2011 IEEE/RSJ International Conference on
Intelligent Robots and Systems
September 25-30, 2011. San Francisco, CA, USA

Combining Radar and Vision for Self-Supervised Ground Segmentation in Outdoor Environments

Annalisa Milella, Giulio Reina, James Underwood, and Bertrand Douillard

Abstract—**Ground segmentation is critical for a mobile robot to successfully accomplish its tasks in challenging environments. In this paper, we propose a self-supervised radar-vision classification system that allows an autonomous vehicle, operating in natural terrains, to automatically construct online a visual model of the ground and perform accurate ground segmentation. The system features two main phases: the training phase and the classification phase. The training stage relies on radar measurements to drive the selection of ground patches in the camera images, and learn online the visual appearance of the ground. In the classification stage, the visual model of the ground can be used to perform high level tasks such as image segmentation and terrain classification, as well as to solve radar ambiguities. The proposed method leads to the following main advantages: (a) a self-supervised training of the visual classifier, where the radar allows the vehicle to automatically acquire a set of ground samples, eliminating the need for time-consuming manual labeling; (b) the ground model can be continuously updated during the operation of the vehicle, thus making it feasible the use of the system in long range and long duration navigation applications. This paper details the proposed system and presents the results of experimental tests conducted in the field by using an unmanned vehicle.**

I. Introduction

AUTONOMOUS vehicle operations in outdoor environments challenge robotic perception and make integration of information from multiple sensors a major requirement. Among other sensors, optical devices, either active or passive, have proved to be especially effective to provide the robot with the ability to understand its surroundings and successfully accomplish its tasks.

Manuscript received March 15, 2011. 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 research was undertaken through the Centre for Intelligent Mobile Systems (CIMS), and was funded by BAE Systems as part of an ongoing partnership with 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.

A. Milella is with the Institute of Intelligent Systems for Automation (ISSIA), National Research Council (CNR), via G. Amendola 122/D, 70126, Bari, Italy (corresponding author; phone: +39 080 5929453; e-mail: milella@ba.issia.cnr.it).

G. Reina is with the University of Salento, Department of Engineering for Innovation, Via Arnesano, 73100 Lecce, Italy (e-mail: giulio.reina@unisalento.it).

J. Underwood and B. Douillard are with the Australian Centre for Field Robotics, University of Sydney, Rose Street Building (J04), NSW 2006, Australia (e-mail: {j.underwood, b.douillard}@acfr.usyd.edu.au).

As a type of active imaging sensor, millimeter-wave radars have been increasingly used in autonomous vehicle systems, since they provide relatively accurate measurements of obstacles in low visibility conditions, including the presence of dust, fog, and rain. Radar also provides a rich source of information allowing for multiple object detection within a single beam whereas other range sensors are limited to one target return per emission. However, radar has shortcomings as well, including large footprint, specularity and reflection effects, and limited range resolution, all of which may result in poor environment survey or make it difficult to extract object features for classification and scene interpretation tasks. Consequently, to expand the range of possible applications, it is necessary to combine radar with other sensors. Video sensors lend themselves very well to this purpose. Being passive devices, cameras are affected by environmental factors, such as lighting conditions. Nevertheless, 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. Due to the complementary characteristics of the two sensors, it is reasonable to combine them in order to get improved performance [1], [2].

Radar and vision fusion has been discussed mostly in the context of driver assistance systems featuring object detection and classification modules [1]-[7]. For instance, in [1] radar and vision independently detect targets of interest; 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 non-vehicle is developed in [2]. It uses radar data to select visual attention windows, which are then assigned a label and processed to extract features to train a Multi-layer In-place Learning Network (MILN). In [3], a vehicle detection system fusing radar and vision data is proposed. First, radar data are used to locate areas of interest on images. Then, 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.

In this paper, we propose a novel radar-vision combination for accurate ground segmentation by an autonomous vehicle operating in natural terrain.

