the sunset, in an area adjacent to

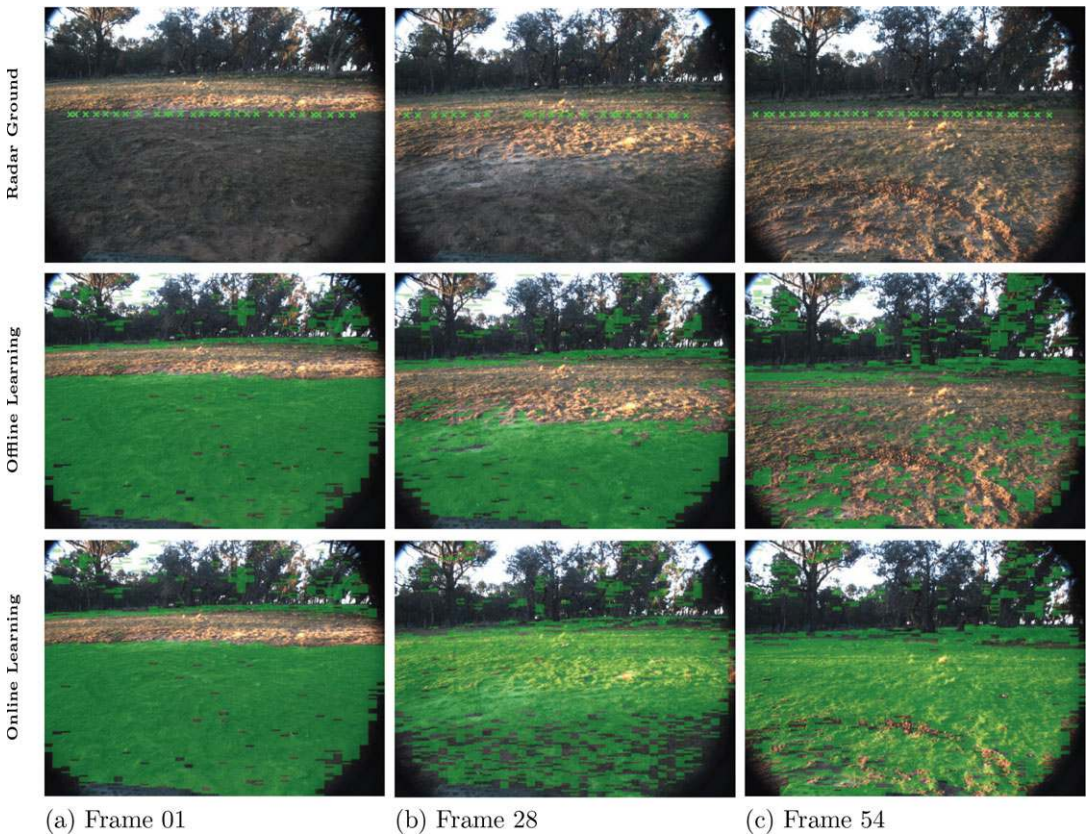


Figure 14. Adaptation to lighting variations and shadows: sample images comparing online and offline learning. First row: radar-based ground labels (green crosses); second row: output of the offline classifier; third row: results of a classifier using the online strategy. Note that, in these images, only the superclass of ground is shown using green pixels.

