International Journal of Electrical, Electronic and Communication Sciences
ISSN: 2517-9438
Vol:11, No:6, 2017

Terrain Classification for Ground Robots Based on Acoustic Features

Bernd Kiefer, Abraham Gebru Tesfay, Dietrich Klakow

***Abstract*—**The motivation of our work is to detect different terrain types traversed by a robot based on acoustic data from the robot-terrain interaction. Different acoustic features and classifiers were investigated, such as Mel-frequency cepstral coefficient and Gamma-tone frequency cepstral coefficient for the feature extraction, and Gaussian mixture model and Feed forward neural network for the classification. We analyze the system's performance by comparing our proposed techniques with some other features surveyed from distinct related works. We achieve precision and recall values between 87% and 100% per class, and an average accuracy at 95.2%. We also study the effect of varying audio chunk size in the application phase of the models and find only a mild impact on performance.

***Keywords*—**Terrain classification, acoustic features, autonomous robots, feature extraction.

I. Introduction

WHEN using autonomous ground robots, e.g, in disaster area exploration tasks, detecting the terrain type the robot is currently operating on has many useful applications, such as the adaptation of motion control, avoiding hazardous areas, and also autonomous report generation for a human surveillance person to support decision making.

In the robotics literature, there are several approaches to terrain detection based on a variety of sensors, such as motor slip measurement [1], velocity and acceleration features [2], [3], visual detection [4]-[6], vibration [7]-[9], sound [10], or a mixture of several sensors [11], [12]. Many of the aforementioned approaches are unsupervised and aim at automatically acquiring models to improve, e.g., the accurate navigation of the robot, or its autonomy.

Not all sensors are reliable in every situation. Visual sensors are quite sensitive to different lighting conditions, fog, vegetation, or other factors that change or obscure the appearance of objects. Exploiting as much information sources as possible will therefore result in a more robust and usable system.

The TRADR project [13], which gave rise to this study, works on robots for disaster response scenarios. The goal is to create persistent environment models of the disaster area, which are useful to robots and humans. One aspect in this mission is the description of the area that a robot explores.

Aside from the use of vibration, there is not that much work exploiting sound for the recognition of different terrain types. In this paper, we are studying the recognition of the terrain based on the sound that a robot generates when moving, and use a supervised approach with humanly labelled data for training the classifiers for five different terrain types. These choices were made because the module was primarily intended for the generation of human-readable reports. While this may limit the use of the method for sensor fusion or the avoidance of hazardous terrain, the results strongly suggest that the sound features used here may also be useful for other applications.

Bernd Kiefer is with the German Research Center for Artificial Intelligence (DFKI), Saarbrücken.

Abraham Gebru Tesfay is with the German Research Center for Artificial Intelligence (DFKI), Kaiserslautern (e-mail: abraham$_g$ebru.tesfay@dfki.de).

Dietrich Klakow is with the Saarland University, Saarbrücken.

II. Hardware Setup and Data Collection

The robot used in our experiments is a version of Bluebotics' Absolem platform [18], with two side tracks and front and rear flippers on each side for operation in difficult terrain, to surmount uneven surfaces or small crevices, and also climb stairs. The robot is equipped with several sensors, such as a laser scanner and omnicam, but we will not go into details because they are not used in the experiments described here.

For the sound recording, we mounted a laptop onto the robot, equipped it with a small USB sound card (a Sennheiser USB adapter) and used an onmidirectional electret capacitor microphone (Elecom table microphone), which was placed next to the left rear motor and relatively close to the ground, as shown in Fig. 1. The recording was done with standard Linux software (audacity). All audio was recorded in 16 bit mono pulse-code modulation (PCM) with a sampling rate of 11025Hz.

Fig. 1 Microphone Setup on the Robot

The data collection was performed at the campus of the Saarland University, Saarbrücken, piloting the robot manually with a wireless joystick. The recordings were done in different parts of the campus, with different locations for each terrain type. The robot ran with varying speed and also performed several turning manoeuvres, in flat as well as ascending and descending places.

