Array 15 (2022) 100195

Contents lists available at ScienceDirect

Array

journal homepage: www.elsevier.com/locate/array

Road scene classification based on street-level images and spatial data

Roman Prykhodchenko *, Paweł Skruch

Department of Automatic Control and Robotics, AGH University of Science and Technology, Cracow, Poland

ARTICLE INFO

Keywords:
Autonomous vehicles
Deep learning
Scene classification
Scene categorization
Spacial data

ABSTRACT

Understanding the context of the scene is one of the most important aspects for new generation of autonomous vehicles. It is a very trivial task for a human to recognize the scene context by only a single look at the picture, however, for a computer, it is still a challenging task. This problem can be solved by automatic data labeling using deep-learning models for scene classification. Relying on scene type labels we can select relevant scenes to prepare a balanced dataset to train more advanced instance detection models using data from a specific road condition.

This study presents a novel framework based on a deep convolutional neural network (CNN) for the automatic road scene classification of street-level automotive images. For the evaluation of our approach, we use a well-known autonomous benchmark dataset, from which we extract geo-position data and combine them with predictions from the scene classification model to get ground truth labels to train and evaluate a ResNet-50 model for scene classification. The results and comparison with state-of-the-art methods are presented.

1. Introduction

Automotive scenes contain many details and the number of scenarios is infinite. The ability of the computer vision algorithm to classify certain scenario and predict correct label is very critical to provide curtain level of safety of an autonomous vehicle.

Current achievements in image classification are significant, although the majority of efforts build to classify a wide range of image categories [1,2]. Deep convolutional neural networks have been showing impressive classification results and continuous improvement on the object-centric ImageNet classification benchmark [3]. However, scene categorization is an important step for further improvement of visual perception for autonomous driving. The scene-centric Places-CNN, for scene classification/recognition introduced in work [4], which has been trained on 2.5 million images from 205 scene categories of places, including outdoor scenes. An extended version of the dataset with 2.1 images for classification of 365 categories has been presented in work [5]. The scene-level task is more challenging in comparison to the object-centric task feature learning due to the larger diversity of scenes and the number of possible combinations in scenarios.

Scene classification topic has been studied broadly in the robotic field, such as semantic mapping [6,7], where semantics information was used to help a mobile robot with high-level navigation in an indoor environment. Many methods infer scene category by detecting certain objects that usually belong to a certain scene.

In the case of automotive datasets, a scene semantic can be used for scene classification, such as buildings, sidewalks, parking lots, and other construction that could represent an urban road condition. On other hand, typically for the rural scene could be an abundance of plants or vegetation. Where for the highway scene a road will be a dominating class in the scene.

Scene labels (i.e. categories) of these datasets are mainly defined on the following aspects:

- weather type and light conditions [8].
- static objects on the scene, constructions, bridges, tunnels e.g. [9],
- category of special traffic scenes.
- intersection classification.

In most cases, autonomous datasets do not provide traffic scene labels, but only object-level attributes and a general description of the dataset in original paper [10]. In some cases, we have scene categories per video clip without precise per frame annotation [10]. However, the annotation of traffic condition is not provided. Visual analysis and labeling of road scenes would be very time-consuming for automotive datasets.

In this paper, we study image classification in the context of road scene classification for automotive datasets. Our focus is on scene categorization/classification is based on a single image for categories as urban, rural, highway road scenes. In order to achieve it, we use two convolutional neural networks, one to obtain a scene labels out of object-level classes. The image-based scene labels have been combined with the output of our map-based approach, which is using spatial

* Corresponding author.
E-mail addresses: prykhodc@agh.edu.pl (R. Prykhodchenko), pawel.skruch@agh.edu.pl (P. Skruch).

https://doi.org/10.1016/j.array.2022.100195
Received 15 May 2022; Accepted 24 May 2022
Available online 30 May 2022

data for scene classification. The second residual deep convolutional network [11] was trained based on labels from previously mentioned methods.

We believe that scene condition labeling can be used for both online and offline applications such as semantic mapping of autonomous datasets and driving assistance.

The remainder of the paper is organized as follows: Section 2 gives a review of related works and Section 3 explain our proposed methodology and system architecture. Section 4 presents the results of scene classification for proposed methods. Finally, we conclude by discussing of our results and avenues for further research.

2. Related work

Our work is related to three interconnected research areas: (i) image classification in general, (ii) traffic scene classification, and (iii) image classification for spatial data.

