Outdoor terrain recognition based on transfer learning

Song Zeng¹, Hao Huang² and Zhenyun Shi¹*

¹ School of Mechanical Engineering and Automation, Beihang University, Beijing, 100191, PR China
² School of Mechanical Engineering and Automation, Beihang University, Beijing, 100191, PR China
*Corresponding author’s e-mail: shichong1983623@hotmail.com

Abstract. Terrain recognition exerts an extremely important role in outdoor mobile robot gait planning, speed control, environment perception, etc. Compared with the traditional terrain recognition process that uses color, texture, and other underlying features to describe terrain images, the present study starts from the perspective of transfer learning. MobileNet and DenseNet are employed for high-level feature extraction, and the voting integrated learning algorithm is used to classify high-level feature data sets. In the meanwhile, we have established an outdoor terrain data set that conforms to the traveling process of outdoor mobile robots, and processed the collected video data with key frames and sliding windows. The accuracy of the classification results reached 97%, basically satisfying the needs of actual terrain recognition.

1. Introduction

Regarding terrain image recognition, the three aspects that have the greatest impact on recognition accuracy include the quality of the terrain data set, the extraction method of terrain features, and the classification algorithm of terrain features[1]. In order to obtain a better terrain classification effect and generalization ability of the actual test effect, the three aspects are complementary, requiring optimization and control throughout the process.

The existing open-source terrain classification data sets are relatively limited. The current data sets mainly concentrate on remote sensing image data sets, such as RSSCN7[2] and LU[3] data sets, as well as some commonly used texture data sets such as GTOS[4], DTD[5]. However, for the recognition of the actual terrain of the outdoor robot, remote sensing, and related texture data cannot better reflect the image collection situation of the robot when it is running. Shandong University has produced an SDU-Terrain dataset containing six types of terrain[6], covering sand, grass, and other terrain conditions. However, it is not open source. Therefore, it is of necessity to establish an image dataset that conforms to the robot's operation process for training.

In terms of visual image feature extraction, a lot of progress has been made. Since Li Feifei established ImageNet[7], various novel neural networks have emerged in an endless stream. In addition, better results have been achieved on the ImageNet data set. The proposal of transfer learning reveals that data that can be trained on other training sets based on the deep learning network structure can be reused in similar instances[8]. Nevertheless, in terms of terrain image recognition, it is mainly based on low-level features such as color, texture, shape and location using SIFT, ORB, and other artificial design features for feature extraction[9, 10]. These low-level artificial design features are complicated to use and require a large amount of calculation. Additionally, they are also subject to color, light, and other environmental information. Research[11] shows that in the feature extraction of neural networks, the...
first few layers can almost achieve the purpose of feature extraction at the bottom. At the same time, the weights of neural networks also contain more information. The use of existing neural networks for feature extraction will significantly improve the effect of terrain image extraction.

In terrain image classification, traditional machine learning methods such as support vector machines, K-nearest neighbors, random forests are extensively applied[12-15]. However, there are numerous parameter settings, and it takes a lot of time to adjust to the optimal settings. The use of neural networks to establish a fully connected layer for classification is currently directly integrated into the feature extraction network. However, due to its low computational efficiency and slow learning speed, it is easy to fall into local minimums[16]. To improve the accuracy of traditional machine learning classifiers, these weak classifiers can be integrated, and the accuracy of classification results can be improved using integrated learning[17].

Based on the above analysis, to meet the adaptability of outdoor mobile robots to different terrains and improve their autonomous operation capabilities in outdoor environments, it is necessary to select appropriate terrain perception methods, establish reasonable data sets, and adopt appropriate terrain feature extraction methods and classifiers. The present study uses visual perception to perceive the terrain and establishes a terrain data set that conforms to the traveling process of outdoor mobile robots. The camera is offset to the ground by an angle of about 60 degrees, and images of 9 common outdoor terrains including sand and grass are also collected. The transfer learning method is used to extract terrain features using a variety of trained neural networks, and multiple classification algorithms are integrated through the ensemble learning method to enhance classification accuracy. In the current experiment, the robot terrain picture and video recognition module are designed, and the overall design scheme is shown in the Figure1.