We distinguish four outdoor terrain types (gravel, pavement, grass and sand) and carpet (indoor), as shown in Fig. 2. The terrain also varied inside the classes, e.g., from dense structured gravel to dirt road with wet crushed stones mixed with varying level of mud, or from wholesome grassland to grass covered with leaves or mixed with stones or sandy patches.

The recordings for the sand class had to be stopped prematurely because the sand entered between the belt of the main track and the driving wheel and dislocated the belt. Therefore, the amount of data available for the sand class is much lower than for the other terrain types. For all types except sand, we recorded around 24 minutes of audio, for sand, only a bit more than two minutes.

(a) Gravel (b) Grass

(c) Gravel in hilly surface (d) Grass with leaves

(e) Pavement (f) Sand

Fig. 2 Examples of various environments for the four outdoor terrain types

III. Feature Extraction and Classifiers

A. Feature Extraction

We divided the original files into chunks of four seconds length, resulting in 315 chunks for each class except for the *sand* class where we ended up with only 31 chunks. To evaluate the spectrogram for each chunk, we use frames of 256 samples with 60% overlap, starting the next frame 100 samples after the current one. This amounts to a frame duration of about 25 ms each. We then compute a FFT with 256 coefficients, applying a Hamming window beforehand. In order to decrease the execution time, we only keep the first 128 coefficients. Finally, the power spectrum is computed by squaring the absolute value of the FFT coefficients.

To be able to compare our results with that of [10], we used a 6D feature vector proposed in [14], comprised of zero crossing rate (ZCR), short time energy (STE), energy entropy, spectral centroid, spectral rolloff, and spectral flux, and a 9D feature vector by adding three more shape features, i.e., the spectral moments (standard deviation, skewness and kurtosis) [15]. They use a support vector machine (SVM) classifier and a k nearest neighbour classifier to separate benign interactions of the robot, like driving on grass, pavement or gravel from hazardous ones like driving into water, hitting hard objects or driving on slippery ground. When looking at all six interactions, their average accuracy is 92%. Since they are more interested in the hazardous interactions, they collapsed the benign classes into one class, and achieve an average accuracy of 96% for the resulting four classes.

Furthermore, we implemented a Gamma-tone frequency cepstral coefficients (GFCC) and a Mel-frequency cepstral coefficients (MFCC) feature extraction. For these, after applying the Gamma-tone resp. Mel filter banks, an additional discrete cosine transform is applied to convert the spectrum back into the time domain. In both cases, the lowest coefficient is dropped because it does not contain useful information.

To build the GFCC models, we use 23 linear phase Gammatone filters of order 4 and use all but the first for the classification task. Lastly, we created a MFCC feature vector using 26 filters, with the first and last triangle filters are centred at 100Hz and 5512.5Hz, respectively. For classification, we only used the coefficients 2–23 from the DCT output.

B. Classifiers

We also experimented with two different classifiers, namely a gaussian mixture model (GMM), and a feed-forward neural net (FFNN).

1) GMM: We trained our system using the GMM model and training algorithm in Matlab's *Statistics Toolbox*™ software using the *gmdistribution* class. This class fits the data using an expectation maximization algorithm. We conducted experiments with the different numbers of Gaussian components, getting best results with a mixture of at most three Gaussians.

In rare cases, the *gmdistribution* converged to a solution with an ill-conditioned singular or close-singular covariance matrix for one or more gaussian components, which could be overcome by using slightly different initial values.

A separate GMM is trained for every class; in the application phase, the extracted features of the test item are applied to every model, which returns a probability, and the one with the highest confidence is then taken as result. Comparing the probabilities the models return can provide a confidence measure for the classification, an added value of the GMM method.

2) FFNN: We use the Matlab R2013a neural network pattern recognition toolbox for classification. The number of inputs corresponds to the number of features, in the output layer there is one neuron for each class. We used a two layer FFNN with 11 neurons in the hidden layer, with a tanh sigmoid activation function. The training algorithm used was scaled conjugate gradient back-propagation *(Trainscg)*, with a mean square error training criterion.