Image classification in general. The problem of image classification was initially on developing classification systems which would archive the best possible accuracy results on PASCAL VOC competitions [1]. One of the best where best approaches [12,13] are based on extraction of many low-level local features, coined to one histogram and non-linear kernel classifiers such as Support-vector machine (SVM) are used to perform classification. The next step in performance has been obtained with spatial Fisher vectors (SFV) [14,15] which aggregate local descriptors features, encoding them with statistics, fit a Gaussian Mixture Model (GMM), and feeding them to kernel SVM classifiers. However, in 2012 it was shown that Convolutional Neural Networks (CNNs), so-called AlexNet [16] trained in a supervised fashion on ImageNet dataset outperformed the Fisher vectors. Next-generation of depth network architectures further improved classification accuracy: VGGNet [17], GoogLeNet [18], ResNet [11] shown that network depth is a crucially important for reach a success.

Traffic scene classification. Early work in the image-based scene classification proposes to use the probability of each object detected on the image converted to a vector in order to classify the scene, where each object consider as a typical for one of the scene classes [19]. Another common approach for the scene classification is using semantic labels. In the work, [20] scene semantic segmentation labels were used to classify 13 urban texture classes, including different types of road layouts, the presence of cars, pedestrians, and a pedestrian crossing in front of the vehicle. In the papers [21,22] semantic labels used for the understanding of traffic scenes, CNN semantic segmentation network was trained on images taken from the same location, but under different weather or illumination conditions.

In the paper, [9] pooled features extracted from DenseNet-121 deep network architecture were used to train SVM classifier with RBF kernel for such traffic scene classification: highway, road, tunnel, exit, settlement, overpass, booth, traffic. Another recent work [23] presented a traffic scene classification method based on resent-50, fine-tuned and pretrained on the Places365 dataset. In addition, standard Long short-term memory modules (LSTM) architectures were added, which allow improving results by the temporal aspect.

Image classification for spatial data. An urban scene recognition based on spatial data, which in most cases satellite images, which has important application value in improving urban land use rate and land use monitoring [24]. A deep-network strategy actively used in order to predict geographical objects by exploring deep features of the high-resolution imagery [25]. Such an approach helps to extract semantic contents of the area and generate high-resolution semantic maps, which might play a crucial role in carving out the path ahead for autonomous vehicles. Another way to get semantic labels is using OpenStreetMap (OSM) [26], the project is established to share a knowledge collective that provides user-generated street maps. Each map is consists from small parts called as tile represented as a set of raster and vector markers, such as lines, polygons etc. In most cases, tiles generated by manual labeling [27] or in the semi-automatic way using the Normalized Difference Vegetation Index (NDVI) for vegetation detection from aerial or satellite imagery [28].

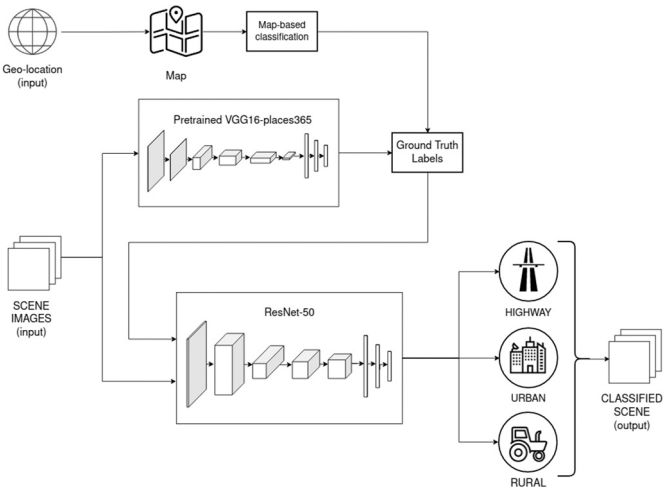

Fig. 1. The road scene classification framework.

3. Methodology

To fulfill the current knowledge gap, we introduce a framework of parallel deep CNN models to detect urban/rural/highway scenes from street-level images of automotive datasets.

The framework is presented on Fig. 1 comprises three different methods for scene classification:

- Map-based classification methods use the color of the map zones to distinguish the type of area based on a predefined color palette typical for different areas.
- Convolutional neural networks (CNNs) trained on Places365 dataset. It was trained on set of Places365-Standard dataset with ~1.8 million images from 365 scene categories, where there are at most 5000 images per category. Out of 365 classes we select 38, those are typical for urban, rural, highway scenes and sum their prediction scores to make final decision about scene class. It aims to understand the subtleties of street-level images despite the dynamics of weather conditions and diversities of the scene, with a high level of attention to details.
- The residual network ResNet-50, which was trained on a custom dataset containing images from popular autonomous datasets (e.g., KITTI, nuScenes) to detect the occurrence of urban, rural, and highway scenes, despite weather and light conditions, with distributed attention to a different level of details on the scene.