![Figure 1. Overall design](image)

## 2. Materials and Methods

### 2.1. Dataset Production

This project uses the RunCam2 camera to simulate the actual car body operation process for image collection. Through controlling the installation height and tilt angle of the camera, the pictures under different terrain conditions are collected. The RunCam2 camera has a camera and video mode and the size of the picture and video is 1920*1080@30fps. With digital anti-shake function, the data can be easily saved through the built-in SD card.

Each image is randomly cropped to obtain a uniform size which is (256,256). The small area image is manually classified, the data label is done, and the initial data set is produced. The data set of this project consists of 9 different terrain categories, including asphalt, dirt, floor, grass, gravel, rock, sand, wood chips, and water. Each category contains 300 pictures, all of which are under picture normalization, limiting the size of a single-pixel in the picture to between 0 and 1. As a result, it will be beneficial for the subsequent terrain feature extraction and training.
For supervised learning, the larger the data set, the better the classification performance of the final classifier. Therefore, on the basic data set established in this article, the image augmentation process is performed during the running process, and the image is randomly cut, rotated, and scaled. Changing the brightness and contrast of the image to expand the image data set will contribute to improving the performance of the final classifier.

2.2. Terrain feature extraction

Highly distinguishable features can significantly improve the accuracy of classification. Traditional terrain feature extraction mostly depends on the color and texture of the image to manually design feature operators. Manually designed features can only reflect the information of a certain dimension in the image. At the same time, only the training data is similar to the test data, and the color and texture are visualized. Features can achieve higher accuracy. In comparison with the images captured in the dataset, the images collected in the actual scene are much more complex. To obtain more comprehensive features more simply and conveniently, we use deep learning methods for feature extraction. We all know that deep learning is in a leading position in the field of image classification. It can be applied in two ways[5]. One is to directly provide end-to-end classification results, and the other is to extract deep features from the middle layer of the deep learning network. Two different application methods are used, using transfer learning methods, and the DenseNet and MobileNet networks that perform better in ImageNet training and have corresponding representativeness and project adaptability are also selected for experiments.

The advantage of deep learning refers to that it can realize self-learning on large-scale data and overcome the shortcomings of human intervention in traditional visual feature extraction. The structure of the deep network makes it possible to extract features hierarchically, and the deep features of fully connected layers are usually the preferred method. However, because high-dimensional features are directly mapped to the fully connected layer, the spatial structure information of the terrain image is lost, and the pooling layer is another method which can be used to extract deep features. In this paper, the deep features of the largest pooling layer are used to express the terrain image, and the deep feature data set is constructed to test the classification performance of different classifiers.

![Figure 2. Feature extraction fusion and classification architecture](image)

2.3. Terrain Classification

Additionally, the process of extracting high-level features of the terrain is also a process of dimensionality reduction. Directly inputting the original image into the classifier will cause the running speed to be slow due to the high dimensionality and too many parameters, and the classification effect will be also reduced. Using the extracted depth features in low-dimensional classification, the types of optional classification algorithms will be greatly expanded.

For the previously extracted high-level features, we first build a fully connected layer for classification test using some operations such as pooling and dropout to improve the generalization performance of the model. Then, we compare the effectiveness of the features extracted by different neural networks for classification. According to the obtained test results, the best extracted high-level
data set is used for different traditional classifiers comparative experiments. In this paper, SVM, Logistics Regression, and KNN, which are often used in the classification process, are selected for testing, and the classification accuracy and training speed of high-level data sets are compared respectively with the aim to find the optimal results in traditional classifiers.

Ensemble learning is a learning task that combines multiple single learners and aggregates the prediction results. Based on the three previously selected classifiers, the results are aggregated, and stacking fusion and voting classifiers are used to perform statistics based on the output results of different types of classifiers. The result with the largest proportion is output as the final result of this voting.

3. Results & Discussion

3.1. Data set feature extraction process

The acquisition process mainly considers the following factors: 1) the sensing range of the camera; 2) mutual interference between different terrains; 3) identification and perception of dangerous environments. The camera is mainly used for terrain recognition. It needs to be installed at a safe height for the collection during the actual operation. Different heights correspond to different sensing ranges. The installation height is approximately 30cm in combination with the actual vehicle body height. To avoid the impact of more complex mixed terrain in the picture area on the classification performance, the camera is tilted to the ground to guarantee that the field of view is mainly the terrain area where the robot is located. For some dangerous areas, such as water bodies, strictly speaking, they do not belong to the terrain environment. However, they are often encountered by robots in outdoor environments. Therefore, they need to be perceived and confirmed in advance. The angle between the camera and the ground is about 60 degrees. It can ensure the recognition of the terrain to be passed within the safe range of the robot to guarantee the safety of the robot.

