Free Space Detection with Deep Nets for Autonomous Driving

Joel Pazhayampallil
Stanford University
Stanford, CA
jpazhaya@stanford.edu

Abstract

In this paper we train a network based on the GoogLeNet and OverFeat architectures to detect free road surface in highway settings. This can potentially be used as a computer vision based sensor for autonomous vehicles to detect generic obstacles and safe driving surfaces. The trained network performs very well on the free space detection task, achieving an F1 score of 0.9912 on a testing set of 1,300 images of highway driving.

1. Introduction

Autonomous driving has the potential to greatly reduce traffic accidents, road congestion, and associated economic loss. Safe autonomous driving requires detection of surrounding obstacles, moving objects, and identifying drivable areas. This must be done for all nearby objects for urban driving as well as objects at a distance for high speed highway driving. Current autonomous vehicles use a mixture of radar, lidar, camera, high precision GPS, and prior map information to determine obstacles around the car and safe driving areas on the road surface [3][8]. However, the critical sensor in these autonomous vehicles is the Velodyne multi-beam laser sensor. This high performance sensor enables detection and classification of objects as well as drivable surface [7]. Unfortunately, the sensor costs about $75,000 which makes it impractical for consumer vehicles. Additionally, the sensor faces significant performance degradation in rain, fog, and snow, limiting its applicability to fair weather. A vision based system could complement the other sensors by providing long range object and road surface detection and classification. However, extracting this information from camera images remains a difficult problem [1].

Given the recent success of deep convolutional neural networks in computer vision tasks [6] [5], DNNs can be a good candidate for tackling the perception challenges in autonomous driving. Prior work to apply CNNs to perception tasks in autonomous driving has produced specialized nets that detect other vehicles and lanes in the camera frame. However, the variety of possible road obstacles and road structure make it impractical to train specific networks for each possible obstacle and road scenario. In particular, obtaining training data to cover all possible scenarios would be very difficult to obtain. Instead, a more generic detector to determine free road surface that is safe to drive over would be preferred.

2. Problem Statement

In this paper, the network is tasked with detecting free road surface given a single 640 by 480 resolution camera frame. Each 4 by 4 pixel non-overlapping patch of the image is classified as free road surface or not. A patch should be classified as free road surface if it contains a part of the road that the vehicle can safely drive on. In particular, the patch should not contain an obstacle *e.g.* another vehicle, or surfaces that the vehicle cannot drive on, *e.g.* sidewalks/shoulders.

The dataset is composed of over 30 hours of 1 megapixel resolution video, recorded from a front facing camera on a vehicle driven on bay area highways. The data is annotated with lane and car labels. The lane labels (see Figure 1) indicate the position of all lanes on the highway within 80 meters of the vehicle, including lanes that are occluded by other vehicles. The lane labels were generated from camera and lidar point cloud data. The lanes were first extracted automatically, then reviewed and corrected by human labellers. The car labels (see Figure 2) consist of bounding boxes of all visible and partially occluded vehicles within about 100 meters of the vehicle (see Figure 3 for final combined label). The car labels are generated by a team of human labellers through Amazon Mechanical Turk.

From this dataset, free drivable space is defined as areas between two lane labels that do not contain a vehicle. The network is expected to produce a pixel mask that indicates the same areas as defined by above as free space. Qualita-

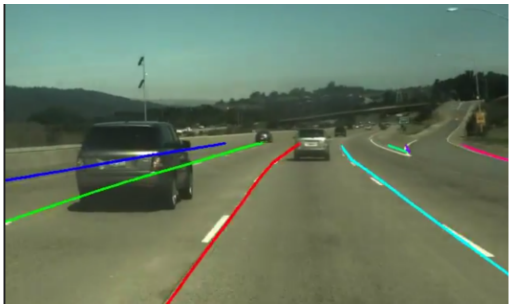

Figure 1. Example of lane labels in dataset. All lane markings within 80 meters are labelled, regardless of occlusion.