Next, we define typical features to establish ground truth labels for each class.

Urban condition. By urban area condition, we assume that it is a built-up area, with a high population density and infrastructure of the human-made environment. Mostly, it is a public road in the city district or residential neighborhood, with dens residential buildings, where potentially could be largely populated by people and many public places.

Rural condition. In the rural area condition, we could consider an opposite situation to urban, with a low-density population and with a low number of total man-made infrastructure. Such areas are away from the city, on the scene you can see a lot of vegetation.

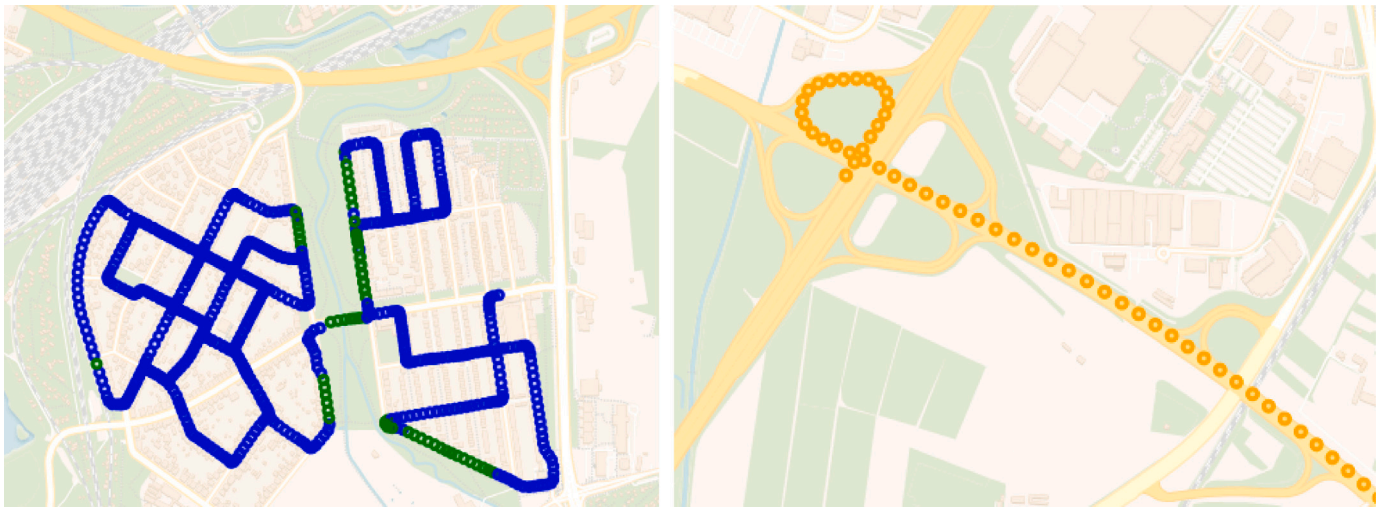

Fig. 2. This figure shows the Global Positioning System (GPS) traces plotted on map, colored accordingly to the map-based classification scene label, where (a) urban scenes is blue, rural scenes is green (b) highway scenes is orange. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Highway condition. The highway definition can be simply described as the main or public road. Most public roads have at least two lanes, one for traffic in each direction, separated by lane markings. The highway road is comparably easy to distinguish from urban and rural since the most distinctive feature is that the dominant place is taken by the road itself. However, highways can lay down through a city or countryside, which implies many elements inherent in urban and rural conditions. On most road maps the highway roads are usually marked as a dark yellow or orange color.

3.1. Map-based classification

The KITTI (Karlsruhe Institute of Technology and Toyota Technological Institute) dataset by default providing a raw data collected from different sensor sources, that are built-in into the car, including position data from Global Positioning System (GPS) sensors. It is provided in such a way that for each visual frame we have an additional set of metadata, such as latitude, longitude, altitude, velocities, and accelerations that are stored in a separate file but has the same timestamp which allows us to make a correlation between visual frame data and metadata. Based on GPS location data (latitude, longitude) we can plot positions of the car on the map, which allows us to see where the certain frame was collected. In Fig. 2, where GPS traces colored based on different road condition.

