Assessment of Deep Convolutional Neural Networks for Road Surface Classification

Marcus Nolte, Nikita Kister and Markus Maurer
Institute of Control Engineering
Technische Universität Braunschweig
Braunschweig, Germany
Email: {nolte, maurer}@ifr.ing.tu-bs.de

***Abstract*—When parameterizing vehicle control algorithms for stability or trajectory control, the road-tire friction coefficient is an essential model parameter when it comes to control performance. One major impact on the friction coefficient is the condition of the road surface. A camera-based, forward-looking classification of the road-surface helps enabling an early parametrization of vehicle control algorithms. In this paper, we train and compare two different Deep Convolutional Neural Network models, regarding their application for road friction estimation and describe the challenges for training the classifier in terms of available training data and the construction of suitable datasets.**

I. Introduction

Systems for vehicle dynamics control have been implemented in series vehicles for several decades. A central challenge for the implementation of well-performing control algorithms is the estimation of the road-tire friction coefficient μ which models the maximal adhesive force between the vehicle's tires and the road surface. While its exact value depends on a variety of factors, such as tire and road temperatures and the composition of the tire, the road surface condition has major impact on the maximal transmittable drive or brake force. Thus proper estimation of the friction coefficient is a widely discussed topic in the field of vehicle dynamics.

Many presented approaches (such as [1], [2]) are of reactive nature, which means that e.g. the current measured vehicle dynamics are utilized in observer-based systems to estimate the friction coefficient. An alternative for reactive estimation is the utilization of sensors under the vehicle (microphones, radar, optical sensors) which are used for recording the road surface under the vehicle. While such reactive approaches have been shown to increase control performance [2], predictive approaches promise additional benefits for control performance, as a look-ahead estimation allows an early adoption of control algorithms to upcoming road conditions. Also when going beyond vehicle control and regarding trajectory planning for automated vehicles, knowledge about the road conditions in front of the vehicle is beneficial, because the gained knowledge allows an adoption of planning strategies e.g. when approaching wet or snowy road patches.

While lidar and radar sensors allow detection of wet surfaces, due to different reflectivity, camera images provide high-resolution texture information. Texture information does not only allow the detection of wet surfaces, but also allows to differentiate e.g. between concrete and cobblestone roads. This additional information has already been exploited for predictive estimation of the road friction coefficient [3]–[5]. As Deep Convolutional Neural Networks (CNNs) have been successfully applied to different classification tasks, also with applications in the field of automated driving, it seems promising to use a CNN-based approach for surface classification.

However, the performance of learned classifiers heavily relies on the design of training data. Most available datasets are recorded in dry conditions in inner cities. Hence, the datasets provide unbalanced class information for training a CNN for the given task. We create a mixed dataset from publicly available datasets for automated driving (KITTI [6], Oxford Robocar Dataset [7]), own recorded data from the project *Stadtpilot* [8] as well as images from datasets not particularly designed for automated driving ([7], [9], [10] as well as images from web search. Based on this mixed dataset, this paper presents two different convolutional network architectures based on ResNet50 [11] and InceptionNetV3 [12] to differentiate six classes of road surface conditions. The results acquired from both networks are discussed with respect to the design of the training dataset.

The remainder of the paper is organized as follows: section II summarizes previous work on predictive and reactive μ-estimation. section III describes the challenge of creating balanced datasets for training a CNN for the given task. After describing the general approach for classification in section IV, section V describes the training parameters for the evaluated network architectures. The obtained results are discussed in section VI before concluding with an outlook on future applications for the presented method (section VII).

II. Related Work

Regarding related work, Khaleghian *et al.* present a literature review about recent publications for road friction estimation. For this paper, we will concentrate on publications discussing camera-based surface classification and related learning-based applications.

A combined predictive approach for road surface classification from audio and video data is presented in [3]. They use luminance-based co-occurrence matrices to differentiate texture properties in parts of the image. They differentiate six different classes (dry, humid, and wet asphalt, and cobblestone, respectively) of road surfaces in daylight and nighttime conditions. The image-based classification is fused with audio information, which is processed to extract characteristic frequencies of the different road surfaces. Unfortunately, no information about the performance of both methods is provided.

Omer and Fu present a pure image-based approach for detecting snow-covered roads [4]. A region of interest in front of the vehicle is used to train a support vector machine with partial image histograms as features. Using geolocation information as additional features, an average classification accuracy of 85 % could be achieved.