Persistent ground segmentation is critical for a robot to improve perception under all conditions, with many important applications, including environment classification

(a) (b)

Fig. 1. The CORD UGV employed in this research (a), and its sensor suite (b).

and scene interpretation. While in structured environments, such as in urban contexts, the task of ground identification can be effectively performed by exploiting some distinctive roadway markings, in natural terrain, no *a priori* information about the ground surface is usually available. Furthermore, ground structure and appearance may significantly change during the vehicle operation; therefore, road detection algorithms based on specific cues are not suitable, unless re-tuning of road markers or re-training of classifiers is performed, generally with human supervision.

To overcome the limitations of these methods, self-supervised terrain classification approaches have been developed, whereby the output of a sensor is used to train another sensor. For instance, in [8], self-supervised terrain classification is performed using a previously trained vibration-based classifier, which provides labels to train online a visual classifier. In [9], data from a stereo camera is used to train a monocular image classifier that segments an image into obstacle and ground patches in the submodular Markov Random Field (MRF) framework. Another notable example of self-supervised ground segmentation can be found in [10], using a laser scanner and a monocular camera. Specifically, the laser is employed to scan for flat, drivable surface in the vicinity of the vehicle. Then, this area is 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.

In this work, we exploit a similar concept to develop a novel online, self-supervised ground segmentation method, using a radar-vision system. The main contribution of the proposed approach relies on the combination of a radar-based segmentation method with a vision-based classification system to incrementally construct a visual model of the ground during the operation of the vehicle. Specifically, first, the radar image is analyzed to identify ground returns; then, the radar ground-labeled points are projected in the camera image and are used to automatically select and label visual ground patches. That results in a self-supervised ground modeling system, since visual ground samples are provided by radar, thus eliminating the need for time consuming manual labeling. In addition, since the ground model can be continuously updated based on the most recent radar scans, this approach is suited to long range and long duration navigation conditions.

Once constructed, the visual model of the ground can be either used to perform high level tasks, such as terrain characterization and visual scene segmentation, or to supplement the radar sensor by solving radar ambiguities deriving from reflections and occlusions.

The system was validated in the field using the CAS Outdoor Research Demonstrator (CORD), an 8 wheel skid-steering all-terrain unmanned vehicle (Fig. 1(a)). The robot sensor suite is shown in Fig. 1(b) including a Prosilica Mono-CCD megapixel Gigabit Ethernet camera, pointing down (a few degrees of pitch) and a 94 GHz Frequency Modulated Continuous Wave (FMCW) Radar, custom built at the Australian Center for Field Robotics (ACFR) for environment imaging [11]. The camera acquires images of 1360×1024 pixels, with resolution of 72×72 ppi, and frame rate of 10 fps. The radar has maximum range of 120 m, raw resolution range of 0.25 m, horizontal FOV of 360 deg, and angular scan rate of 3.0 Hz. 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 during the experiments.

The rest of the paper is organized as follows. Section II details the developed approach. Experimental results are presented in Section III. Conclusions are drawn in Section IV.

II. Description of The Approach

Our objective is that of providing an autonomous vehicle, operating in natural terrain, with the ability of performing precise ground segmentation. Specifically, we propose a self-supervised method for online ground modeling and segmentation using radar and vision data.

The overall architecture of the system is shown in Fig. 2. The processing pipeline takes as input the raw data obtained by the radar and the camera, and features two main phases: a training phase and a classification phase. The training phase takes advantage of a module previously proposed by the authors [12], referred to as the Radar Ground Segmentation