IV. Experimental Results

A. Using GMM as Classifier

In the first experiment, we evaluated the feature sets using a GMM classifier with a 5-fold cross-validation, splitting the 315 audio chunks for the four major classes into a training set of 140 items and a test set of 175 items. For the *sand* class, we used 20 training and 11 test items.

We trained GMMs using the different feature sets described in the previous section. The optimal number of Gaussians for the MFCC, 6D and 9D features was two, for the GFCC, a single Gaussian gave the best results. The mean results are summarized in Table I.

TABLE I
Precision and Recall for the GMM Classifier with Different Feature Sets

Precision	6D	9D	GFCC	MFCC
grass	76.0	72.3	78.0	**99.4**
pavement	64.9	65.9	82.5	**92.4**
gravel	79.1	62.8	80.2	**100**
carpet	56.2	66.5	67.6	**100**
sand	**47.6**	**47.6**	12.7	35.7
Recall	6D	9D	GFCC	MFCC
grass	73.3	73.1	89.1	**100**
pavement	65.9	55.4	64.9	**100**
gravel	30.1	46.2	59.4	**81.8**
carpet	88.3	89.7	70.3	**100**
sand	**90.1**	**90.1**	72.7	90.0

Due to the little amount of data available for the sand class, the results for sand are not very reliable, and that the 6D and 9D vectors outperform the MFCC features for this class might be due to an overfitting or ceiling effect. We report this class only for sake of completeness, and will drop it in the forthcoming experiments.

It is also possible that the 6D and 9D features would generally profit from audio preprocessing such as noise reduction, or larger FFT window sizes, but we did not test this due to time limitations. For the best feature set, the MFCC, Table II shows the detailed confusion matrix, which may give an insight into which classes are harder to separate.

TABLE II
MFCC Confusion Matrix with 95.2% Accuracy. These are Absolute Numbers of Test Samples Classified Correct or Wrong

		Actual label				
		Grass	Pavement	Gravel	Carpet	Sand
Predicted	Grass	175	0	1	0	0
	Pavement	0	175	14	0	1
	Gravel	0	0	142	0	0
	Carpet	0	0	0	175	0
	Sand	0	0	18	0	10

1) Varying Training Data Size: To check the influence of the training data size on the performance of the classification, we conducted two experiments with GFCC and MFCC and GMM as classifier. In one the test data size was fixed to 175 samples, in the other we took all data that was not used for training to test the model.

The number of training samples varied from 20 to 140. The results are shown in Fig. 3. When using the MFCC features, the accuracy varies from 88.9% to 95.6%, for the GFCC features from 49.6% to 71%.

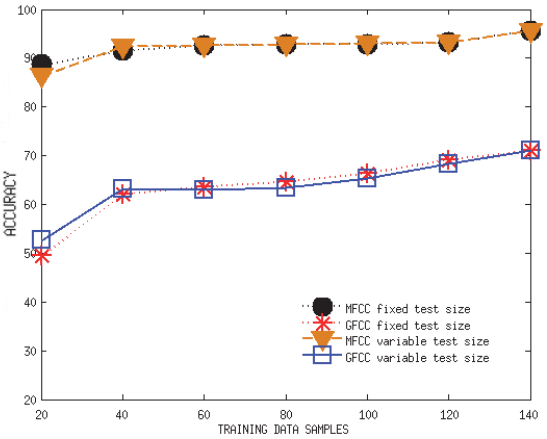

Fig. 3 Accuracy with fixed and variable test data size. The x axis shows the number of training samples, the y axis the average accuracy

The second experiment reveals change in accuracy perceived by varying both training data size and testing data size. In this experiment, we used the data removed from training for test. As can be seen from Fig. 3, in MFCC the accuracy varies from 86.4% to 95.6%, while in GFCC the variation is significant which ranges from 52.8% to 71%. We can deduce that, in MFCC more than 60 training samples and in GFCC more than 120 training samples are enough to achieve a satisfactory performance.