The next step would be to use geolocation data and a High-definition map to define a class in which conditions a frame was taken. In this work, we worked with raster spatial data, where each pixel represents a discrete value which it could be converted into a class of road condition based on heuristically defined thresholds for each of RGB (red, green, blue) channel. Within application to autonomous data labels provided by OpenStreetMap can be filtered on a pixel-level, where colors represent different types of areas, such us gray — urban area, green or white — wood, pole, orange, and yellow as highway and public roads. The result of using OpenStreetMap for pixel-based classification is shown in Fig. 3.

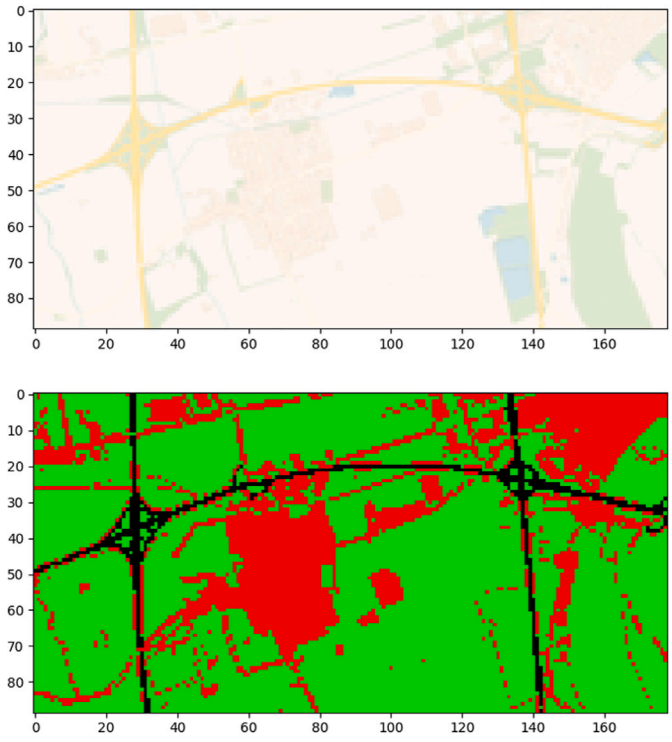

Fig. 3. An example of raster map converted in to labeled area. On top the origin raster map without text annotations. On the bottom is labeled map, where colors: red as urban areas, black as highway and green as rural areas. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.2. Places365-CNN architecture

An architecture of Places-CNNs based on popular object detection CNNs architectures, AlexNet, GoogLeNet, and VGG16 convolutional-layer CNN. Visualization of the attention zone of the CNNs trained on Places shows that object detectors emerge as an intermediate representation of scene classification. Places-CNN trained on a highly diverse number of scenes, more than 1.8 million images, to predict scene out of 365 scene categories, indoor/outdoor type, scene attributes, and the class activation map together from PlacesCNN. The original version of the dataset Places365-Standard was extended up to 8 million training images and presented for Places Challenge 2016 held as part of the ILSVRC Challenge.

As an input layer, VGG16-Places365 takes the resized version of the image ($224 \times 224 \times 3$). Then out image is passed through a sequence of convolutional layers, where filters were used with a 3×3 receptive field with preserved spatial resolution. Spatial pooling is carried out by five max-pooling layers, which follow some of the conv. layers, and

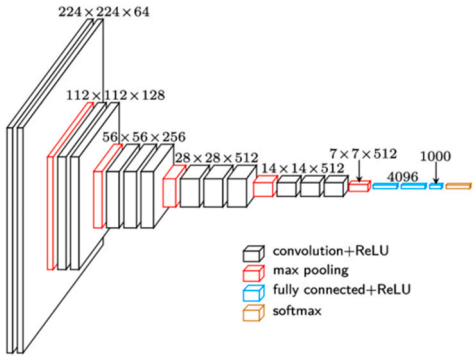

Fig. 4. The VGG16-Places365 architecture.

performed over a 2 × 2 pixel window, with stride 2. Then, flatten vector which comes out of the convolutions comes to the two dense layers with a random dropout of weights and the final dense which predicts values for 365 place classes. The VGG16-Places365 architecture presented in Fig. 4.