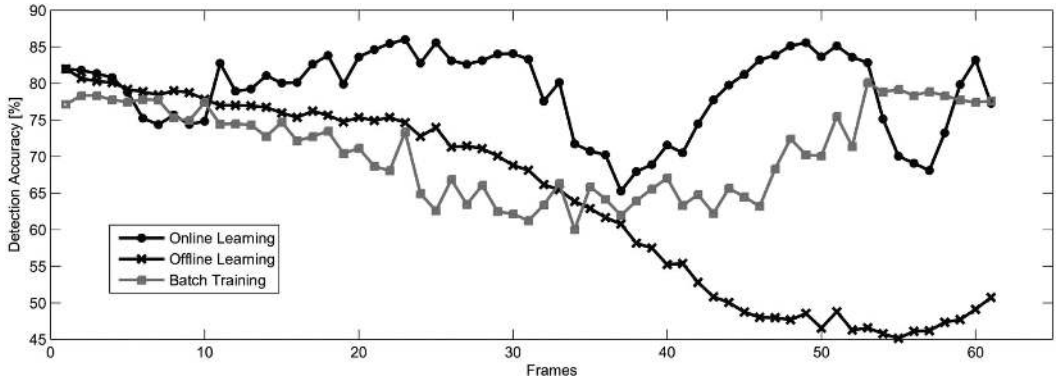


Figure 15. Classification accuracy for a sequence acquired in the field in the presence of heavy shadowing. Results obtained using the online learning strategy (black line with dots) are compared with those obtained by training the classifier once at the beginning of the robot operation (black line with crosses), as well as with the performance of a MOG classifier trained in batch mode (gray line with square markers).

Table IV. Performance of the radar-supervised visual classifier for the sequence acquired at sunset in the *Eucalyptus Area*. First column: results obtained using online learning (i.e., with ground model update). Second column: results of the offline learning classifier (i.e., trained once at the beginning of the sequence). Third column: performance of the batch trained classifier. Fourth column: performance of the manually tuned classifier.

	Online learning	Offline learning	Batch training (prior dataset)	Manually tuned classifier (prior dataset)
Mean accuracy (%)	78.90	64.21	70.80	74.89
St. dev. of accuracy (%)	5.53	12.96	6.10	7.05
Mean F1 score (%)	80.22	60.38	70.30	75.13
St. dev. of F1 score (%)	7.08	19.76	8.84	11.08

a small eucalyptus forest. Due to the presence of high trees, and since the experiment was performed before dusk, the environment encountered by the vehicle was characterized by long shadows and low-lighting conditions. These specific aspects make it suitable for this analysis. The sequence includes 61 radar scans and corresponding visual images. Some key frames are shown in Figure 14 comparing online and offline learning results. It can be noticed that the online learning approach allows for a rapid adaptation to the changing appearance of the ground due to illumination variations. Figure 15 shows the classification accuracy obtained by using the online learning approach, compared to the classification accuracy of the fixed classifier (i.e., trained only once at the beginning of the sequence). Again, the solid line with dots refers to the online learning algorithm, while the solid line with crosses denotes the results of the offline learning approach. The accuracy of the fixed classifier deteriorates steadily. In contrast, the accuracy of the online learning system remains relatively high, demonstrating its capability of adapting to the changing environment. Results are presented in the first and second columns of Table IV in terms of accuracy and F1 score. Overall, the online learning approach resulted in an average accuracy of 78.90% with a standard deviation of 5.53%, while using the offline learning system yielded a lower average accuracy of 64.21% with a standard deviation of 12.96% was obtained.

As a final remark, it should be noted that since many false positives arise from an erroneous classification of the pixels belonging to the upper parts of the trees (see the examples of Figure 14), the system's performance may be improved by adopting an algorithm for horizon detection and sky removal, as explained, for example, in Dahlkamp et al. (2006).

7.4. Online Learning Versus Batch Training and Manual Tuning

Supervised classification techniques rely on the availability of reference samples to be used in the training phase of the classification algorithm. Reliability of the training set depends on both the quantity and quality of the available samples. In many autonomous vehicle applications, such as in exploration of unknown environments, the use of a batch trained classifier is not a viable option, as the training set is typically not available *a priori*. On the other hand, onsite generation of the training set would entail a significant delay between the time training images were collected and the time the classifier could be implemented. What would typically happen is that the training set would be collected and labeled at a certain time and spatial location, and it would be applied successively for prediction with a delay of hours or days, without any possibility of online

