# A Pedestrian And Vehicle Rapid Identification Model Based On Convolutional Neural Network

Ruochen Wang
Beijing University of Technology,
100124, Beijing, China
297348785@qq.com

Zhe Xu
Beijing University of Technology,
100124, Beijing, China
xu2002@bjut.edu.cn

## ABSTRACT

Image recognition technology based on convolutional neural network (CNN) has been widely used in the field of intelligent transportation in recent years. Since the image recognition in the field of intelligent transportation needs high real-time performance, this requires improving the speed of CNN. We refer to Overfeat, which was proposed in the ImageNet Large Scale Visual Recognition Challenge, to build a vehicle and pedestrian recognition model. We do not use the traditional sliding window method. Instead, we apply each convolution over the extent of the full image, eventually producing a map of output class predictions. This method ensures the accuracy of image recognition, while enhancing the operational efficiency and the real-time performance of CNN. In this paper, we use both a new method and the traditional sliding window method for the recognition of pedestrians and cars on the road. Then, we compare the advantages and disadvantages of the two methods in terms of their recognition effect and speed.

## Categories and Subject Descriptors

I.4.8 [**Scene Analysis**]: Object recognition

## General Terms

Algorithms

## Keywords

Image Recognition, Convolutional Neural Network

## 1. INTRODUCTION

In deep learning, a convolutional neural network (CNN) is a very common image recognition model. Many state-of-the-art image recognition models have been used in CNN [1,2]. CNN are an example of specialized neural network (NN) architectures that incorporate knowledge about the invariance of two-dimensional (2-D) shapes by using local connection patterns and by imposing constraints on the weights. A weight-shared network architecture greatly reduces the number of weights [3]. Images can be used directly as input, and the complex process of feature extraction from traditional identification algorithms can be avoided. Compared with the classical characteristics extraction algorithm like SIFT [4] or HOG [5], CNN can learn the required characteristics. This avoids the shallow learning artificial selection features' many drawbacks. Moreover, CNN have a total height of invariance for translation, scaling, tilting, or other forms of deformation.

In 2013, in the ImageNet Large Scale Visual Recognition Challenge [6], Pierre Sermanet, Yann LeCun et al. proposed a new model titled Overfeat [7]. This represent-ed an improvement of the sliding window method for image recognition; the model applies each convolution over the extent of the full image, eventually producing a map of output class predictions. Then, Overfeat use 1 to 5 convolution network layers as a classifier coupled with three regression layers constructed as a locator. To overcome the problem of the resolution declining in the pooling, Overfeat takes an approach similar to that introduced by Giusti et al [8]. Because Overfeat imports full images, it does not configure a separate detector to detect the region of interest. In this experiment, we reference the Overfeat to build an image recognition model based on CNN to recognize an image in a transport field. Our findings indicate that this model doubles the recognition speed compared to the traditional sliding window method.

Compared to the classifier recognition of the entire image by sliding windows, we imported full image can enhance the recognition speed. However, traditionally, CNN add a separate detector before the classifier to detect the region of interest and then use the classifier for recognition. We imports full images; this means that the classifier indiscriminately recognizes the full image. So, we let the detector apply each convolution over the extent of the full image, producing an area-of-interest map. Then we use a separate classifier for classification. This further improves the recognition speed, improving the real-time performance of the model.

## 2. NEURAL NETWORK MODEL

CNN is generally divided into two parts: convolution layers and fully connected layers. The convolution layers are responsible for extracting features, and turning abstract low-level features into high-level features. The fully connected neural network's task is to classify image by the use of high-level features. However, with the increase in the number of convolution layers, the speed of the whole network will be greatly reduced. Therefore, we took some measures to enhance the speed of image recognition.

### 2.1 Sliding Window in High-Level Features

Traditionally, an image recognition model first uses a detector to determine the approximate location of the area-of-interest. Following this step, we use the classifier to determine the category of the target and finally, its precise positioning. A CNN can handle a fixed size image. So detector and classifier built by CNN input a fixed size image. In contrast, the scene image will be

*ICIMCS '15, August 19-21, 2015, Zhangjiajie, Hunan, China*
© 2015 ACM. ISBN 978-1-4503-3528-7/15/08…$15.00
DOI: http://dx.doi.org/10.1145/2808492.2808524

much larger. Therefore, a sliding window is often used to process the scene image.