3.3. ResNet-50 architecture

The residual network architecture archived top results for object detection and localization tasks in the 2015 ImageNet Large Scale Visual Recognition Challenge (ILSVRC), with 3.57% error on the ImageNet test set. ResNet-50 is a convolutional neural network that is 50 layers deep with a skip connection to fit the input from the previous layer to the next layer without any modification in shape. This allows improving the speed and performance of the network comparing with other

As an input layer for ResNet-50, a training and testing images are resized to (224 × 224 × 3) feed-forward via transfer learning. As a second layer, we have a convolutional operation with a size 7 × 7 and a MaxPooling layer with a size 3 × 3. Following layers are presented as a sequence of residual modules in two modes: Convolution block (CB) which changes the dimension of the input, Identity Block (IB) which will not change the dimension of the input. Next, the AveragePooling layer down-sampled each 7 × 7 square to the average value in the square. Finally, the Softmax classifying the class. The architecture of ResNet-50 is shown in the Fig. 5.

4. Experiments and results

In this experiment, we use two methods: a map-based and an image-based method for the automatic generation of ground truth labels. After visual inspection of the results, we decided to combine predictions from both methods to compromise results, which finally used to train a ResNet-50 network to classify road scenes by ground-level image.

4.1. Ground truth labels agreement

In order to verify the validity of tho methods for ground truth estimation, we plot a confusion matrix to see mismatching between the map-based approach and the Places365 network approach. In Fig. 6 we can see that the rural class has many of mismatches with the urban class comparing the map-based method to the Places365 method. This probably happens due to the limitation of the heuristic map-based approach, where boundaries of each class could be wrongly defined by the supervisor or the color scheme on the map is wrongly representing the road type or surrounding area.

Examples of correctly classified in Fig. 8 and incorrectly classified images presented in Fig. 9.

In order to compromise both methods, we combine predictions from map-based and Places365 CNN predictions and decrease the mismatching rate. On Fig. 7, on left confusion matrix for combined predictions with map-based predictions, and on right the combined predictions with image-based predictions. As we can see, the map-based approach has in general a good match with the combined version, more than 80 percent for urban and highway scenes and more than 70 percent for rural scenes. However, in comparison, the Places365 approach demonstrates high matching with a combined version for urban class, but classifying some rural scenes as urban. This is happening because some classes present in Places365 dataset can belong to urban scene. Such classes as 'alley', 'street', 'driveway', 'forest_road', 'promenade', could represent a rural scene too.

Table 1
Classification results for pretrained ResNet-50 model trained on raw dataset.

	Precision	Recall	f1-score	Accuracy
Urban	0.88	0.77	0.82	
Rural	0.75	0.84	0.79	0.83
Highway	0.86	0.89	0.87	

4.2. Experimental settings

The experiment is implemented using PyTorch v1.8.1. Out of 3276, we left 1080 images to have the same number of images for each class. Then data was divided into train, validation, and test sets, in proportion 70/10/20, where 756 images belong to train, 108 images to validation, and 216 images to test sets respectively. A test set was used to estimate the final results. All images are resized and cropped to 224 × 224 pixels. The hyper-parameters for training are set as follows. The learning rate is initialized to 0.001 with momentum 9%, a decay learning rate by a factor of 0.1 every 7 epochs. The size of the batch is 4. Also, the last layer of the pretrained model is replaced with a new Softmax layer, which is relevant to the number of 3 classes.

4.3. Training on raw dataset

In order to improve the training process, we apply random transformation functions. At first, with a certain probability algorithm crop a random portion of an image and resize it to size 224 to 224 pixels. Secondly, with a probability of 0.5, it horizontally flips our images. Finally, we do normalization of the images.

Classification results for each class of pretrained ResNet-50 model with final training on custom KITTI dataset are presented in Table 1. Our best accuracy was 83% achieved after 10 epochs of the training. The f1-score is a single measurement for a system showing the best results for the highway class. In Fig. 10 the confusion matrix for the train set predictions, where similarly highway has matching ground truth labels slightly better than rural class, and much better than urban class.

4.4. Training on extended dataset

The paper [29] shown that data augmented by adding new images provide better results of accuracy and the training process goes faster comparing to the traditional way of data augmentation with the same amount of images to train on. In order to improve the result of our model, we decide to extend our dataset up to 1779 images, where 1419 images belong to train, and the same 177 images for validation and test sets respectively. Adding horizontally flipped images as a new image in combination with random crop and resize should provide the model more features and generalize the model.

