Sensing, Perception and Decision for Deep Learning Based Autonomous Driving

Takayoshi Yamashita(✉)

Chubu University, 1200, Matsumoto-cho, Kasugai, Aichi, Japan
takayoshi@isc.chubu.ac.jp

Abstract. Toward the realization of autonomous driving, deep learning has attracted the most attention, and it is seen as indispensable technology. AlexNet which consists of eight layers and incorporating ideas for improving generalization, was able to accomplish substantial improvement in accuracy with image recognition task. Since then, not only image recognition but also various tasks and applications to various fields are dramatically advanced. Even in the autonomous driving field, many manufacturers have taken aggressive efforts and are pushing ahead with practical application. In this paper, we will introduce the tasks being tackled for autonomous driving. The tasks introduced here are object detection, human pose estimation, and semantic segmentation from images and other sensors. By combining these methods, it is possible to realize safer automatic operation system.

Keywords: Deep learning · Object detection
Semantic segmentation · CNN

1 Introduction

In the latter half of the 1990 s, with the evolution of a general-purpose computer, since it became possible to process a large amount of data at high speed, a feature amount vector called image local feature amount was extracted from the image and an image Methods to realize recognition have become mainstream. Machine learning requires a large amount of samples with class labels, but since researchers do not need to design several rules as in the rule-based method, image recognition with high versatility can be realized. In the 2000s, features such as Scale-Invariant Feature Transform (SIFT) and Histogram of Oriented Gradients (HOG), which were designed based on the hand crafted, had studied as image local features. And in the 2010s, deep learning which automatically acquires feature extraction process by learning is attracted attention. Handcrafted feature is not necessarily optimal because it extracted and expressed features based on human knowledge. Deep learning is a new approach that can automatically extract effective feature for recognition. Image recognition by deep learning has

N. Streitz and S. Konomi (Eds.): DAPI 2018, LNCS 10922, pp. 152–163, 2018.
https://doi.org/10.1007/978-3-319-91131-1_12

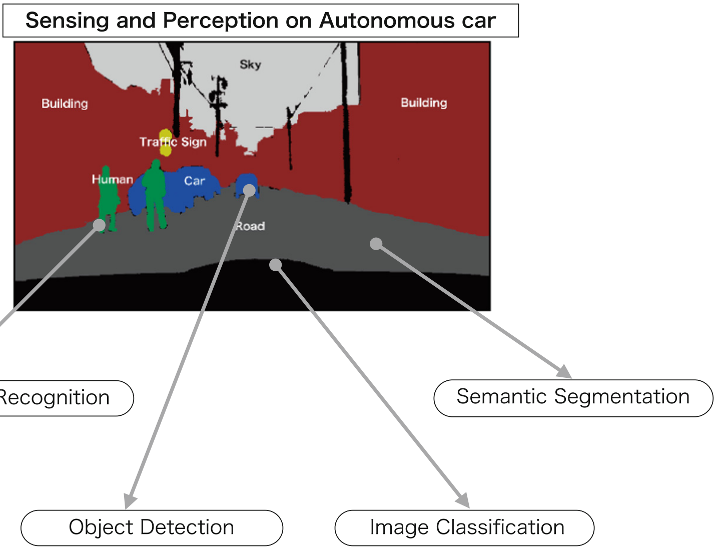


Fig. 1. Subdividing tasks on autonomous driving.

overwhelming results in object recognition challenge and since then, deep learning approach is applied in various fields is progressing. In this paper, we explain how deep learning is applied in the computer vision and autonomous driving, how it is solved problems, and the latest trend of deep learning.

2 Problem Setting in Autonomous Driving

In conventional machine learning, it is difficult to solve general object recognition directly from an input image as shown in Fig. 1, so this problem has been solved by subdividing it into each task of image classification, object detection, scene understanding and specific object recognition. The following describes the definition of each task and the approach to each task.

2.1 Image Classification

Image classification is a task that recognized the belonging class of image. In conventional machine learning, an approach called Bag-of-Features (BoF) [4] has been used which vector-quantizes the image local feature amount and expresses the feature of the entire image as a histogram. Thereafter, the feature includes the Fisher vector which expresses richer information and the Vector of Locally Aggregated Descriptors (VLAD) which reduces the amount of memory have been proposed [5,6]. On the other hand, deep learning has achieved accuracy exceeding human recognition performance in 1,000 class image classification task.