A sliding window generally continues to take some fixed-size blocks input CNN from the scene image to get the consequence of detection or classification in each position, as shown in Figure 1. This causes a lot of double counting, as there are many duplicated parts among the sliding window's acquired image, and this duplication can be avoided. For CNN, most of the calculations are coming from calculation convolution. The network structure cannot be changed, but the size of the characteristic pattern in each layer can be changed. Overfeat applies each convolution over the extent of the full image, eliminating most of the repetitive computation. Our model also refers to the practice of Overfeat, applying each convolution over the extent of the full image.

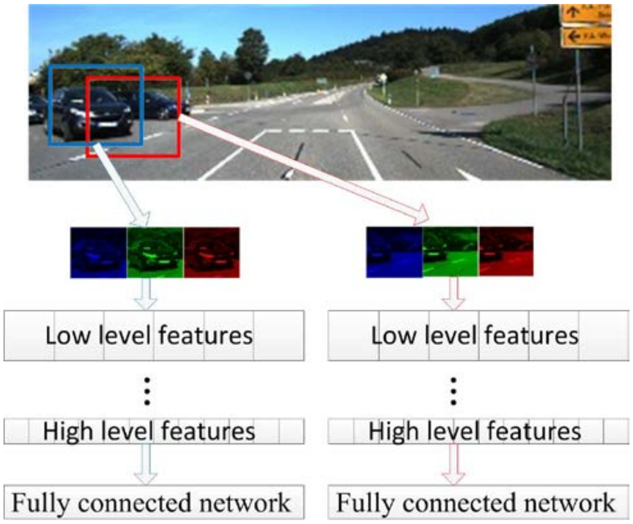

**Figure 1: CNN traditional model based on a sliding window. This program uses a sliding window on the scene image, receiving a fixed sample size image input to a CNN. As an identical pixel will be sampled by several sliding windows, this causes a significant amount of waste, thereby reducing the computational speed of the entire CNN.**

When recognizing the actual scene, in order to produce a map of output class predictions, the NN is actually split into two parts. The first part is made up of 1 to 5 convolution network layers. It is responsible for the convolution of the whole picture, obtaining feature maps that are much larger than those used in training. Then, using a similar approach to sliding windows for this feature map, we obtain a stand-ard size of data for the second part of the fully connected network. As shown in Figure 2, the part connected to the full convolution and connection layers is also uses sliding window. However, unlike ordinary sliding windows, the sliding window of our model is not in the bottom of the pixel-level image, but instead in the high-level features that are extracted by the convolution layer.

## 2.2 The Structure of Convolution Network

Overfeat did not join the background image for training, as it uses the ImageNet data set and there are as many as 1000 classes of the sample. In our model, the positive sample category is far smaller than ImageNet. We find that if we do not join the background image as a negative sample, our model will produce a high error rate when it handles background. Therefore, we randomly selected part of the background image as a negative sample added to the training set.

Next, we build four image recognition models to compare them with the traditional model in recognition performance and real-time performance, as shown in Figure 3. Due to the high need for real-time performance, the size of our network is relatively small. The first model uses a fully connected NN sliding window in the high-level feature map. It only contains a larger convolution network and both detection and classification functions. The structure of CNN is shown in Table 1. The second model establishes a detector and a classifier.

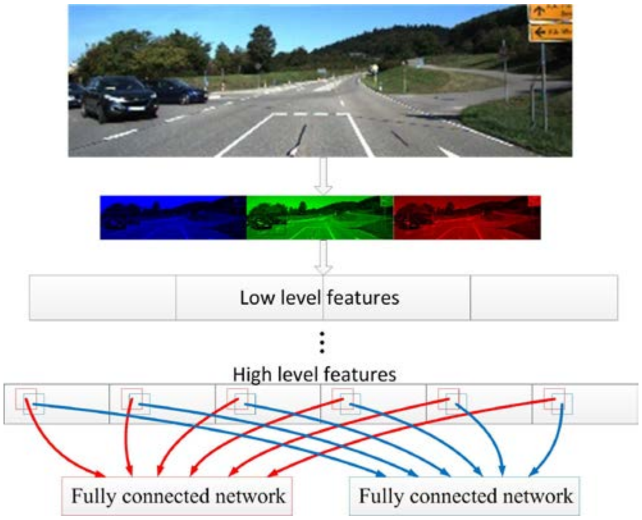

**Figure 2: Our image recognition model. During the feature extraction, the whole image is convolved, resulting in a large map of the high-level features. This avoids a lot of repetitive computation. We then use a fully connected NN sliding window in the high-level feature map.**