Qian *et al.* evaluate different learning-based approaches for surface classification in a dynamic region of interest [5]. Evaluated features were MR8 [14] as well as SURF features [15]. For classification, Fisher Vectors were compared to Bag of Visual Words as well as a Texton dictionary. The classifiers were trained to perform binary classification (asphalt vs. snow- or ice-covered road) over three-class (dry vs. wet asphalt vs. snow- or ice-covered road) to five-class classification (dry vs. wet asphalt vs. snow covered vs. snow packed). MR8 features with a bag of words classifier provided the best results with 98 % average classification accuracy for the binary classification problem with a manually defined region of interest. For the five-class problem, classification accuracy dropped to 62 %.

Valada and Burgard present an audio-based approach and train a convolutional neural network with recurrent long short-term memory (LSTM) units to differentiate nine classes (asphalt, mowed grass, high grass, paving, cobblestone, dirt, wood, linoleum, and carpet) [1]. Input data for the convolutional layers are spectrograms extracted via Short Term Fourier Transform. The approach reaches an average classification accuracy of 97.52 % (CNN only) and 98.67 % (CNN+LSTM), respectively

III. Challenges Regarding Available Datasets

A challenge for training deep neural networks is the availability of suitable, annotated training data. One challenge for training neural networks for classification tasks is the class imbalance problem caused by over-represented classes (majority classes) and under-represented classes (minority classes) in a dataset: If single classes dominate a training set or single classes are only represented by a small number of samples, classification performance can degrade significantly [16].

For the application of deep convolutional networks to road surface classification this has the following consequences: While there are many datasets available for general image classification (ImageNet [17]) or autonomous driving in general, such as KITTI [6], a specific dataset for road surface classification is not available. Resorting to those general datasets for automated driving results in heavily imbalanced datasets, as the majority of images was recorded in sunny or overcast weather, for reasons of better illumination and less optical obstructions due to rain on the windscreen. Furthermore, the majority of recorded road surfaces is flat asphalt, while surfaces such as dirt road or sand are not represented, as they do not appear on city roads.

TABLE I: Available classes in the individual datasets. Numbers in parentheses denote total number of selected samples.

	asphalt (10273)	dirt (8547)	grass (2887)	wet asphalt (3668)	cobble-stone (1082)	snow (3075)
Robocar	X		X	X		X
Stadtpilot	X			X	X	
NREC		X	X			
New College	X	X	X			
Giusti et al.	X	X	X		X	X
KIITI	X		X		X	

A. Composition of Dataset

For the selection of suitable training data we thus looked at a variety of available datasets which should provide a more balanced set of images for surface types as a whole. In addition, composing training data from multiple datasets has the advantage of covering several different cameras which helps avoiding learning features specific to the camera-setup of a single research vehicle. One restriction we applied was that the perspective towards the surface should be vaguely similar to the perspective of a windscreen mounted camera, to avoid applying artificial distortion of the chosen images. For the composition of the dataset we used images from the following datasets:

- NREC Human Detection & Tracking in Agriculture [18]
- KITTI Vision Benchmark Suite [6]
- Oxford Robocar Dataset [7]
- New College Vision and Laser Data Set [10]
- Imageset published by [9]
- Image sequences captured in the Stadtpilot project by our research vehicle *Leonie* [8]

The obtained classes available in each individual dataset are presented in Table I.

Analyzing these datasets and comparing the majority class asphalt with the minority class cobblestone yields an imbalance ratio of 10:1: The class *asphalt* consists of over 10000 images, while the class *cobblestone* is represented in just over 1300 images.

To counteract the imbalance, instead of applying over or under sampling, we added further images from Google image search, following the example of [19] for fine grained image classification.

B. Selection of Test and Training Data

All used datasets provide frame sequences rather than a random collection of independently recorded frames. Thus the road conditions vary only slightly between frames from a single sequence. When dividing the selected images into test and training sets, we did not only split single sequences, but also selected images from different sequences for test where possible.

The finally used test set consisted of 300 images per class. The remaining images were used for training, building three different training sets.

A first set only consisted of the images from the datasets mentioned above. To create a balanced set, 700 images per class were chosen randomly. 300 images were used for validation.