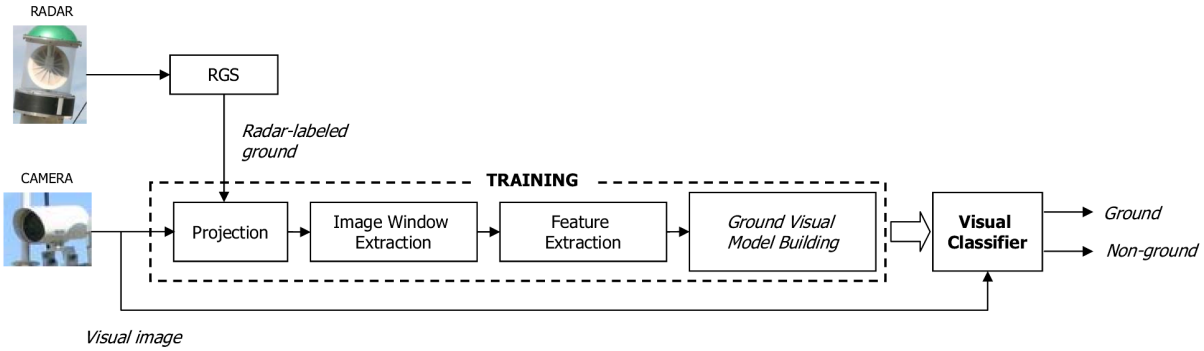
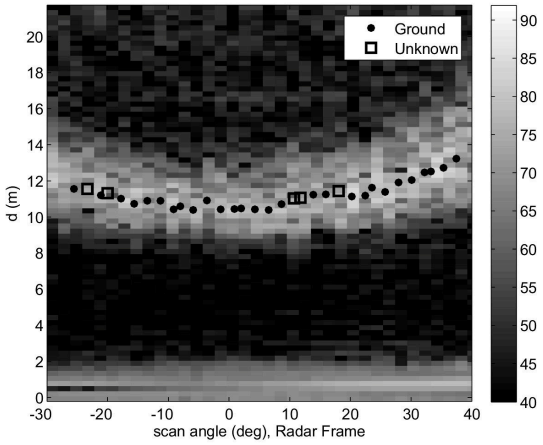
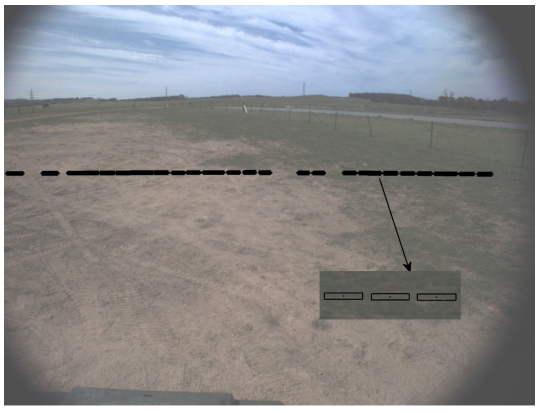


Fig. 2. Architecture of the proposed system. The training stage is supervised by the radar, allowing continuous update of the ground model during robot operation.

(RGS) system. The RGS system can be applied to the radar-generated image of the environment to detect objects belonging to three broad categories, namely *ground*, *non-ground* (i.e., obstacles), or *unknown* (i.e., occluded areas and, more generally, areas for which the RGS system generates a low confidence estimate due to radar misreading or low resolution). Successively, the radar-labeled ground is used to guide the selection of terrain patches in the camera image to construct a visual model of the ground. Specifically, the radar-centered points labeled as ground are projected in the camera image through the camera perspective transformation, and are used to define attention windows. Their associated sub-images are then processed to extract visual features, incorporating the visual appearance of the ground, which is finally modelled as a multivariate Gaussian distribution.



(a)



(b)

Fig. 3. (a) Radar image with overlaid RGS results. Black dots represent ground returns, while outlined squares are used for unknowns. (b) Projections of ground-labeled radar returns in the corresponding camera image with a close up of some attention windows. Visual features extracted from these windows provide the training set to build a visual model of the ground. Note that only the points lying in the field of view common to both sensors are shown.

In the classification phase, the visual ground model is employed to develop a Mahalanobis distance-based one-class classifier for scene segmentation. The proposed approach can be also used to solve radar ambiguities by classifying unknown radar returns, through comparison of the visual feature vectors extracted from unknown labelled visual patches with the ground model. In this sense, the visual classifier serves as a supplement to the radar system to solve uncertain situations. In addition, using the constructed visual model of the ground, the vehicle can perform more complex high-level tasks, including terrain classification and road finding.

A. Radar Ground Segmentation

The performance of a ground classifier is tightly connected with the choice of the model for the training of the system. This is particularly challenging at the start of the vehicle motion, when no prior information is available, and whenever a significant change in the ground properties occurs.