The classification results presented in the Table 2, our model achieves 86%, which is demonstrating that proposed augmentation by extension of the dataset improves the model by three percent. Also, the f1-score for each of the three classes is improved, with a slightly better improvement for the highway class. In Fig. 11 the confusion matrix for train set predictions. Two predicted classes demonstrate very high

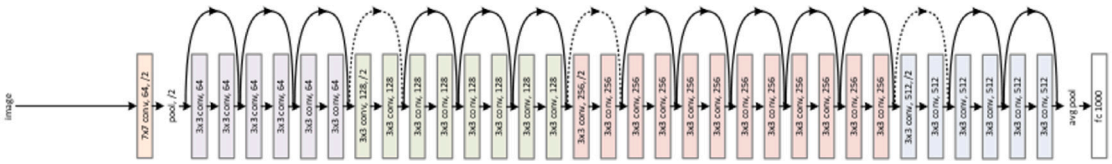

Fig. 5. The ResNet-50 architecture.

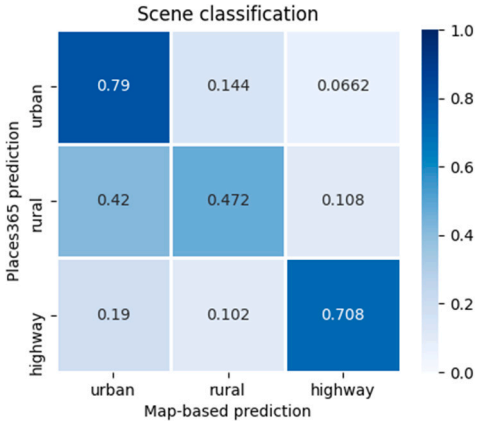

Fig. 6. Confusion matrix for scene classification between map-based and image-based Places365 methods. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 2
Classification results for pretrained ResNet-50 model trained on extended dataset.

	Precision	Recall	f1-score	Accuracy
Urban	0.89	0.80	0.84	
Rural	0.74	0.90	0.81	0.86
Highway	0.96	0.90	0.93	

Table 3
Classification results for pretrained models trained on augmented dataset.

Models	Parameters	Model size MB	Train time	Accuracy
ResNet-34	21M	178	0.72	0.84
ResNet-50	23M	376	1.0	0.86
ResNet-101	42M	592	1.71	0.88
MobileNetV3	2M	45	0.66	0.80
VGG16	138M	747	2.44	0.87

matching scores, rural and highway both 89% matches with ground truth labels. An urban class has a small improvement, however, we can see that it was still confused with rural scenes.

We compare performance of ResNet-50 with different state-of-the-art deep learning architectures, including ResNet variations with different number of depth layers and with other well known architectures preordained on ImageNet dataset, MobileNet version 3 [30] and VGG architecture with 16 layers. Model parameters and results of 50 epochs of training present in 3.

From the Table 3 that small model ResNet-34 demonstrate slightly lower performance then ResNet-50, but training is taking 25% less time. Oppositely deeper ResNet-101 model slightly outperforms ResNet-50 but takes twice more times to train. For a fair comparison, we added two models with different architectures. One is small MobileNet V3 and second is VGG16 model. Those architectures was on top of ImageNet benchmark performance graph.

As we can see ResNet-50 looks as a reasonable compromise between high performance and fast training.

4.5. Test on a data outside the dataset

Our network is generalized enough to classify road scene images that are not from the KITTI dataset on a relatively good level, which is presented in Fig. 12. Those images have been extracted from the road trip videos available on YouTube. The Fig. 13 presented examples of scenes, which our model not correctly classified. On the left picture scene with the snow, which our model has never seen snow before and classify picture as urban, which is the most diverse class in the training set. In the middle picture, the highway is most probably predicted as urban due to the fact that the road has many road markings which are most typical for urban scenes. On the right picture, we can observe that the rural scene was predicted as urban, where in fact it is hard to say visually that it is an urban or rural scene, however, the model classified it as urban with a lower probability of about 70 percent.

5. Conclusion

In this work, we proposed a method for labeling road scenes in a semi-automatic fashion, which might be used to prepare metadata for automotive datasets.

Vehicle GPS location was used for scene classification together with street-level image features, extracted by the VGG16-Places365 convolutional neural network, in order to generate labels to train residual networks with 50 convolutional layers to classify urban/rural/highway conditions by a single image. The network has been pre-trained on the ImageNet dataset then the transfer learning technique was used to train it on a custom KITTI dataset.

We show that spatial data can be used together with a pre-trained deep network pre-trained on the Places365 dataset as a semi-automatic approach to label road scenes classification task, instead of performing labeling manually.

As a baseline for image classification, we use a residual neural network with 50 layers with skip connections to demonstrate 86% accuracy on an augmented and extended dataset for scene classification with three classes.