For the second dataset, the classes *cobblestone* and *wet asphalt* were extended with images from Google image search as mentioned earlier. The class *wet asphalt* was available in the fewest sequences, while the class *cobblestone* had the lowest number of samples. As the image search resulted in too few usable samples (*grass*) or a sufficient number of images was available (*asphalt*), only the first two classes were extended. Using the Google image search, the class *cobblestone* was extended such that each class consisted of 2500 images. The class *wet asphalt* was extended to increase variation of images within the class. The training set was thus more than doubled. 500 images were selected for validation.

For a third dataset all classes from the basic dataset were extended with 300 images from Google image search, which corresponds to the number of returned usable images for the class *grass*.

In order to overcome the issue of lacking variation between consecutive frames in the sequences, the used sequences were subsampled, using only every n^{th} frame, with n depending on the length of the sequence.

IV. Approach for Surface Classification

In order to classify the road conditions in front of the vehicle, several strategies were evaluated, including running the classification on the whole image and selecting regions of interest. Performing the classification task on the whole image provides additional information about the environment, such as light conditions and the sky. However, evaluations showed that this approach resulted in severe overfitting, providing a validation accuracy of 80 %. Therefore, classification was performed on a region of interest (cf. Fig. 1) which increased validation accuracy to over 90 % as will be presented in the results section.

As the position of the road differs in each of the chosen data sets due to different cameras with different fields of view, the position of the region of interest was defined individually for each dataset. The extracted texture patches were resized to 224 x 224 px for training and classification.

V. Training Parameters

As mentioned above, we evaluated ResNet50 and InceptionV3 for the classification task. Both architectures were initialized with pre-trained weights from the ImageNet dataset and trained using cross-entropy as a cost function minimized by stochastic gradient descent. Batch normalization was applied. The initial learning rates for both architectures were set to $3 \cdot 10^{-5}$ in order to protect the pre-trained weights. Training was performed with a batch size of 48. Analysis of the validation accuracy showed no significant gain after five epochs, thus early stopping was applied in order to avoid overfitting

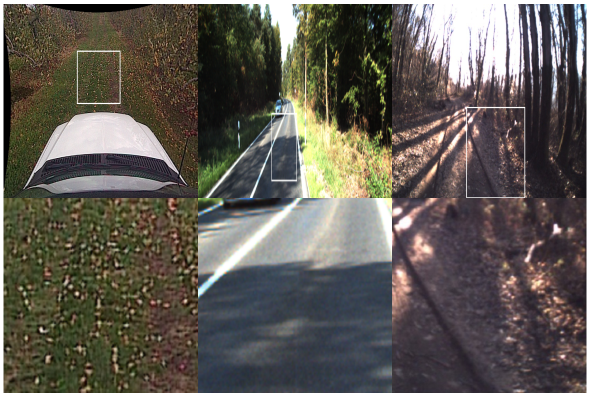


Fig. 1: Some example images for the choice of the region of interest. The cropped images were resized to 224 by 224 px for the classification.

In order to account for the vehicle's motion and the resulting changes of perspective towards the road surface, we also applied data augmentation for each batch. For this purpose the texture patches were mirrored horizontally, randomly rotated in an interval of ±40° and scaled by a random factor of 0.9 to 1.1.

As the regions of interest can contain ambiguous texture information, e.g. if multiple surface types are visible in the patch, the selected labels can be incorrect to a certain degree. For this reason we applied label smoothing with a factor of 0.1.

VI. Results

A. Training & Classification Performance

This section presents an overview about the achieved training and classification results with both implemented architectures on the three datasets (basic, image search augmentation for two classes, augmentation for all classes) described above.

Considering training performance, the InceptionV3 model terminates after seven (second training dataset) to ten (first training dataset) epochs. The maximum validation accuracy is reached after the third epoch as shown in Figure 2 (left hand side). The average validation accuracies of the models trained on the basic dataset and the second dataset are comparable.

Extending the basic dataset with images from Google image search for all classes leads to a decrease of ≈1.5 % in validation accuracy on training data. The extension of the training dataset does not impact the duration of the training. The inference run on an NVIDIA Titan X GPU takes 153 ms in average.

When evaluating the performance on the test dataset, the InceptionV3 architecture behaves differently: Training the model on the first and second dataset resulted in a comparable test accuracy of 90 %. Extending all classes with images from image search, however, resulted in a an test accuracy of only 84 %. The behavior of the model provides a hint, that the additional variation of training data does not provide any benefit. In contrary, the network starts to overfit

Training of the ResNet50 architecture takes longer than the training of the InceptionV3 model: Training terminates after ten (basic dataset) to twenty epochs (second and third dataset). In contrast to the InceptionV3 architecture, adding images from Google image search, speeds up the training process (cf. Fig.2, right hand side).