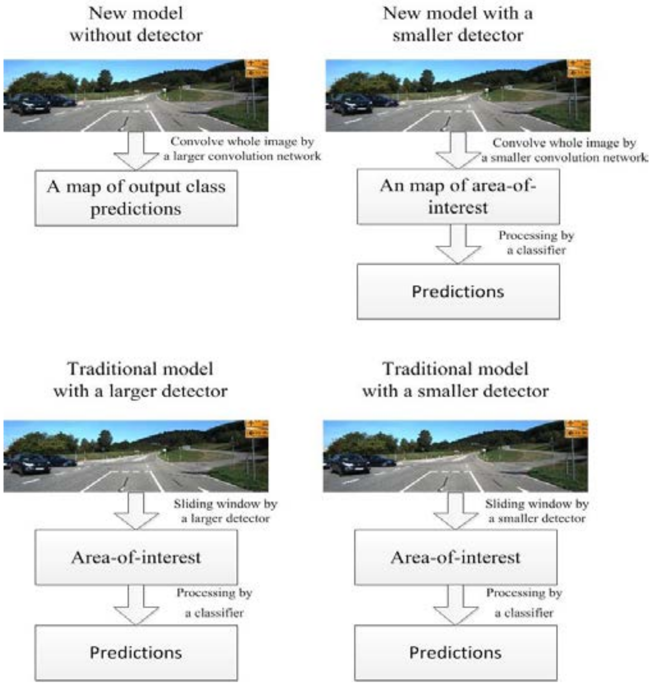

**Figure 3: four image recognition models.**

In fact, since the task of the detector is relatively simple, its network structures tend to be much simpler than those of the classifier. As such, we can considerably reduce the recognition time. The detector network model is shown in Table 2, and the classifier continues to use a relatively large network model, as shown in Table 1. The detector convolutes the whole image, taking a sliding window in a high-level feature map to determine the area-of-interest. The last two models are the traditional image recognition model and use sliding windows on the original image. They used to be compared with the previous two models. One model uses both larger CNN on the detector and classifier, as shown in Table 1. The other uses a smaller detector, as shown in Table 2, and a larger classifier.

In the training of the network, we used the ADADELTA [9] method. This is a first-order adaptive learning rate approach that greatly improves the efficiency of NN training. We also added weight decay to enhance the generalization ability [10]. This model does not require a high recognition resolution; therefore, we did not use the Δ pooling.

## 3. EXPERIMENTAL AND RECOGNITION RESULTS

We used the KITTI Vision Benchmark Suite [11] as the data set. This comprises data from autonomous driving platforms provided by the Karlsruhe Institute of Technology and Toyota Technological Institute (KITTI). This data set is captured by driving around the mid-size city of Karlsruhe, in rural areas and on highways. In order to facilitate training, this experiment compressed the sample into a unified picture of 50*50.

We attempted to reduce unnecessary differences between the four models and to enhance validity. Among each experiment, all models shared the same convolution layer initialization parameters. At the same time, the weight decay  was 0.01 in both cases.

Despite these similarities, there are still some inevitable differences between them. For the CNN in which detection and classification are carried out together, a negative output and corresponding samples had to be added for training. The other model uses these negative samples and the rest of the sample to train the detector. Because convolution of the full image does not impact the validation set error, we only compared the new model without detector and the traditional model with larger detector in terms of their validation set error rate. Using different

**Table 1: Larger CNN structure.**

| Layer | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| Stage | conv+max | conv+max | conv | conv | conv+max | full | full | full |
| Channels | 8 | 12 | 16 | 10 | 6 | 24 | 12 | 10 |
| Filter size | 6*6 | 3*3 | 3*3 | 3*3 | 3*3 | - | - | - |
| Conv. Stride | 1*1 | 1*1 | 1*1 | 1*1 | 1*1 | - | - | - |
| Pooling size | 2*2 | 2*2 | - | - | 2*2 | - | - | - |
| Pooling stride | 2*2 | 2*2 | - | - | 2*2 | - | - | - |
| Spatial input size | 50*50 | 22*22 | 10*10 | 8*8 | 6*6 | - | - | - |