We divide the terrain data set into a training set and test set according to 8:2 and use the data enhancement function in Keras to expand the terrain image before entering the neural network. The input dimensions of DenseNet and MobileNet are both (256, 256). The training parameters in the two networks are set to not update respectively, that is, the network parameters obtained by training ImageNet are trained on the terrain data set, and the global max pooling layer before the fully connected layer is intercepted as the output layer of the neural network in order to construct a new truncated network. We use the truncated network to predict and output the Terrain9 data set to obtain the depth features of the terrain data, and the output depth features are all 1024-dimensional vectors. The depth

![Nine types of terrain images in the dataset](image-url)
features are saved and the depth feature data set is constructed for subsequent testing and analysis of different classifiers. We fuse the deep feature sets extracted by DenseNet and MobileNet, which can expand the feature set. Subsequently, we compare the classification performance of single network features and fusion features in different classifiers.

3.2. Model classification effect comparison

This paper uses MobileNet and DenseNet networks to extract high-level feature data sets of terrain images, constructs different fully connected networks to classify the data sets, and selects the obtained high-level feature data based on training time and training effect. Initially, a three-layer fully connected layer was constructed to classify the feature data set. From the changes in the training loss function, it was found that the training process is overfitted. In turn, the dropout layer was added to the fully connected layer to reduce the number of fully connected layers. Finally, it is found that using a single fully connected layer to classify high-level features is the best without overfitting. It shows that the high-level features extracted by CNN remain basically in the linear dimension, and the use of a multi-layer neural network for training will cause over-fitting.

Therefore, a single-layer fully connected layer is used to learn the high-level fusion features of MobileNet, DenseNet, and two CNN networks. In addition, the recognition effect of different neural networks to extract high-level features is compared. Among them, the classification effect of different high-level features takes little time. However, the average accuracy rate is not much different. The following table 1 compares the training time comparison of the high-level features obtained by different networks. The longest is the fusion feature classification training time of the two networks. Obviously, the average accuracy of the fusion feature classification and recognition is the highest and the fusion feature facilitates the recognition by the classifier.
Clearly, the classification and recognition of the fully connected layer that the fusion feature contains more information about the terrain, which is beneficial for the classifier to achieve better results. We use SVM, KNN and Logistics Regression to classify and recognize the fusion features as well as obtain the comparison result of model training time and model training accuracy. In this paper, a confusion matrix is used to evaluate the model. The traditional classifier is used to classify high-level features. The accuracy of the classification results is improved compared with the classification of the fully connected layer. Meanwhile, the training speed is faster than the neural network classification method. Because the performance of the platform equipment is limited, we prefer to use traditional classifiers for voting fusion to integrate the results of multiple weak classifiers which can thus balance the contradiction between recognition accuracy and running time.
Table 2. The training time and classification accuracy of different classifiers for fusion features

| Classifier        | Average accuracy | Training time |
|-------------------|------------------|---------------|
| SVM               | 0.9605           | 4.45s         |
| KNN               | 0.9493           | 0.28s         |
| Logistics Regression | 0.9616       | 1.87s         |
| Stacking          | 0.9643           | 23.54s        |
| Voting            | 0.9704           | 4.05s         |

3.3. Video process
In the past, most of the methods dealt with the recognition of a single picture. However, when the robot navigates in the terrain, the camera captures a sequence of frames. Therefore, we have to process a series of continuous video streams. Any two consecutive frames have a lot of common image areas, that is, they look extremely similar visually. To use traditional machine learning-based algorithms to characterize the topography of an image, it is of necessity to extract features from each tile. The high-level features of the image are input into the classifier for recognition and classification, and the classification result is obtained. If you perform block image segmentation and feature extraction and recognition for each picture, the amount of running required is very huge. The recognition results are also very similar and thus we use the fast optical flow method to make a rough comparison between the previous frame and the current frame. If the current frame is consistent with the previous frame, we ignore the recognition of the current image until a new scene environment is detected.