The ResNet model trained on the first dataset achieved a lower test accuracy on the test dataset (80 %) than the corresponding InceptionV3 model. However, the partial addition of images from google image search increased the test accuracy by 4 % to an overall 92 %.

The ResNet50 architecture exposes the same over-fitting behavior as InceptionV3 with performance decreased to 84 % when the basic dataset is extended with images for each class. Inference for ResNet50 takes 94 ms on the NVIDIA Titan X.

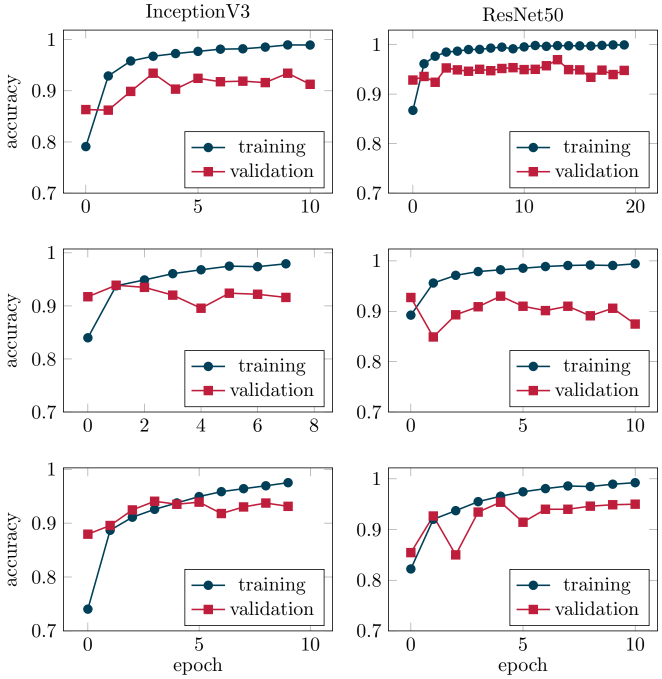
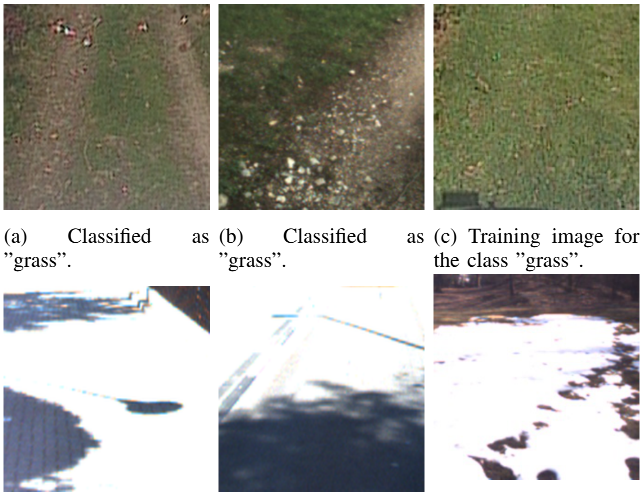


Fig. 2: Training and validation accuracy of InceptionV3 (left) and ResNet50 (right) architectures trained on the three datasets. Top to bottom: basic dataset, dataset with class *cobblestone* and *wet asphalt* extended from image search, dataset with all classes augmented from image search. All data is plotted until training terminated due to early stopping.



(a) Classified as "grass".

(b) Classified as "grass".

(c) Training image for the class "grass".

(d) Classified as "snow"

(e) Classified as "snow"

(f) Training image for the class *snow*.

Fig. 3: The first two images in each row were misclassified. The rightmost images were part of the training set.

B. Analysis of Results

Looking at the confusion matrices in Figure 5, misclassification occurs in a pattern. Images from the class *wet asphalt* are often classified as *asphalt*, but in no case with the class *grass*, when augmenting the classes *cobblestone* and *wet asphalt*. In contrast classification accuracy for the class *cobblestone* increases by 2 % and for *asphalt* by 12 %. Images from the classes *snow* and *grass* are classified with high recall.

By examining the misclassified images, several possible causes for the misclassification could be identified.