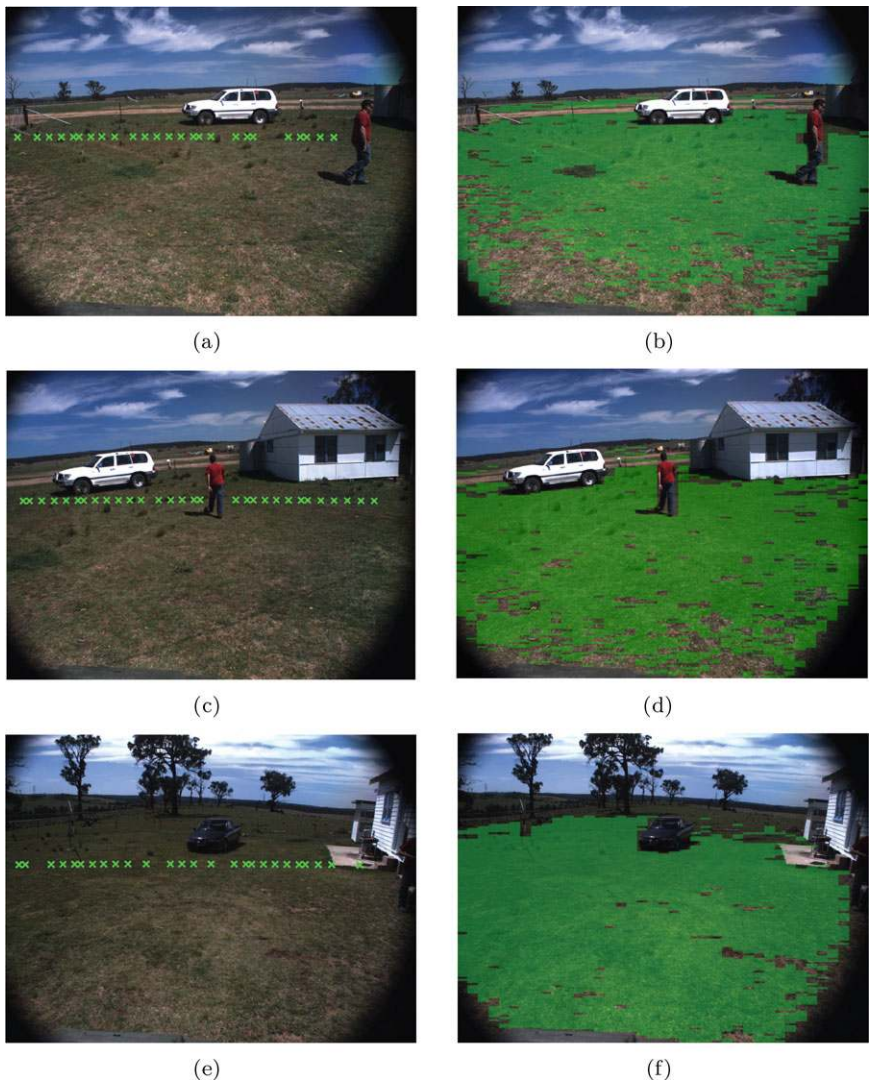
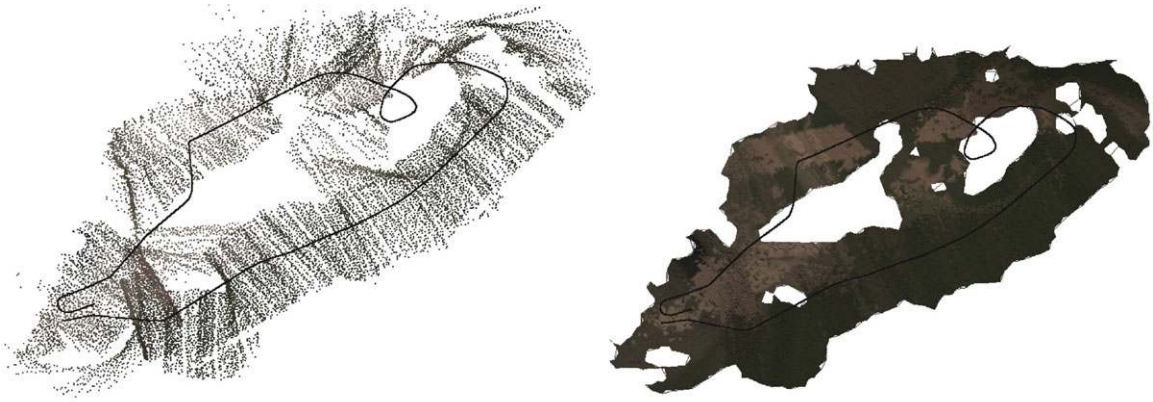


Figure 16. Typical classification results obtained from the radar-vision combined system. Left side: green crosses mark radar-labeled training examples detected in the current scan. Right side: green pixels denote the ground superclass segmented by the visual classifier.