**Table 2: Smaller CNN structure.**

| Layer | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| Stage | conv+max | conv+max | conv | conv | conv+max | full | full | full |
| Channels | 6 | 5 | 4 | 3 | 3 | 12 | 8 | 6 |
| Filter size | 6*6 | 3*3 | 3*3 | 3*3 | 3*3 | - | - | - |
| Conv. Stride | 1*1 | 1*1 | 1*1 | 1*1 | 1*1 | - | - | - |
| Pooling size | 2*2 | 2*2 | - | - | 2*2 | - | - | - |
| Pooling stride | 2*2 | 2*2 | - | - | 2*2 | - | - | - |
| Spatial input size | 50*50 | 22*22 | 10*10 | 10*10 | 6*6 | - | - | - |

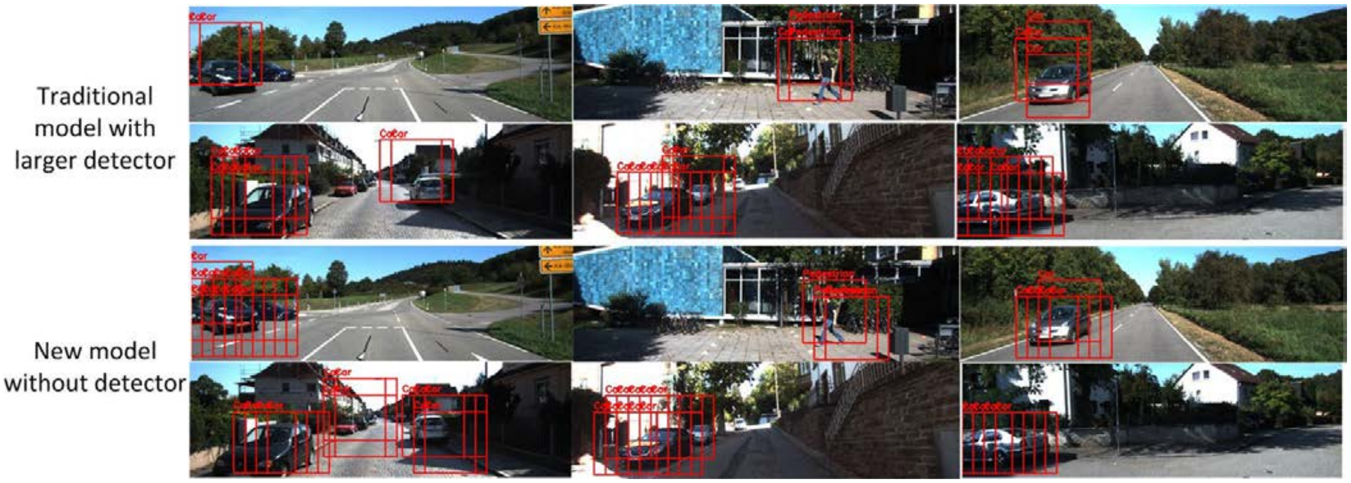
Traditional model with larger detector
New model without detector

**Figure 4: The actual recognition results.**

initialization parameters, we obtained ten groups of comparative data. The recognition error on the validation set is shown in Figure 5.

It can be seen that the models were trained using the same initialization parameters, training set, and stop conditions, but the overall error rate of the new model without detector was lower than the traditional model with larger detectors by about 4%. This may be due to relatively small target category in our experiment. In the actual scene, the recognition effect of these two models is shown in Figure 4.