The first one is the dominating color in the images. Within the given classes *snow* and *grass*, the most distinctive feature is color, as grass is commonly green and a road covered with snow is commonly white. Color as a learned feature can thus result in a high recall score for both classes. Evaluating the false positive classifications, the resulting images consist of images which contain these colors. Samples are shown in Figure 3.

When evaluating images from the class *dirt*, the misclassified samples partially contain patches of grass, which makes the class prone to be misclassified as *grass*. The same reason

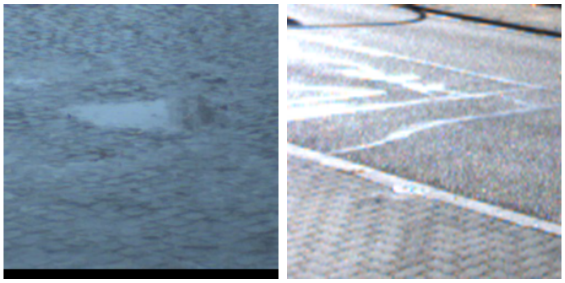


Fig. 4: Examples for images with ambiguous classes: remaining puddles (left), cobblestone and asphalt in the same ROI (right)

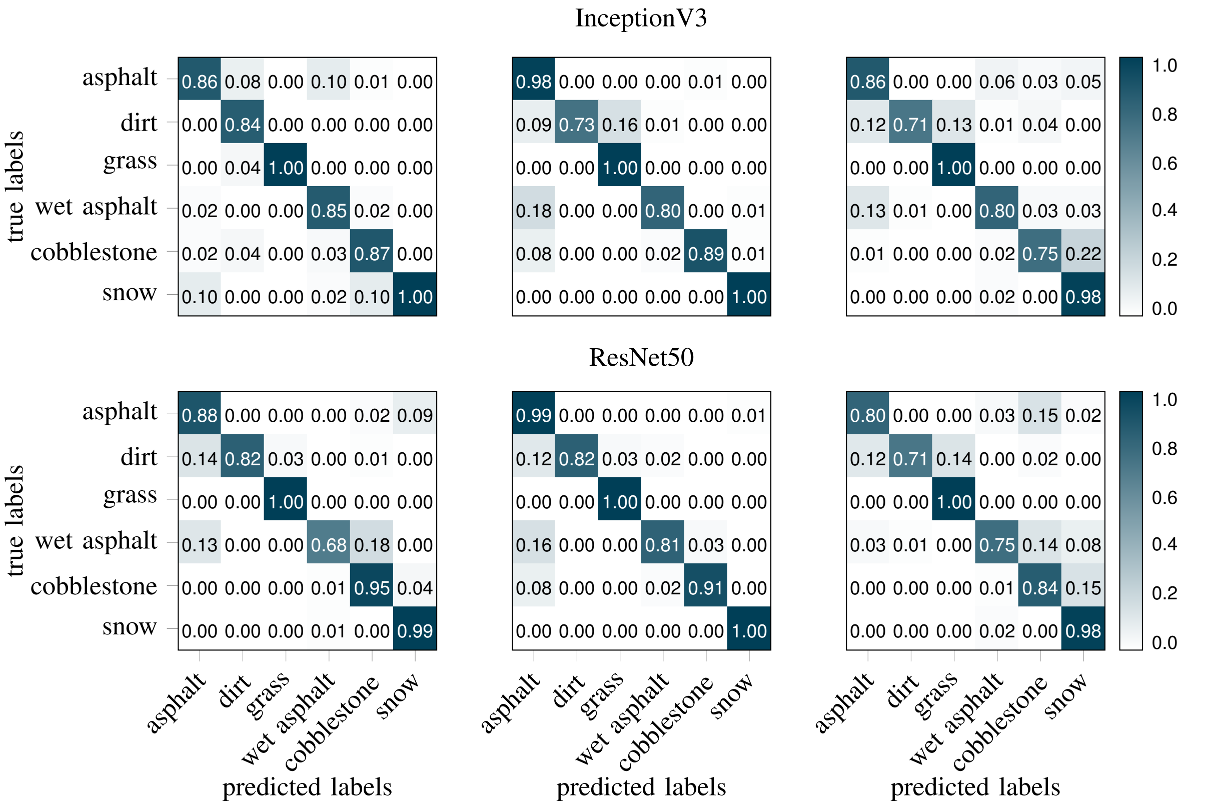


Fig. 5: Evaluation on test data: Confusion matrices for trained InceptionV3 (top) and ResNet50 (bottom) architectures. Left to right: Basic dataset, dataset with classes *cobblestone* and *wet asphalt* extended from image search, dataset with all classes augmented from image search.

for misclassification applies to images containing overexposed regions which appear white and can thus be misclassified as *snow*. Not only the samples containing the class *dirt* are prone to ambiguous texture information, as shown in Figure 4.