2) Varying Duration of Test Samples: Since in an application scenario fast response times decide about the usability of an approach, we also studied the effect of reducing the duration of the audio chunks on the recognition accuracy. Shorter test chunks mean less latency, which can be crucial when the robot enters hazardous terrain. We took the audio test set and created chunks of one and two seconds from the four second chunks.

TABLE III
Varying Test Audio Length for GFCC and MFCC Features and GMM Classifier

	1 sec	2 sec	4 sec
GFCC	65.6	69	71
MFCC	93	94.1	95.6

Table III gives the result for the MFCC and GFCC features with a GMM classifier. Especially for the MFCC features, the drop in accuracy is acceptable in the light of a four times faster response time.

B. FFNN Classifier

For the neural network experiments, we reduced the test sample size to one second. For the GFCC and MFCC feature sets, we also tried two different ways to feed the set of feature vectors to the network. Firstly, we averaged over the vectors of all frames of a sample and used that as input, and secondly, we appended all vectors into one large input vector, creating an input vector of size `nr_features` × `nr_frames`, resulting

in $12 \times 109 = 1308$ inputs. As already said before, the network uses one hidden layer of 11 nodes, four output nodes (for the classes), a tanh sigmoid transfer function and mean squared error as training criterion for the scaled conjugate gradient descent learning method. In the experiment 300 samples for each class were used, which is 1200 samples altogether for all classes. Of these, 70% were taken for training, 15% for testing and 15% for validation. The results of this experiments are summarized in Table IV.

TABLE IV
PRECISION AND RECALL FOR THE FFNN WITH DIFFERENT FEATURE SETS. GFAVG AND MFAVG ARE THE AVERAGED VECTORS OF GFCC AND MFCC, RESPECTIVELY

Precision	6D	9D	GFCC	MFCC	GFavg	MFavg
grass	88.24	87.76	60.78	**97.87**	68.00	87.23
pavement	61.90	71.79	89.13	91.67	91.11	**100**
gravel	75.00	64.71	55.56	**92.11**	77.50	90.91
carpet	97.83	94.83	78.57	93.62	95.56	**100**
Recall	6D	9D	GFCC	MFCC	GFavg	MFavg
grass	90.91	79.63	73.81	92.00	85.00	**93.18**
pavement	68.42	82.35	87.23	**100**	89.13	97.22
gravel	69.86	61.11	34.88	83.33	64.58	**86.96**
carpet	95.74	98.21	91.67	**100**	93.48	**100**

The MFCC features gave the best accuracy also in case of the neural network classifier, with 93.9% for the large input vector, and 94.4% for the averaged input vector.

To compare these results with the GMM classifier, we repeated the GMM / MFCC experiment with audio chunks of one second, and obtained an accuracy of 93.03%. Table V shows the confusion matrices for the neural network with averaged MFCC features and the GMM classifier with MFCC. The number of samples for MFCC & GMM differ from those in Table II because the sample duration used in this experiment is one second instead of four, giving us more training and test samples.

Still, due to differing test and training data sizes, the result are not absolutely comparable. Which method performs better in practice certainly has to be explored.

TABLE V
CONFUSION MATRICES FOR THE BEST NEURAL NETWORK AND GMM CLASSIFIER RESULTS IN ABSOLUTE NUMBER OF SAMPLES

		Actual label			
MFCC avg. & FFNN		Grass	Pavement	Gravel	Carpet
Predicted	Grass	41	0	3	0
	Pavement	0	35	1	0
	Gravel	6	0	40	0
	Carpet	0	0	0	54
MFCC & GMM					
Predicted	Grass	609	3	12	0
	Pavement	3	625	46	0
	Gravel	3	0	476	0
	Carpet	0	0	1	628