Although we have controlled the angle of the camera to the ground during the image collection process, we can avoid excessive terrain interference. During the process of switching between different terrains, there will still exist a combination of multiple terrain environments in the collected terrain images. To identify the real terrain environment where the robot is located and the terrain environment it is about to enter, we employ the sliding window method to scan the entire picture, divide a large-size picture into multiple (256,256) size tiles, and classify different tiles. The results are weighted statistics. Finally, the current terrain recognition results and the upcoming terrain estimation results are obtained.

4. Conclusions
To conclude, in this paper, we independently construct an outdoor mobile robot running terrain data set, and use the migration learning method to extract the high-level features of the terrain image from MobileNet and DenseNet, which lowers the workload of manual design of the underlying terrain features. Through fusing the features extracted by the two networks into different categories for classification and recognition in the machine, the ensemble learning method can achieve a good balance between model classification accuracy and training time. In future work, we will apply our method on actual robots to further confirm the robustness of the model and test generalization in real complex scenes.

Acknowledgments
The work reported in this article is supported by D316 Soft Robot Research Group of Beihang University of Mechanical Engineering.

References
[1] Kozlowski P, Walas K. (2018)Deep neural networks for terrain recognition task, [C]. Warsaw University of Technology.
[2] Zou Q, Ni L, Zhang T, et al. (2015)Deep Learning Based Feature Selection for Remote Sensing Scene Classification[J]. IEEE Geoscience & Remote Sensing Letters, 12(11):2321-2325.
[3] Yang Y, Newsam S. (2010)Bag-of-visual-words and spatial extensions for land-use classification: Sigspatial International Conference on Advances in Geographic Information Systems, [C].
[4] Xue J, Zhang H, Dana K. (2018)Deep Texture Manifold for Ground Terrain Recognition[J].
[5] Xue J, Zhang H, Nishino K, et al. (2020)Differential Viewpoints for Ground Terrain Material
[6] Li X, Ma X, Song P. (2019) Fusion of Deep Feature and Hand-Crafted Features for Terrain Recognition[J]. IOP conference series. Materials Science and Engineering, 646:12052.

[7] Deng J, Dong W, Socher R, et al. (2009) ImageNet: A large-scale hierarchical image database, [C]. IEEE.

[8] Pan S J T I. (2011) Domain Adaptation via Transfer Component Analysis[J]. IEEE Transactions on Neural Networks, 22(2):199-210.

[9] Hui Z, Xiaofang D, Fengchi S, et al. (2016) Terrain classification in field environment based on Random Forest for the mobile robot, [C]. TCCT.

[10] Nampoothiri M G H, Anand P S G, Antony R. (2020) Real time terrain identification of autonomous robots using machine learning[J]. International Journal of Intelligent Robotics and Applications, 4(3):265-277.

[11] Pan S J, Yang Q. (2010) A Survey on Transfer Learning[J]. IEEE Transactions on Knowledge and Data Engineering.

[12] Zou Y, Chen W, Xie L, et al. (2014) Comparison of different approaches to visual terrain classification for outdoor mobile robots[J]. Pattern recognition letters, 38:54-62.

[13] Suykens J A K, Lukas L, Van P, et al. (2000) Least squares support vector machine classifiers: a large-scale algorithm[J].

[14] Keller J M, Gray M R, Givens J A. (2012) A fuzzy K-nearest neighbor algorithm[J]. IEEE Transactions on Systems Man & Cybernetics, SMC-15(4).

[15] Yutaka, Nakasone, Osamu, et al. (2007) Shock Index Correlates with Extravasation on Angiographs of Gastrointestinal Hemorrhage: A Logistics Regression Analysis[J]. Cardiovascular & Interventional Radiology.

[16] Haykin S S. (1994) Neural Networks – A Comprehensive Foundation[J].

[17] Zhou L, Liu Z, Wang W. (2020) Terrain Classification Algorithm for Lunar Rover Using a Deep Ensemble Network with High-Resolution Features and Interdependencies between Channels[J]. Wireless Communications and Mobile Computing, 2020:1-14.