If the region of interest contains transitions between two road surfaces (e.g. *asphalt* and *cobblestone*, as shown on the right in Figure 4), it contains features of two classes and are therefore also prone to misclassification. Another example for this are left-over puddles on an already dry cobblestone road (Fig. 4 on the left). Although part of the surface is wet, the surface should be considered as *cobblestone*.

Although the classifier operates on single frames, the images are part of sequences. In order to get an impression of the stability of the classification results when applied to those sequences, we evaluated the classification on sample sequences from the *Stadtpilot* project, which were not part of the training dataset. No tracking was performed between frames.

For this classification, ResNet50 trained on the second dataset was used. In Figure 6 three of the worst classification results in sequences are shown. Looking at these results, it is visible that misclassification tends to appear in groups of several frames. Fluctuations as shown in the center and bottom sequence can render the trained classifiers unsuitable for adapting control algorithms.

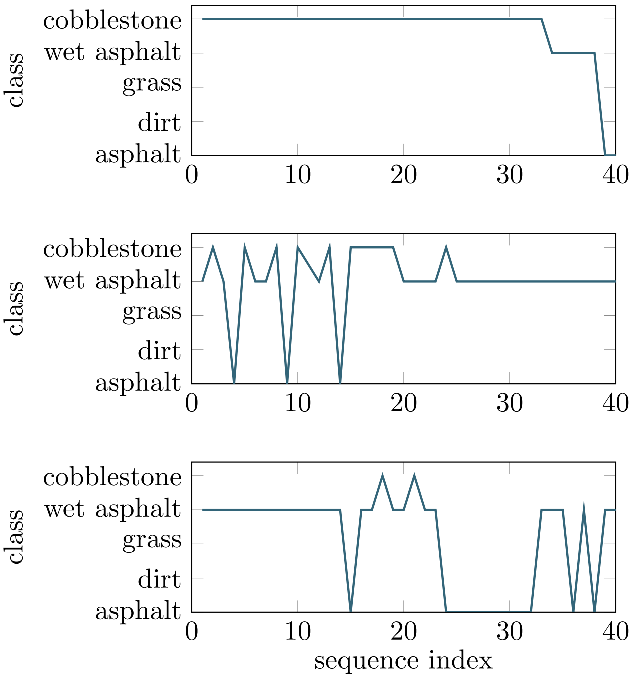


Fig. 6: Classification results in sequences. Each image is classified separately, no tracking is performed. Ground truth from top to bottom: cobblestone, wet asphalt, wet asphalt.

VII. Conclusion & Future Work

In this paper we presented an approach for CNN-based road surface classification, which can be used as a basis to predict the road friction coefficient. The trained network models are able to differentiate between six types of surface labels. Augmenting data from publicly available datasets for automated driving with images from Google search for minority classes has helped to increase the overall classification accuracy of ResNet50 by 4 % and reduced confusion when differentiating between wet asphalt and cobblestone.

For the presented task, the trained ResNet50 model outperforms InceptionV3 by 2 % with respect to average classification accuracy on test data, if the basic dataset is extended with additional images for minority classes. With an average accuracy of 92 % on the test dataset, the approach performs better than the image-based approaches using classic features and classifiers [4], [5]. Unfortunately neither [4] nor [5] present more detailed results about precision and recall, such that a comparison based on the average accuracy may be biased by the actual choice of the classes and the available distinctive features.

Compared to the audio-based reactive classification approach presented in [1], our approach is outperformed by 5 % to 6 % (CNN vs. CNN + LSTM), regarding average classification accuracy, but provides a look-ahead in front of the vehicle.

For future work, we will extend the approach to semantically segmented images. This promises fine grained information about the location of surface patches and thus additional information for the parameterization of control algorithms. In order to stabilize classification performance on sequences, we will evaluate the addition of LSTM units to the ResNet50 model.