Various contextual factors should be taken into account for generalized and confident predictions of the road scene classification in which this picture was taken. Such factors as weather, light, geographical or man-made process can make a big influence on the vision-based perception system.

Taking into account all those factors it is planned to experiment with other state-of-the-art datasets where image conditions might be different and more diverse. Another promising direction would be to use unsupervised methods for scene classification based on clustering techniques using embedding features from convolution layers. We believe that scene condition labeling can be used for both online and offline applications such as semantic mapping of autonomous datasets and driving assistance.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

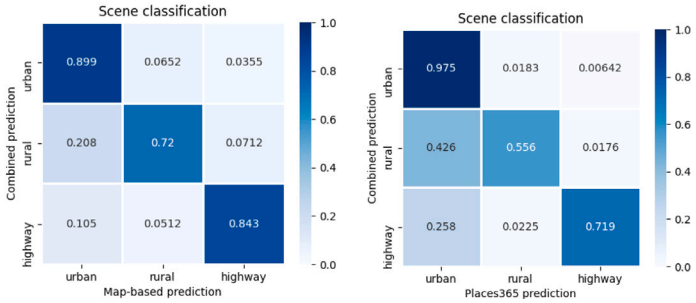

Fig. 7. Confusion matrix for scene classification: (a) combined predictions and map-based approach (b) combined predictions and image-based approach.

Fig. 8. Examples of correctly classified images: (a) urban predicted as urban (b) rural predicted as rural.

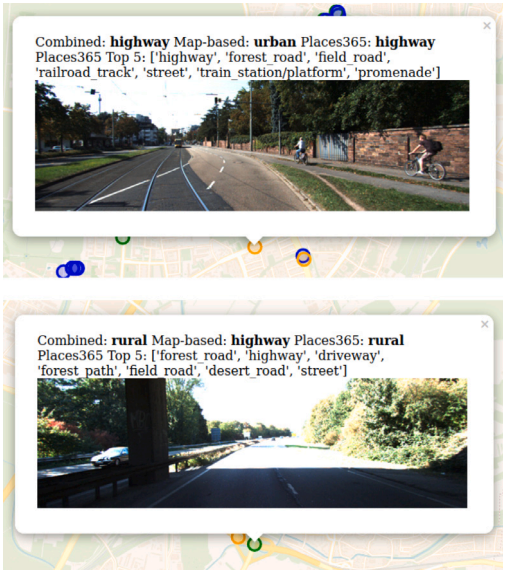

Fig. 9. Examples of wrongly classified images: (a) urban predicted as highway (b) highway predicted as rural.

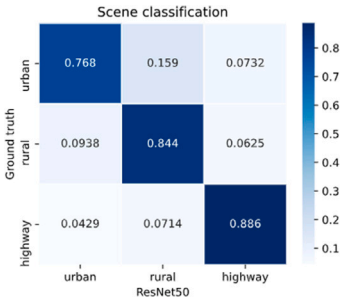

Fig. 10. Confusion matrix for pretrained ResNet-50 model trained on raw dataset.

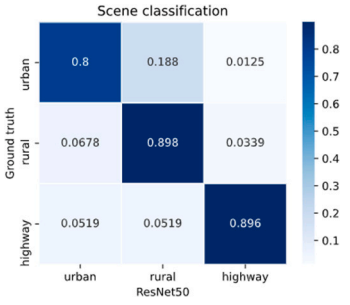

Fig. 11. Confusion matrix for pretrained ResNet-50 model trained on extended dataset.

Fig. 12. Some examples of correctly classified scenes from YouTube videos.

Fig. 13. Some examples of incorrectly classified scenes from YouTube videos.

References

[1] Everingham M, Van Gool L, Williams CK, Winn J, Zisserman A. The pascal visual object classes (voc) challenge. Int J Comput Vis 2010;88(2):303–38.

[2] Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. Imagenet large scale visual recognition challenge. Int J Comput Vis 2015;115(3):211–52.

[3] Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. Ieee; 2009, p. 248–55.

[4] Zhou B, Lapedriza A, Xiao J, Torralba A, Oliva A. Learning deep features for scene recognition using places database. 2014.

[5] Zhou B, Lapedriza A, Khosla A, Oliva A, Torralba A. Places: A 10 million image database for scene recognition. IEEE Trans Pattern Anal Mach Intell 2017;40(6):1452–64.

[6] Rangel JC, Cazorla M, García-Varea I, Martínez-Gómez J, Fromont E, Sebban M. Scene classification based on semantic labeling. Adv Robot 2016;30(11–12):758–69.