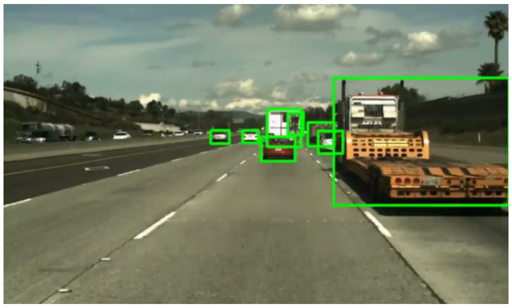

Figure 2. Example of car labels in dataset. All fully visible and partially occluded vehicles within 100 meters are labelled with bounding boxes.

tive results will show test driving images with the drivable road surface indicated. Quantitatively, the results will be evaluated according to the F1 score of correctly classified image patches. We intuitively expect network performance to decrease with increasing distance in front of the vehicle. Ideally, the network will be able to identify free road surface in a variety of road structures and with a variety of obstacle types. The final dataset contains 13,000 training images, 1,300 validation images, and 1,300 test images.

3. Technical Approach

The network uses a modified GoogLeNet [6] style network to generate image features. The modified network follows the GoogLeNet architecture only up to the average pooling layer. The average pooling works well for the ImageNet classification task, but the free space detection task requires local information which would be lost by average pooling reducing the activation volume to a single activation vector.

Localization of free space in the image is accomplished with a modified OverFeat [4] architecture, with the input image features as generated above. The localization is accomplished by performing 1 by 1 convolutions over the image feature activation volume. Then a similar fully connected layer followed by softmax classification is used to determine the correct class. It should be noted that the spatial dimensions of the image feature activation volume are quite small relative to the original image size. Therefore, performing classification over the spatial dimension of the image feature activation volume would provide very poor localization of the detected class within the original image. To address this, additional depth channels are used in the fully connected layers to represent specific locations in the image. For example, the final softmax activation volume would have spatial dimension of 20 by 15, where each location represents a 32 by 32 pixel patch in the original image. However, each spatial location has depth of 128, since there are 8 by 8 patches of 4 by 4 pixels within each 32 by 32 pixel patch in the original image and 2 classes, free space and not free space. This enables localization of the free space patch to a much finer resolution while still being able to gain contextual information from a larger patch of the original image.

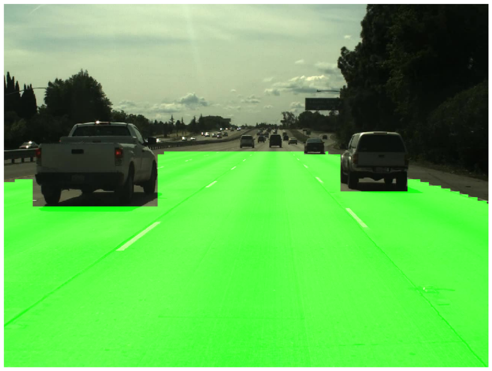

Figure 3. Example of combined car and lane labels showing free road surface label. The green shaded areas show the area of image labelled as free space.

The network implemented within the Caffe framework [2] with some modifications. The network is trained on a single Nvidia GeForce GTX TITAN Black GPU. The network is initialized with weights from the BVLC GoogLeNet trained on ImageNet and fine tuned with the free space dataset.

4. Results

The model was trained with mini batch stochastic gradient descent with momentum of 0.9. Initial learning rate was set to 0.01 and reduced by a factor of 0.96 every 3200 iterations. The final model was trained for 60,000 iterations with a batch size of 10, or 46 epochs. Typical qualitative results from the test set are shown in Figure 4 and 5.

The images show that the network is very successful in

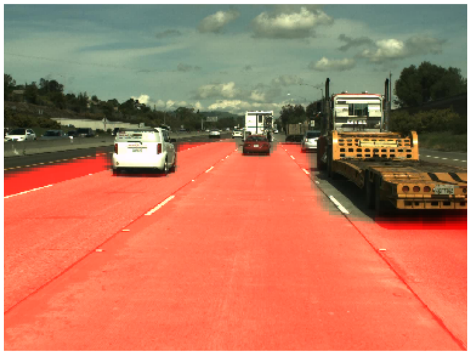

Figure 4. The red shaded regions show the softmax score of the correct class for that pixel patch in the image. The network successfully predicts the open road surface as free while indicating the truck and vehicles as not free space.