In this research, a self-supervised training approach is proposed using the data obtained from radar mounted on a frame attached to the vehicle's body and tilted forward so that the center of the beam intersects the ground at a look-ahead distance of about 11.4 m in front of the robot. A single sensor sweep outputs a bidimensional intensity graph (radar image) as shown in Fig. 3(a), acquired from a large, relatively flat area. The abscissas in Fig. 3(a) represent the horizontal scanning angle. The ordinates represent the range measured by the sensor.

The RGS system was proved to be effective in performing ground segmentation using a physical model of the ground echo that is compared against a given radar observation to assess the membership confidence to ground, non-ground, and unknown. However, it is worth noting that the ground

(a) (b)

Fig. 4. (a) Observations obtained by the radar at scan #778 before visual classification. Black dots and triangles denote radar-labeled ground and non-ground observations, respectively. Outlined squares denote uncertain measurements. (b) Results after classification of the unknowns using the visual classifier. Outlined circles and triangles denote ground and non-ground, respectively.

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 varies with the scan angle between 6-10 m, thus limiting the radar resolution for segmentation purposes. As an example, the RGS system output is overlaid over the radar intensity image of Fig. 3(a). Ground labels are denoted by black dots, while outlined squares mark uncertain terrain. The corresponding visual image is shown in Fig. 3(b). Note that only the points lying in the field of view common to both sensors (approximately 70 deg horizontally) are shown.

B. Radar-Camera Integration

For each radar scan, the RGS module detects and ranges a set of background points in radar-centered coordinates, which we regard as good estimates of ground.

To perform radar-camera integration, these points are, first, projected in the camera image using the perspective transformation, based on the known intrinsic and extrinsic camera calibration parameters. Then, for each projected point, an attention window is fixed in the camera image. In order to perform a local analysis of the visual characteristics of the ground, we analyze small ground portions of 0.30m×0.30m. At an average distance of 10-12 m, this leads to windows of approximately 35×7 pixels (see Fig. 3(b)). Successively, the image patches associated to the windows are processed to extract visual features and build a training set for the concept of ground.

C. Visual Ground Classifier

The ground visual model building phase is formulated as a one-class classification problem [13]. 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 under-sampled or is difficult to model. Typically, the objective of a one-class classifier is that of constructing a decision boundary that separates the instances of the target class from all other possible objects. In the context of this paper, ground samples constitute the target class, while non-ground samples (i.e., obstacles) are regarded as the outlier class. It is worth noting that in principle, both ground and non-ground samples from RGS may be exploited to train a two-class classifier. Nevertheless, in open rural environments non-ground samples are typically sparse; in addition, the variation of all possible non-ground classes is unlimited. That makes it difficult to model the non-ground class, whereas, although it changes geographically and over time, the ground class is generally less variable than random objects. Furthermore, our objective is that of building 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 ground class. Specifically, we adopt a multivariate Gaussian distribution to model positive ground samples, and we implement a Mahalanobis distance classifier [14].

Let us consider N_G ground patterns. The ground pattern i is represented by its m-dimensional row feature vector f_G^i, with m being the number of feature variables. These vectors constitute the training set X, expressed in the form of a $N_G \times m$ matrix. If we compute the sample mean μ and the sample covariance Σ of the data in X, we can denote the ground model as $M(\mu, \Sigma)$. Then, given a new pattern f, the squared Mahalanobis distance between f and $M(\mu, \Sigma)$ is defined as:

$$d^2 = (f-\mu)\Sigma^{-1}(f-\mu)^T \quad (1)$$

The pattern is an outlier, i.e. it is defined as a non-ground sample, if d^2 is greater than a threshold. The latter is computed as the α-quantile $\chi^2_{m;\alpha}$ of the *chi-square* distribution with m degrees of freedom. Note that, in order to update the ground class during the vehicle motion, the model $M(\mu, \Sigma)$ is continuously rebuilt, always using the ground feature vectors obtained by the most recent radar scans.

III. Experimental Results

In this section, experimental results are presented, to validate our approach for ground segmentation using a millimeter-wave radar and a monocular camera.

The system was tested using the CORD UGV (see Fig. 1).