V. CONCLUSION AND OUTLOOK

We have studied the usefulness of audio features for supervised human-labelled terrain recognition on ground robots. We tried several feature sets, including the 6D and 9D features proposed in a very similar setting [10], and two classifier methods, namely Gaussian mixture models and feed-forward neural networks. We showed that, at least in our setting, the very traditional approach using MFCC and GMM performed astonishingly well. Another interesting finding was that GMM and FFNN performed comparably for the MFCC, but that the 6D and 9D features performed much better with the FFNN classifier. An advantage of the GMM over the FFNN classifier, although its performance is slightly worse, is the comparability of the probabilities of the single outcomes, making it possible to judge if the given result is more or less reliable. The recordings and Matlab code we used in our experiments are available for download [16].

It has to be noted that it is quite likely that our method works so well because of the kind of robot we used. A wheeled robot will certainly produce less distinctive sound patterns given different terrain types. What is still missing are field recordings to confirm the results of the experimental recordings. Especially the effect of a lot of different driving speeds and the differing motion patterns of an autonomous robot on the classification is still untested.

We did not take special measures to record the data in a quiet environment; there is some background noise on the recordings, but it is so much lower than the noise coming from the motors due to the carefully chosen position of the microphone, that it does not seem to pose a problem for the classifier. Also the noise from cooling fans and other sound sources of the robot is insignificant. Thus, we did not see the need for elaborate noise reduction techniques.

In the future, we would like to study the fusion of other sensors with the audio features, e.g., to improve odometry, as in [17]. This would also be of importance because the audio can only capture the *current* terrain, which is not very helpful to completely avoid hazardous areas.

Also, there is a wide range of research possibilities from multi-channel recording to noise reduction which may become necessary to achieve good results if the number of classes increases or the mechanical setting is different.

ACKNOWLEDGMENT

This work was supported by the EU-funded projects TRADR (FP 7-ICT-609763) and ALIZ-E (FP 7-ICT-248116).

REFERENCES

[1] Debangshu Sadhukhan, Carl Moore, and Emmanuel Collins, "Terrain estimation using internal sensors," in *Proceedings of the 10th IASTED International Conference on Robotics and Applications (RA)*, 2004.

[2] David Tick, Tauhidur Rahman, Carlos Busso, and Nicholas Gans, "Indoor robotic terrain classification via angular velocity based hierarchical classifier selection," in *Robotics and Automation (ICRA), 2012 IEEE International Conference on*. IEEE, 2012, pp. 3594–3600.

[3] Eric Coyle, Emmanuel G Collins Jr, and Rodney G Roberts, "Speed independent terrain classification using singular value decomposition interpolation," in *Robotics and Automation (ICRA), 2011 IEEE International Conference on*. IEEE, 2011, pp. 4014–4019.

[4] Ayanna Howard and Homayoun Seraji, "Vision-based terrain characterization and traversability assessment," *Journal of Robotic Systems*, vol. 18, no. 10, pp. 577–587, 2001.

[5] Jann Poppinga, Andreas Birk, and Kaustubh Pathak, "Hough based terrain classification for realtime detection of drivable ground," *Journal of Field Robotics*, vol. 25, no. 1, pp. 67, 2008.