Table V. Performance of the radar-supervised visual classifier on 105 images taken from different datasets.

Mean precision (%)	93.50
St. dev. of precision (%)	6.98
Mean recall (%)	77.87
St. dev. of recall (%)	13.39
Mean accuracy (%)	83.47
St. dev. of accuracy (%)	8.20
Mean F1 score (%)	84.18
St. dev. of F1 score (%)	9.65

update during the vehicle travel to deal with environmental changes. Here, for the purpose of comparison, the results of the online learning approach for the sequence acquired in the eucalyptus area are compared with the results obtained using a MOG-based classifier trained in batch mode using a dataset previously collected in a similar context. In the batch training experiment, first the radar classifier was run on the whole training sequence in order to build a MOG ground model. Afterward, this model was used to classify the entire scene in all frames of the test sequence. Overall, 24,644 radar-labeled ground samples, acquired along a path of about 150 m, were used for training. The classification accuracy of the batch trained classifier for each frame of the



(a) (b)

Figure 17. (a) Radar-generated map, shown as raw data obtained from the classifier with associated RGB data obtained from the visual classifier. (b) Same radar-vision data after Delaunay triangulation.

sequence is shown in Figure 15 as a gray line with square markers. It can be seen that accuracy is generally lower than the online learning approach with a mean value of 70.80% and a standard deviation of 6.10% (refer to the third column of Table IV for numerical results).

For the same sequence, it is also interesting to compare the proposed self-learning approach with that previously presented by the authors [see Milella, Reina, Underwood, & Douillard (2014)], where parameters were hand-tuned for both the radar and the visual classifier. In detail, the radar module was based on a set of "expert" classification rules with manually tuned thresholds [please refer to Table I in Reina et al. (2011a)], whereas the visual classifier used a single multivariate Gaussian to model the ground. Results are reported in the fourth column of Table IV. One should note that the online learning approach outperforms the hand-tuned one (note that the tuning was based on a prior dataset) with the further advantage of being automatic at any stage from training to prediction without any *a priori* information.

7.5. Ground Detection and Mapping

An overall assessment of the system's performance was performed using a subset of salient images ($s_b = 105$) taken from different data sets. Some typical results obtained from the classifier are shown in Figure 16. Images on the left side show the original image overlaid with the training instances provided by the radar-based classifier for the current scan, while the results of the visual classifier are shown on the right side, where pixels associated with the ground superclass are marked in green. The numerical results summarizing the classification performance are presented in Table V as average value and statistical spread. Specifically, the precision resulted in 93.50% with a standard deviation of 6.98%, the recall was 77.87% with a standard deviation of 13.39%, the accuracy resulted in 83.47% with a standard deviation of 8.20%, and the F1 score was 84.18% with a standard deviation of 9.65%. Finally, it should be recalled from Section 3 that when a single radar observation, i, is successfully labeled as ground, an estimate of its range distance $R_{0,i}$ is also returned by the fitting process. When combined with the localization estimation of the vehicle, this provides a 3D georeferenced position for the labeled point. This aspect highlights an additional advantage of combining radar with vision, that is, the generation of "rich" 3D data, where radar provides range information and vision color-based subground separation toward an augmented map of the environment. For a complete overview, the results obtained from the same sequence analyzed in Section 7.3.1 are shown in Figure 17(a). The ground labeled observations are denoted by the average RGB color associated with the detected Gaussian component. The path followed by the robot is also shown by a solid black line. Figure 17(b) depicts the same data after a post-processing step applying a Delaunay triangulation. This figure demonstrates that the system is capable of providing a clear understanding of the environment, suitable for robotic applications including scene interpretation and autonomous navigation.

8. CONCLUSIONS

In this paper, a unified self-learning framework for online ground detection in outdoor environments was proposed. It can be applied to a single sensor or to combine multiple sensors. Within this framework, a radar-supervised visual classifier was developed that allows an autonomous vehicle, operating in natural terrains, to construct online a visual model of the ground and perform ground segmentation.