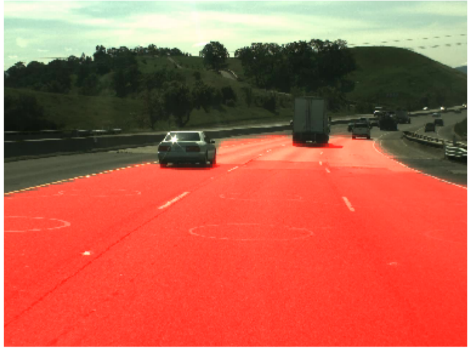

Figure 5. The red shaded regions show the softmax score of the correct class for that pixel patch in the image. The network successfully predicts the open road surface as free space. Notice that the road shoulder on both sides are not labelled as free even though these two areas are visually very similar.

correctly classifying obstacles in the road as not free space. It also successfully distinguishes between road lanes and road shoulders even though these two road surfaces are quite similar visually. The local pixel information is insufficient to distinguish between road lanes and road shoulder so the higher level features in higher layers of the network are necessary to make this distinction. The network also successfully follows curves in the road. A video of the networks performance on a video clip from the test set can be viewed here: `https://youtu.be/b-e4YhBn6Bo`.

To quantitatively measure performance the network performance, each pixel patch in each image was treated as an individual classification, resulting in $120 \times 160 \times 1300$ classifications for the test set. The F1 metric used proved to not be very useful in tracking performance of network training. All networks trained from 2,000 to 60,000 iterations had F1 scores of about 0.99. The final network trained with 60,000 iterations achieved an F1 score of 0.9912 and accuracy of 0.9926.

The networks can be differentiated qualitatively as detections around other vehicles and on the road edges become crisper. However these improvements are not measured well by the F1 metric since they occupy only a few pixel patches. These subtle differences are overwhelmed by the large number of correct classifications on the bulk of the road surface and on the top of the image containing mostly sky and background.

5. Conclusion

The network trained in this paper achieves good results in detecting free road surface in highway settings. These results were achieved by starting with a GoogLeNet style architecture pretrained on ImageNet data and fine tuned with the free space data. With relatively little training, only 60,000 iterations, the network was able to correctly recognize free road surface while indicating other vehicles as obstacles. The network was also able to distinguish the road boundary from other road markings.

In future work, we intend to collect more data in more varied conditions. Particularly urban situations involving many more obstacle types like pedestrians. Given the results of this experiment, extending this network architecture to general free space detection in much more complicated urban situations will likely be successful.

Furthermore, a more appropriate quantitative metric must be developed to better distinguish network improvement. While training, the network quickly learns to detect the bulk of the free road surface. In our dataset, most of the area at the bottom of the image is free road surface, while most of the area at the top of the image is not road surface. However, the more interesting areas are the boundaries between free road surface and not. Further training helps the network make better predictions at these boundaries, eventually providing very clear boundaries between free space and obstacles or the road edges. In future work, a network trained on only the boundaries of free space might provide better results. In particular, the training of this network will focus the entire loss function on the boundaries instead of weighting all pixel patches equally.

References

[1] D. Held, J. Levinson, and S. Thrun. A probabilistic framework for car detection in images using context and scale. In *Robotics and Automation (ICRA), 2012 IEEE International Conference on*, pages 1628–1634. IEEE, 2012.

[2] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolu-

tional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[3] M. Montemerlo, J. Becker, S. Bhat, H. Dahlkamp, D. Dolgov, S. Ettinger, D. Haehnel, T. Hilden, G. Hoffmann, B. Huhnke, et al. Junior: The stanford entry in the urban challenge. *Journal of field Robotics*, 25(9):569–597, 2008.

[4] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229*, 2013.

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

[6] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. *arXiv preprint arXiv:1409.4842*, 2014.

[7] A. Teichman, J. Levinson, and S. Thrun. Towards 3d object recognition via classification of arbitrary object tracks. In *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, pages 4034–4041. IEEE, 2011.

[8] C. Urmson, J. Anhalt, D. Bagnell, C. Baker, R. Bittner, M. Clark, J. Dolan, D. Duggins, T. Galatali, C. Geyer, et al. Autonomous driving in urban environments: Boss and the urban challenge. *Journal of Field Robotics*, 25(8):425–466, 2008.