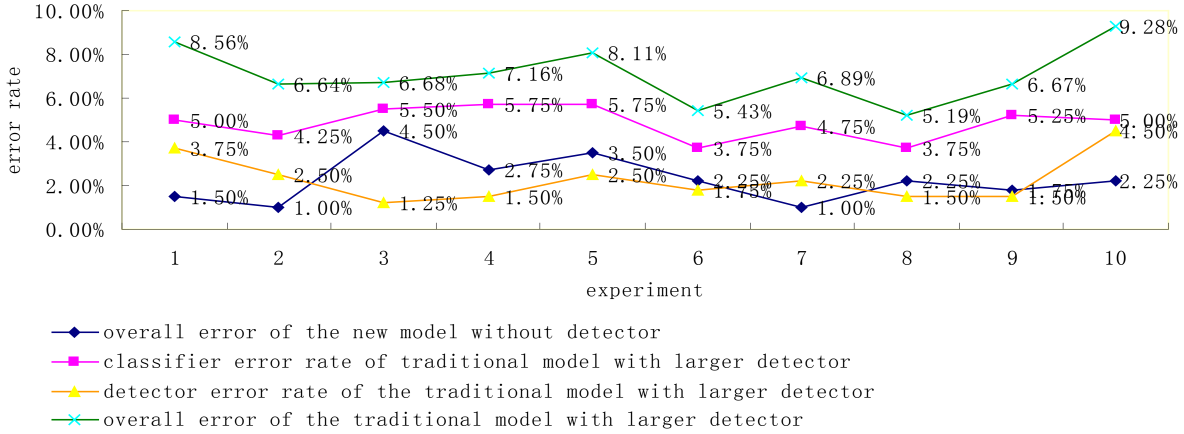

**Figure 5: The validation set error rate.**

In the speed of recognition, this experiment compared all four models with 200 images at a resolution of 1242*375 under the same hardware condition. The new model without a detector convoluted full images in 120 seconds. The model with a smaller detector convoluted full images in 105 seconds. The traditional model using a sliding window in the original picture with larger and smaller detectors took 265 seconds and 195 seconds, respectively. To compare them with the traditional model, the relative times of the recognition models are shown in Figure 6.

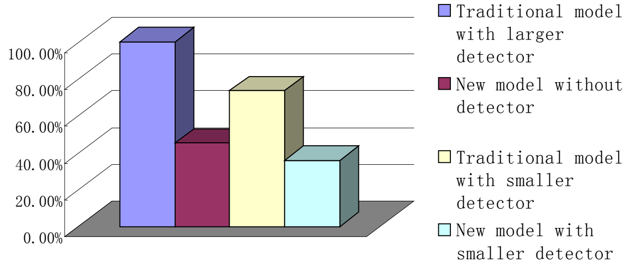

**Figure 6: The relative time of each recognition model.**

## 4. CONCLUSIONS

We reference Overfeat to build a pedestrian and vehicle identification model based on CNN. Compared to the traditional model, the new model significantly improves the real-time model's performance, and it has a lower recognition error rate at fewer target categories. Further, if we set up a separate detector and classifier and convolute the full image using the detector, real-time performance can significantly increase. In terms of accuracy, this was identical to the sliding windows method. When the image recognition model runs on a limited-performance hardware platform such as ARM or DSP, a faster recognition speed is undoubtedly important.

## 5. REFERENCES

[1] Girshick R, Donahue J, Darrell T, et al. “Rich feature hierarchies for accurate object detection and semantic segmentation.” arXiv preprint arXiv: 1311.2524, 2013.

[2] Szegedy C, Liu W, Jia Y, et al. “Going deeper with convolutions.” arXiv preprint arXiv:1409.4842, 2014.

[3] LeCun, Yann, Léon Bottou, Yoshua Bengio, and Patrick Haffner. “Gradient-based learning applied to document recognition.” Proceedings of the IEEE 86, no. 11 2278-2324, 1998.

[4] Lowe D G. “Distinctive image features from scale-invariant keypoints.” International journal of computer vision, 60(2): 91-110, 2004.

[5] Dalal N, Triggs B. “Histograms of oriented gradients for human detection.” CVPR 2005.

[6] Deng J, Dong W, Socher R, et al. “Imagenet: A large-scale hierarchical image database.” CVPR 2009.

[7] Sermanet, Pierre, David Eigen, Xiang Zhang, Michaël Mathieu, Rob Fergus, and Yann LeCun. “Overfeat: Integrated recognition, localization and detection using convolutional networks.” arXiv preprint arXiv: 1312.6229, 2013.

[8] A. Giusti, D. C. Ciresan, J.Masci, L.M. Gambardella, and J. Schmidhuber. “Fast image scanning with deep max-pooling convolutional neural networks.” In ICIP, 2013.

[9] Zeiler, Matthew D. "ADADELTA: An adaptive learning rate method." arXiv preprint arXiv: 1212.5701, 2012.

[10] Moody J E, Hanson S J, Krogh A, et al. “A simple weight decay can improve generalization.” Advances in neural information processing systems, 4: 950-957, 1995.

[11] Geiger, Andreas, Philip Lenz, and Raquel Urtasun. “Are we ready for autonomous driving? The KITTI vision benchmark suite.” In CVPR, 2012.