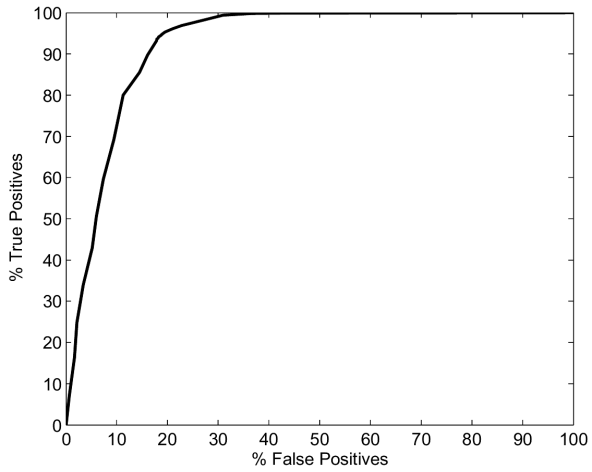


Fig. 5. ROC curve of the visual classifier, formed as the cutoff threshold for terrain detection is adjusted.

The test field was located in a rural environment at the University of Sydney's test facility near Marulan, NSW, Australia [15]. It was mainly composed of a relatively flat ground made of natural terrain with sparse low grass, delimited by fences. A few obstacles were present in the field, including a static car, a trailer, and a metallic shed. During the experiment, the CORD vehicle was remotely controlled to follow an approximately closed-loop path, with an average travel speed of about 0.5 m/s and a maximum speed of 1.5 m/s. Overall, 868 radar images and corresponding visual images were acquired, with a total number of 28,326 observations in the field of view common to both sensors. Of these observations, 23,535 were labeled as ground, 41 as non-ground, and 4,750 as unknown by the RGS module. In order to provide a quantitative evaluation of the system performance, we measured the true positive and false positive rates of the visual classifier for the unknown-labeled observations returned by the RGS module. The ground-truth was constructed manually, yielding to 3,786 ground patterns (the first seven of which remained unclassified, as belonging to the first frames used for initial model building) and 964 non-ground patterns.

In this implementation, we used a four-dimensional feature vector resulting from the concatenation of visual textural descriptors (i.e., contrast and energy), along with colour descriptors (i.e., mean intensity values in the normalized red and green colour planes). More complex visual descriptors may be also used without altering the rest of the algorithm, as long as the hypothesis of normally distributed features holds. The visual classifier was continuously re-trained along the sequence, using the last 100 radar-labeled ground samples, corresponding approximately to a 1.2-sec temporal window. As an example, Fig. 4 shows, for an image of the sequence, the radar output and the result of the classification of the unknown-labeled observations using the visual classifier. RGS results (Fig. 4(a)) are denoted by black dots and triangles for returns labeled as ground and non-ground, respectively, and by outlined squares for returns labeled as unknowns. The output of the visual classification of radar unknowns (Fig. 4(b)) is shown using outlined circles for patterns classified as ground and outlined triangles for patterns classified as non-ground.

In order to establish the optimal threshold value for the Mahalanobis distance-based classifier, we constructed the receiver operating characteristic (ROC) curve of the system, formed as the cutoff value for terrain detection is adjusted by varying α between 0 and 1. The resulting ROC curve is shown in Fig. 5. The vertical axis indicates the true positive rate (i.e., the fraction of ground patches which were correctly classified as ground), while the horizontal axis indicates the false positive rate (i.e., the fraction of non-ground samples which were erroneously classified as ground by the system). Note that random assignment of a patch to the ground class would yield a diagonal line from (0, 0) to (1, 1). We can observe that the point of maximum difference between the true positive rate and the false positive rate is reached approximately at α=0.993, and corresponds to a true positive rate of 94.1% and a false positive rate of 18.3%. For this point, the overall accuracy, i.e. the fraction of correct detections with respect to the total number of classifications is of 91.6%.