The proposed system presents two main characteristics of interest: 1) it is fully self-supervised, as both the radar module and the visual module take advantage of an

automatic training procedure, thus avoiding time-consuming manual labeling; 2) it uses an online learning approach, i.e., the ground model is learned and updated in the field, which may be useful for long-range navigation in unknown environments.

Experimental results obtained with an unmanned vehicle operating in a rural environment were presented, demonstrating the capability of the system to adapt to environmental changes, such as variations in the illumination conditions and of the ground appearance, after an automatic initialization phase, with no need of human supervision for training.

Self-learning systems may be the only option when no prior datasets are available for training, and wherever the use of a static ground model would rapidly lead to poor classification outcome due to highly variable environmental conditions. Nevertheless, the use of a completely self-supervised procedure brings intrinsic limitations as well. First, for the system to work properly, the two sensors have to be accurately calibrated and synchronized in order to have a coherent data association, which is a prerequisite for correct data fusion. Furthermore, the overall accuracy of the classifier depends on both the ability of the supervising module to produce a reliable training set, and on the robustness of the visual classifier; therefore, the overall system performance is affected by error propagation. With respect to a batch trained classifier, whereby all training examples are simultaneously available, incremental learning systems have to learn sufficient information to accommodate new classes that may be introduced with new data before performing proper scene recognition. For instance, in the specific case of ground segmentation, parts of the ground that have not yet been included in the training set would not be properly recognized as ground and would be erroneously labeled as obstacles. This issue is particularly critical for the sensor configuration adopted in this research, where the supervising sensor (i.e., the radar) has a field of view much narrower than the supervised sensor (i.e., the camera), so that the area where the training samples are collected covers a small portion of the environment, while classification is performed on the entire video frame. On the other hand, the system may suffer from forgetting previously acquired knowledge. In the proposed framework, this problem may be partly mitigated by setting an appropriate size of the training rolling window; however, implementation of an efficient strategy to preserve previously acquired knowledge would be beneficial and will be part of further investigation by the authors. Future work will also include experimental validation in the presence of hills, ditches, trees, and vegetation, which make ground estimation challenging, as well as in cluttered urban settings. This will require specific research on more complex visual features to better deal with the underlying structures of ground and obstacles. Another focus of the research will address the analysis of the system performance under failure conditions of one sensor (e.g., when the radar is affected by specularity and reflection, or the camera fails due to visual obscurants).

ACKNOWLEDGMENTS

The authors are thankful to the Australian Department of Education, Employment and Workplace Relations for supporting the project through the 2010 Endeavour Research Fellowship 1745_2010. The authors would like also to thank the National Research Council, Italy, for supporting this work under the CNR 2010 Short Term Mobility program. This work is supported in part by the Australian Centre for Field Robotics (ACFR) at the University of Sydney. The financial support of the ERA-NET ICT-AGRI through the grant Ambient Awareness for Autonomous Agricultural Vehicles (QUAD-AV) is also gratefully acknowledged.

REFERENCES

Alessandretti, G., Broggi, A., & Cerri, P. (2007). Vehicle and guard rail detection using radar and vision data fusion. IEEE Transactions on Intelligent Transportation Systems, 8(1), 95–105.

Atreya, A., Cattle, B., Momen, S., Collins, B., Downey, A., Franken, G., Glass, J., Glass, Z., Herbach, J., Saxe, A., Ashwash, I., Baldassano, C., Hu, W., Javed, U., Mayer, J., Benjamin, D., Gorman, L., & Yu, D. (2007). DARPA Urban Challenge Princeton University. Technical report.

Balkenius, C., & Johansson, M. (2007). Finding colored objects in a scene. In LUCS Minor, 12.

Baraldi, A., & Parmiggiani, F. (1995). An investigation of the textural characteristics associated with gray level cooccurrence matrix statistical parameters. IEEE Transactions on Geoscience and Remote Sensing, 33(2), 293–304.

Brooker, G., Hennessey, R., Bishop, M., Lobsey, C., Durrant-Whyte, H., & Birch, D. (2006). High-resolution millimeter-wave radar systems for visualization of unstructured outdoor environments. Journal of Field Robotics, 23(10), 891–912.