For an application of the proposed CNN model to road friction estimation, the occurring misclassification of *wet asphalt* and *dirt* as *asphalt* is a critical issue, as this can possibly lead to an over-estimated road friction coefficient, which can in turn reduce control performance in critical situations. For this reason, we will further investigate the features learned by the ResNet50 model in order to resolve misclassification of those classes.

Acknowledgement

We would like to thank our colleague Felix Grün for valuable discussions.

References

[1] A. Valada and W. Burgard, "Deep spatiotemporal models for robust proprioceptive terrain classification," *The International Journal of Robotics Research*, vol. 36, pp. 1521–1539, 13-14 2017.

[2] K. Han, E. Lee, M. Choi, and S. B. Choi, "Adaptive Scheme for the Real-Time Estimation of Tire-Road Friction Coefficient and Vehicle Velocity," *IEEE/ASME Transactions on Mechatronics*, vol. 22, no. 4, pp. 1508–1518, 2017.

[3] F. Holzmann, M. Bellino, R. Siegwart, and H. Bubb, "Predictive estimation of the road-tire friction coefficient," in *2006 IEEE Intern. Conference on Control Applications*, 2006, pp. 885–890.

[4] R. Omer and L. Fu, "An automatic image recognition system for winter road surface condition classification," in *2010 IEEE Intern. Conference on Intelligent Transportation Systems (ITSC)*, 2010, pp. 1375–1379.

[5] Y. Qian, E. J. Almazan, and J. H. Elder, "Evaluating features and classifiers for road weather condition analysis," in *2016 IEEE Intern. Conference on Image Processing (ICIP)*, 2016, pp. 4403–4407.

[6] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets Robotics: The KITTI Dataset," *International Journal of Robotics Research (IJRR)*, vol. 32, no. 11, pp. 1231–1237, 2013.

[7] W. Maddern, G. Pascoe, C. Linegar, and P. Newman, "1 Year, 1000km: The Oxford RobotCar Dataset," *The International Journal of Robotics Research (IJRR)*, vol. 36, no. 1, pp. 3–15, 2017.

[8] T. Nothdurft, P. Hecker, S. Ohl, F. Saust, M. Maurer, A. Reschka, and J. R. Böhmer, "Stadtpilot: First Fully Autonomous Test Drives in Urban Traffic," 2011 IEEE Intern. Conference on Intelligent Transportation Systems (ITSC), 2011, pp. 919–924.

[9] A. Giusti, J. Guzzi, D. C. Ciresan, F.-L. He, J. P. Rodriguez, F. Fontana, M. Faessler, C. Forster, J. Schmidhuber, G. D. Caro, D. Scaramuzza, and L. M. Gambardella, "A Machine Learning Approach to Visual Perception of Forest Trails for Mobile Robots," *IEEE Robotics and Automation Letters*, vol. 1, no. 2, pp. 661–667, 2016.

[10] M. Smith, I. Baldwin, W. Churchill, R. Paul, and P. Newman, "The new college vision and laser data set," *The International Journal of Robotics Research*, vol. 28, no. 5, pp. 595–599, 2009.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Intern. Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[12] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *2016 IEEE Intern. Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2818–2826.

[13] S. Khaleghian, A. Emami, and S. Taheri, "A technical survey on tire-road friction estimation," *Friction*, vol. 5, no. 2, pp. 123–146, 2017.

[14] M. Varma and A. Zisserman, "A Statistical Approach to Material Classification Using Image Patch Exemplars," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 11, pp. 2032–2047, 2009.

[15] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded Up Robust Features," in *Computer Vision – ECCV 2006*, A. Leonardis, H. Bischof, and A. Pinz, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 404–417.

[16] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *CoRR*, vol. abs/1710.05381, 2017.

[17] J. Deng, W. Dong, R. Socher, L.-J. Li, Kai Li, and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Intern. Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2009, pp. 248–255.

[18] Z. Pezzementi, T. Tabor, P. Hu, J. K. Chang, D. Ramanan, C. Wellington, B. P. W. Babu, and H. Herman, "Comparing Apples and Oranges: Off-Road Pedestrian Detection on the NREC Agricultural Person-Detection Dataset," *CoRR*, vol. abs/1707.07169, 2017.

[19] J. Krause, B. Sapp, A. Howard, H. Zhou, A. Toshev, T. Duerig, J. Philbin, and L. Fei-Fei, "The Unreasonable Effectiveness of Noisy Data for Fine-Grained Recognition," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., Cham: Springer International Publishing, 2016, pp. 301–320.