The advantage of using an adaptive online learning approach with respect to a batch training system can be shown by computing the rolling average of the true positive rate for the same sequence, with and without ground model update. In the latter case, the ground model was constructed at the beginning of the sequence, and was never updated. If we denote with TP_i the percentage of correct ground identifications at scan i, the rolling average of the true positive rate at scan j can be defined as:

$$\overline{TP}_j = \frac{1}{N} \sum_{i=j-N+1}^{j} TP_i \qquad (2)$$

where N is the size of the average window. The resulting

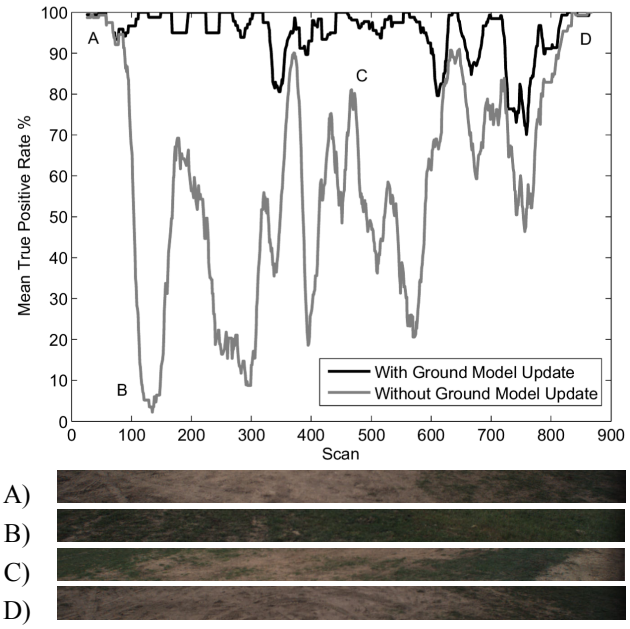


Fig. 6. Rolling average of the true positive rate with (black line) and without (gray line) ground model update. Some samples where significant change in the terrain appearance occurs are indicated, and the pertinent visual images are shown in the lower part of the graph: A-mostly sandy, B-mostly grass, C-sand/grass, D-mostly sandy.

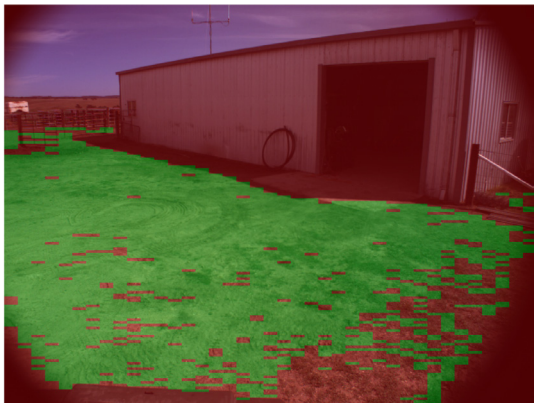


Fig. 7. Result of classification of the entire camera field of view at scan #778. Green pixels denote image patches classified as ground, while red pixels represent non-ground patches.

graph, obtained using α =0.993 and N=20, is shown in Fig. 6. It allows one to highlight the trend of the true positive rate of the classifier, either when the ground model is continuously updated (black line in the figure) or in case of no update (gray line in the figure). We can observe that the true positive rate remains almost constant if the ground model is continuously re-learnt, while it decreases if the model is constructed only once at the start of the vehicle operation due to terrain changes. For instance, a significant reduction of the average true positive rate occurs starting from approximately the 95th scan where the terrain starts changing from sandy (A) to mostly grass (B). In contrast to this, an increment is observed in the last part of the sequence, where we have sand/grass (C) and mostly sandy soil again (D), as the vehicle returns to its starting position.

Once the classifier has been trained, the vision algorithm can be run on the entire field of view of the camera. To this aim, the image is divided in small patches, and individual sub-images are classified as ground or non-ground according to their Mahalanobis distance from the current ground model. The extension of the classification to the entire scene for the sample image of Fig. 4 is displayed in Fig. 7. In this picture, the results of visual classification are overlaid on the original image: each patch is assigned either a red color if classified as non-ground or a green color if classified as ground.

IV. Conclusion

In this paper, we proposed a self-supervised radar-vision classification system that allows an autonomous vehicle, operating in natural terrains, to automatically construct online a visual model of the ground and perform accurate ground segmentation. With respect to radar-vision systems previously developed in literature, our work presents the following main novelties:

1) Use of radar output to fix attention windows in the camera images and extract training data for the concept of ground, instead of detecting and classifying obstacles.