Brooks, C., & Iagnemma, K. (2008). Visual detection of novel terrain via two-class classification. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems.

Brooks, C., & Iagnemma, K. (2012). Self-supervised terrain classification for planetary surface exploration rovers. Journal of Field Robotics. Special Issue: Special Issue on Space Robotics, Part I, 29(3), 445–468.

Cossu, R. (1988). Segmentation by means of textural analysis. Pixel, 1(2), 21–24.

Dahlkamp, H.A., Kaehler, D. S., Thrun, S., & Bradski, G. (2006). Self-supervised monocular road detection in desert terrain. In Proceedings of the Robotics Science and Systems Conference.

DeSouza, G., & Kak, A. (2002). Vision for mobile robot navigation: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(2), 237–267.

Dima, C., Vandapel, N., & Hebert, M. (2004). Classifier fusion for outdoor obstacle detection. In Proceedings of the International Conference on Robotics and Automation (ICRA) (vol. 1, pp. 665–671).

Filitchkin, P., & Byl, K. (2012). Feature-based terrain classification for LittleDog. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (pp. 1387–1392).

Gevers, T., Weijer, J. V. D., & Stokman, H. (2006). Color feature detection. Chapter on Color image processing: Emerging applications. CRC Press.

Haralick, R., Shanmugam, K., & Dinstein, I. (1973). Textural features for image classification. IEEE Transactions on Systems, Man and Cybernetics, 3(6), 21–24.

Hong, T., Chang, T., Rasmussen, C., & Shneier, M. (2002). Road detection and tracking for autonomous mobile robots. In Proceedings of SPIE Aerosense Conference (pp. 1194–1200).

Huertas, A., Matthies, L., & Rankin, A. (2005). Stereo-based tree traversability analysis for autonomous off-road navigation. In Proceedings of the Workshop of Applications of Computer Vision.

Ji, Z., Luciw, M., Weng, J., & Zeng, S. (2011). Incremental online object learning in a vehicular radar-vision fusion framework. IEEE Transactions on Intelligent Transportation Systems, 12(2), 402–411.

Jocherm, T., Pomerleau, T., & Thorpe, C. (1995). Vision-based neural network road and intersection detection and traversal. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems.

Khan, Y., Masselli, A., & Zell, A. (2012). Visual terrain classification by flying robots. In IEEE International Conference on Robotics and Automation.

LeCun, Y., Huang, F. J., & Bottou, L. (2004). Learning methods for generic object recognition with invariance to pose and lighting. In Proceedings of the Conference on Computer Vision and Pattern Recognition (vol. 2, pp. 97–104).

Manduchi, R., Castano, A., Talukder, A., & Matthies, L. (2003). Obstacle detection and terrain classification for autonomous off-road navigation. Autonomous Robot, 18, 81–102.

Mardia, K., Kent, J., & Bibby, J. (1979). Multivariate analysis. London: Academic Press.

Mateus, D., Avina, G., & Devy, M. (2005). Robot visual navigation in semi-structured outdoor environments. In Proceedings of the IEEE International Conference on Robotics and Automation.

Milella, A., Reina, G., Underwood, J., & Douillard, B. (2011). Combining radar and vision for self-supervised ground segmentation in outdoor environments. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (pp. 255–260).

Milella, A., Reina, G., Underwood, J., & Douillard, B. (2014). Visual ground segmentation by radar supervision. Robotics and Autonomous Systems, 62(5), 696–706.

Pagnot, R., & Grandjean, P. (1995). Fast cross country navigation on fair terrains. In Proceedings of the IEEE International Conference on Robotics and Automation (pp. 2593–2598).

Permuter, H., Francos, J., & Jermyn, I. (2006). A study of Gaussian mixture models of color and texture features for image classification and segmentation. Pattern Recognition, 39, 695–706.

Peynot, T., Scheding, S., & Terho, S. (2010). The Marulan data sets: Multi-sensor perception in natural environment with challenging conditions. International Journal of Robotics Research, 29(13), 1602–1607.

Peynot, T., Underwood, J., & Scheding, S. (2009). Towards reliable perception for unmanned ground vehicles in challenging conditions. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, St. Louis, MO.