[7] Kostavelis I, Gasteratos A. Semantic mapping for mobile robotics tasks: A survey. Robot Auton Syst 2015;66:86–103.

[8] Wang S, Li Y, Liu W. Multi-class weather classification fusing weather dataset and image features. In: CCF conference on big data. Springer; 2018, p. 149–59.

[9] Sikirić I, Brkić K, Bevandić P, Krešo I, Krapac J, Šegvić S. Traffic scene classification on a representation budget. IEEE Trans Intell Transp Syst 2019;21(1):336–45.

[10] Geiger A, Lenz P, Stiller C, Urtasun R. Vision meets robotics: The kitti dataset. Int J Robot Res 2013;32(11):1231–7.

[11] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016, p. 770–8.

[12] Van De Sande K, Gevers T, Snoek C. Evaluating color descriptors for object and scene recognition. IEEE Trans Pattern Anal Mach Intell 2009;32(9):1582–96.

[13] Zhang J, Marszałek M, Lazebnik S, Schmid C. Local features and kernels for classification of texture and object categories: A comprehensive study. Int J Comput Vis 2007;73(2):213–38.

[14] Krapac J, Verbeek J, Jurie F. Modeling spatial layout with fisher vectors for image categorization. In: 2011 international conference on computer vision. IEEE; 2011, p. 1487–94.

[15] Sánchez J, Perronnin F, Mensink T, Verbeek J. Image classification with the fisher vector: Theory and practice. Int J Comput Vis 2013;105(3):222–45.

[16] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Adv Neural Inf Process Syst 2012;25:1097–105.

[17] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014, arXiv preprint arXiv:1409.1556.

[18] Tang P, Wang H, Kwong S. G-MS2F: GoogleNet based multi-stage feature fusion of deep CNN for scene recognition. Neurocomputing 2017;225:188–97.

[19] Bosch A, Munoz X, Oliver A, Marti R. Object and scene classification: What does a supervised approach provide us? In: 18th international conference on pattern recognition, Vol. 1. IEEE; 2006, p. 773–7.

[20] Ess A, Müller T, Grabner H, Van Gool L. Segmentation-based urban traffic scene understanding. In: BMVC, Vol. 1. Citeseer; 2009, p. 2.

[21] Di S, Zhang H, Mei X, Prokhorov D, Ling H. A benchmark for cross-weather traffic scene understanding. In: 2016 IEEE 19th international conference on intelligent transportation systems. IEEE; 2016, p. 2150–6.

[22] Di S, Zhang H, Li C-G, Mei X, Prokhorov D, Ling H. Cross-domain traffic scene understanding: A dense correspondence-based transfer learning approach. IEEE Trans Intell Transp Syst 2017;19(3):745–57.

[23] Narayanan A, Dwivedi I, Dariush B. Dynamic traffic scene classification with space-time coherence. In: 2019 International conference on robotics and automation. IEEE; 2019, p. 5629–35.

[24] Xia J, Ding Y, Tan L. Urban remote sensing scene recognition based on lightweight convolution neural network. IEEE Access 2021;9:26377–87.

[25] Zhao W, Bo Y, Chen J, Tiede D, Blaschke T, Emery WJ. Exploring semantic elements for urban scene recognition: Deep integration of high-resolution imagery and OpenStreetMap (OSM). ISPRS J Photogramm Remote Sens 2019;151:237–50.

[26] Haklay M, Weber P. Openstreetmap: User-generated street maps. IEEE Pervasive Comput 2008;7(4):12–8.

[27] Alirezaie M, Längkvist M, Kiselev A, Loutfi A. Open GeoSpatial data as a source of ground truth for automated labelling of satellite images. In: The 9th international conference on geographic information science. CEUR Workshop Proceedings; 2016, p. 5–8.

[28] Ludwig C, Hecht R, Lautenbach S, Schorcht M, Zipf A. Mapping public urban green spaces based on OpenStreetMap and sentinel-2 imagery using belief functions. ISPRS Int J Geo-Inf 2021;10(4):251.

[29] Liu S, Tian G, Xu Y. A novel scene classification model combining ResNet based transfer learning and data augmentation with a filter. Neurocomputing 2019;338:191–206.

[30] Howard A, Sandler M, Chu G, Chen L-C, Chen B, Tan M, et al. Searching for mobilenetv3. In: Proceedings of the IEEE/CVF international conference on computer vision. 2019, p. 1314–24.