[6] Liang Lu, Camilo Ordonez, Emmanuel G Collins Jr, and Edmond M DuPont, "Terrain surface classification for autonomous ground vehicles using a 2d laser stripe-based structured light sensor," in *Intelligent Robots and Systems, 2009. IROS 2009. IEEE/RSJ International Conference on*. IEEE, 2009, pp. 2174–2181.
[7] Christopher Brooks, Karl Iagnemma, et al., "Vibration-based terrain classification for planetary exploration rovers," *Robotics, IEEE Transactions on*, vol. 21, no. 6, pp. 1185–1191, 2005.
[8] Christian Weiss, Holger Fröhlich, and Andreas Zell, "Vibration-based terrain classification using support vector machines," in *Intelligent Robots and Systems, 2006 IEEE/RSJ International Conference on*. IEEE, 2006, pp. 4429–4434.
[9] Chris C Ward and Karl Iagnemma, "Speed-independent vibration-based terrain classification for passenger vehicles," *Vehicle System Dynamics*, vol. 47, no. 9, pp. 1095–1113, 2009.
[10] Jacqueline Libby and Anthony J. Stentz, "Using sound to classify vehicle-terrain interactions in outdoor environments," in *Robotics and Automation (ICRA), 2012 IEEE International Conference on*. IEEE, 2012, pp. 3559–3566.
[11] Karl D Iagnemma and Steven Dubowsky, "Terrain estimation for high-speed rough-terrain autonomous vehicle navigation," in *AeroSense 2002*. International Society for Optics and Photonics, 2002, pp. 256–266.
[12] Karl Iagnemma, Hassan Shibly, and Steven Dubowsky, "On-line terrain parameter estimation for planetary rovers," in *Robotics and Automation, 2002. Proceedings. ICRA'02. IEEE International Conference on*. IEEE, 2002, vol. 3, pp. 3142–3147.
[13] Ivana Kruijff-Korbayová, Francis Colas, Mario Gianni, Fiora Pirri, Joachim de Greeff, Koen Hindriks, Mark Neerincx, Petter Ögren, Tomáš Svoboda, and Rainer Worst, "TRADR project: Long-term human-robot teaming for robot assisted disaster response," *KI-Künstliche Intelligenz*, pp. 1–9, 2015.
[14] Theodoros Giannakopoulos, Dimitrios Kosmopoulos, Andreas Aristidou, and Sergios Theodoridis, "Violence content classification using audio features," in *Advances in Artificial Intelligence*, pp. 502–507. Springer, 2006.
[15] Mark C Wellman, Nino Srour, and David B Hillis, "Feature extraction and fusion of acoustic and seismic sensors for target identification," in *AeroSense'97*. International Society for Optics and Photonics, 1997, pp. 139–145.
[16] Abraham Gebru Tesfay, "Audio data collection for terrain classification from sound," Available for Download under http://ox6.dfki.de/publications/infostore/10/Bernd%20Kiefer?secret=694424e304f6dbb14f3b1eaff75b9e83.
[17] Michal Reinstein, Vladimir Kubelka, and Karsten Zimmermann, "Terrain adaptive odometry for mobile skid-steer robots," in *Robotics and automation (icra), 2013 ieee international conference on*. IEEE, 2013, pp. 4706–4711.
[18] Absolem surveillance and rescue, Available: http://www.bluebotics.com/mobile-robotics/absolem/ (Accessed 27 March 2017).

Bernd Kiefer Bernd Kiefer is a Senior Software Engineer and Researcher in the Multilingual Technology Group of the German Institute for Artificial Intelligence in Saarbrücken. His main research interest are the efficient processing of natural language, and conversational systems for multi-modal Human-Robot interaction.

Abraham Gebru Tesfay studied Bsc. in electonics and communication engineering at Mekelle institute of technology (MIT), Ethiopia and received his MSc. in computer science from Saarland university, Germany. His research interests include wireless industrial communication, telecommunication, computer networks and communication. Currently working as a researcher in the Intelligente Netze group at German research center for artificial intelligence (DFKI), Germany.

Dietrich Klakow Prof. Klakow, born 1966 in Nürnberg, studied Physics from 1987 until 1991 at the Universities of Erlangen and York. After a one year research visit to the USA he completed his PhD at the University of Erlangen in 1994. For the next one and a half years he did a post doc at the Weizmann-Institute in Israel. In 1996 he changed to the area of speech and language research and joined the Philips research lab, in the beginning as researcher and two years later as project and team manager. Together with his team he worked on new algorithms for speech recognition. Later his team explored topics in the fields of "dialog systems for the living room", "access to content" and "computer vision". In addition to his work at Philips, he was holding a lecturer position at Aachen University since 1999. Since May 2003 he is professor at Saarland University where he builds up a new research team which deals with algorithms for the human machine interaction for example robust far field speech recognition.