Pomerleau, D. (1989). ALVINN: An autonomous land vehicle in a neural network. In Advances in neural information processing systems: Morgan Kaufmann.

Raftery, A. E. (1995). Bayesian model selection for social research (with discussion). Sociological Methodology, 25, 111–196.

Rasmussen, C. (2002). Combining laser range, color, and texture cues for autonomous road following. In Proceedings of the IEEE International Conference on Robotics and Automation.

Reed, T., & du Buf, J. H. (1993). A review of recent texture segmentation and feature extraction techniques. Computer Vision, Graphics, and Image Processing, Image Understanding, 57(3), 359–372.

Reina, G., Ishigami, G., Nagatani, K., & Yoshida, K. (2010). Odometry correction using visual slip-angle estimation for planetary exploration rovers. Advanced Robotics, 24(3), 359–385.

Reina, G., & Milella, A. (2012). Towards autonomous agriculture: Automatic ground detection using trinocular stereovision. Sensors, 12(9), 12405–12423.

Reina, G., Milella, A., Halft, W., & Worst, R. (2013). LIDAR and stereo imagery integration for safe navigation in outdoor settings. In Safety, Security, and Rescue Robotics (SSRR), 2013 IEEE International Symposium (pp. 1–6).

Reina, G., Milella, A., & Underwood, J. (2012a). Radar-vision integration for self-supervised scene segmentation. In Proceedings of Robotics: Science and Systems Conference, Workshop Beyond Laser and Vision: Alternative Sensing Techniques for Robotic Perception, University of Sydney, Australia.

Reina, G., Milella, A., & Underwood, J. (2012b). Self-learning classification of radar features for scene understanding. Robotics and Autonomous Systems, 60(11), 1377–1388.

Reina, G., Underwood, J., & Brooker, G. (2011a). Short-range radar perception in outdoor environments. In Proceedings of Towards Autonomous Robotics Systems (TAROS) (pp. 265–276).

Reina, G., Underwood, J., Brooker, G., & Durrant-Whyte, H. (2011b). Radar-based perception for autonomous outdoor vehicles. Journal of Field Robotics, 28(6), 894–913.

Schwarz, G. (1978). Estimating the dimension of a model. Journal of the American Statistical Association, 6, 461–464.

Singh, S., Simmons, R., Smith, T., Stentz, A., Verma, V., Yahja, A., and Schwehr, K. (2000). Recent progress in local and global traversability for planetary rovers. In Proceedings of the IEEE International Conference on Robotics and Automation (pp. 1194–1200).

Sofman, B., Lin, E., Bagnell, J. A., Vandapel, N., & Stentz, A. (2006). Improving robot navigation through self-supervised online learning. In Proceedings of Robotics: Science and Systems.

Sole, A., Mano, O., Stein, G., Kumon, H., Tamatsu, Y., & Shashua, A. (2004). Solid or not solid: Vision for radar target validation. In Proceedings of IEEE Intelligent Vehicles Symposium.

Sung, G., Kwak, D., Kim, D., & Lyou, J. (2008). Terrain cover classification based on wavelet feature extraction. In Proceedings of the International Conference on Control, Automation and Systems.

Tax, D. (2001). One-class classification. Concept learning in the absence of counter examples. Ph.D. thesis, Delft University of Technology, Delft, The Netherlands.

Vernaza, P., Taskar, B., & Lee, D. (2008). Online, self-supervised terrain classification via discriminatively trained submodular Markov random fields. In Proceedings of the IEEE International Conference on Robotics and Automation (pp. 2750–2757).

Wu, S., Decker, S., Chang, P., Camus, T., & Eledath, J. (2009). Collision sensing by stereo vision and radar sensor fusion. IEEE Transactions on Intelligent Transportation Systems, 10(4), 606–614.

Zhou, S., Xi, J., McDaniel, M., Nishihata, T., Salesses, P., & Iagnemma, K. (2012). Self-supervised learning to visually detect terrain surfaces for autonomous robots operating in forested terrain. Journal of Field Robotics, 29(2), 277–297.

Zhu, X. (2005). Semi-supervised learning literature survey. Technical Report 1530, Computer Sciences, University of Wisconsin–Madison.