2) Automatic online labeling based on a radar ground segmentation approach prior to image analysis. This avoids time consuming manual labelling to construct the training set. At the same time, no a priori knowledge of the visual terrain appearance is required.

3) Adaptive ground model learning by continuously re-training the classifier using the most recent radar scans. This is particularly useful for long range and long duration navigation applications.

We demonstrated the effectiveness of the proposed approach through field validation. Results were promising, showing overall classification accuracy greater than 90%.

References

[1] A. Sole, O. Mano, G.P. Stein, H. Kumon, Y. Tamatsu, and A. Shashua, "Solid or not solid: vision for radar target validation," in *Proc. IEEE Intelligent Vehicles Symposium*, Parma, Italy, 2004, pp. 819-824.

[2] Z. Ji and D. Prokhorov, "Radar-vision fusion for object classification," in *Proc. 11th International Conference on Information Fusion*, Cologne, Germany, 2008, pp. 1 - 7.

[3] G. Alessandretti, A. Broggi, and P. Cerri, "Vehicle and guard rail detection using radar and vision data fusion," *IEEE Trans. on Intelligent Transportation Systems*, vol. 8, no. 1, pp. 95 – 105, 2007.

[4] R. Grover, G. Brooker, and H.F. Durrant-Whyte, "A low level fusion of millimeter wave radar and night-vision imaging for enhanced characterization of a cluttered environment," in *Proc. Australian Conference on Robotics and Automation*, Sydney, Australia, 2001, pp. 98 - 103.

[5] U. Hofmann, A. Rieder, and E.D. Dickmanns, "Radar and vision data fusion for hybrid adaptive cruise control on highways," in *Proc. International Conference on Computer Vision Systems (ICVS)*, Vancouver, Canada, 2001, pp. 125-138.

[6] B. Steux, C. Laurgeau, L. Salesse, and D. Wautier, "Fade: a vehicle detection and tracking system featuring monocular color vision and radar data fusion," in *Proc. IEEE Intelligent Vehicles Symposium* (IV2002), Versailles, France, 2002, pp. 632-639.

[7] S. Wu, S. Decker, P. Chang, T. Camus, and J. Eledath, "Collision sensing by stereo vision and radar sensor fusion," *IEEE Trans. on Intelligent Transportation Systems*, vol. 10, no. 4, pp. 606 - 614, December 2009.

[8] C.A. Brooks and K. Iagnemma, "Self-supervised terrain classification for planetary rovers," in *Proc. of NASA Science Technology Conference*, Adelphi, Maryland, USA, 2007.

[9] P. Vernaza, B. Taskar, and D.D. Lee, "Online, self-supervised terrain classification via discriminatively trained submodular Markov random fields," in *Proc. IEEE International Conference on Robotics and Automation*, Pasadena, California, 2008, pp. 2750 - 2757.

[10] H. Dahlkamp, A. Kaehler, D. Stavens, S. Thrun, and G. Bradski, "Self-supervised monocular road detection in desert terrain," in *Proc. of Robotics: Science and Systems,* Philadelphia, USA, 2006.

[11] G. Brooker, R. Hennessey, M. Bishop, C. Lobsey, H. Durrant-Whyte, and D. Birch, "High-resolution millimeter-wave radar systems for visualization of unstructured outdoor environments," *Journal of Field Robotics*, vol. 23, no. 10, pp. 891-912, 2006.

[12] G. Reina, J. Underwood, G. Brooker, and H. Durrant-Whyte, "Radar-based perception for autonomous outdoor vehicles," *Journal of Field Robotics*, vol. 28: n/a, doi: 10.1002/rob.20393, 2011.

[13] D.M.J. Tax, "One-Class Classification. Concept Learning in the Absence of Counter Examples," PhD Thesis, Delft University of Technology, Delft, Netherlands, 2001.

[14] E.O. Duda, P.E. Hart, and D.G. Stork. (2001). *Pattern Classification*, 2nd Ed., Wiley.

[15] T. Peynot, S. Scheding and S. Terho, "The Marulan Data Sets: Multi-Sensor Perception in Natural Environment with Challenging Conditions," *International Journal of Robotics Research* (IJRR), vol. 29, no. 13, pp. 1602-1607, November 2010.