2.2 Object Detection

Object detection is a task of finding where objects of a certain class are in the image. Face detection and pedestrian detection are included in this task. Combinations of Haar-like feature and AdaBoost for face detection [2], and combination of HOG feature and Support Vector Machine (SVM) for pedestrian detection [3] are widely used. In conventional machine learning, object detection is realized by learning a two-class classifier corresponding to a certain category and performing raster scan within the image. Object detection by deep learning can realize multi-class object detection targeting a plurality of categories with a single network.

2.3 Semantic Segmentation

Scene understanding is a task of understanding the scene structure in the image. In particular, semantic segmentation for recognizing object class for each pixel has been considered difficult task to solve by conventional machine learning. Therefore, although it was considered as one of the final problems of computer vision, it is beginning to show that it is a task that can be solved by application of deep learning.

2.4 Specific Object Recognition

Specific object recognition is a task of recognizing a specific object class by assigning attributes to objects having proper nouns and is defined as a subtask of general object recognition problem. Specific object recognition is realized by detecting feature points in a image such as SIFT [7] and calculating distance and voting process with feature points of the registration pattern. The Learned Invariant Feature Transform (LIFT) [8] replaces the process of SIFT with deep learning based approach, and realize performance improvement.

In the following, we focus on the tasks of image classification, object detection, scene understanding (semantic segmentation), and describe the application of deep learning and its trends.

3 Object Classification

On ImageNet Large Scale Visual Recognition Challenge (ILSVRC) which performs 1,000 class category classification, deep learning based approaches are winning after 2012. Convolutional Neural Network (CNN) [9] is a basis method with overwhelming results in such large-scale image recognition. AlexNet [10] is constructed 5 convolution layer, 3 fully connection layers and a output layer which outputs probabilities of 1,000 categories, as shown in the Fig. 2. Figure 3 is visualizing convolution filters of 1st layer of AlexNet which obtains from training a large amount of object images of 1,000 classes. These filters are consist of texture, color information and edge having direction components are automatically acquired by learning.

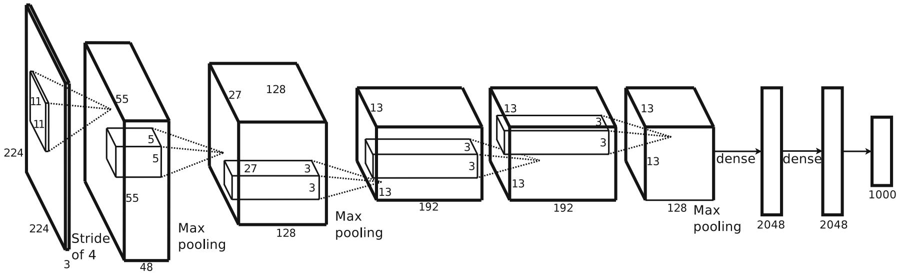


Fig. 2. Structure of AlexNet

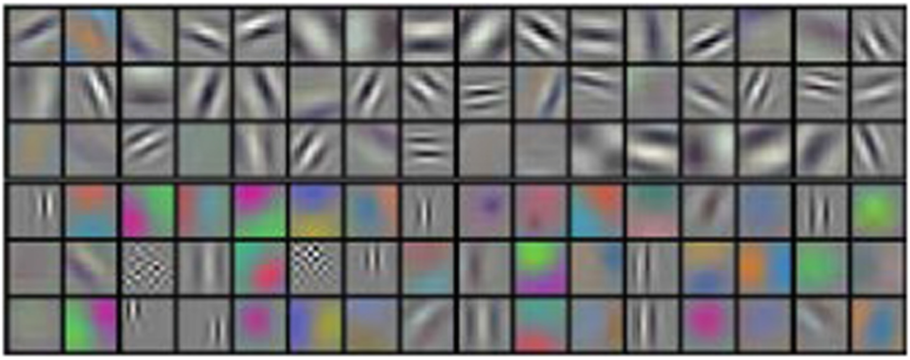


Fig. 3. Convolution filters of 1st layer of AlexNet [10].

Simonyan proposed more deeper network, called VGGNet, and improved classification performance [11]. GoogLeNet [12] has 22 layers with Inception modules. When the layer becomes deeper, the gradient becomes 0 in the middle and a gradient elimination problem occurs where reverse propagation cannot be performed. ResNet [13] proposed a method which has a route to backpropagate through the bypass, realizing 152 layers super deep structure. As a result, the error rate of ResNet improved to 3.56%. When the same task was performed by humans, the error rate was 5.1%, and the approach to deep learning gained recognition performance equal to or better than human ability. This approach is also apply to traffic sign recognition task and is comparable performance to human. Figure 4 shows the result of traffic sign recognition under a clutter background, dark image.

4 Object Detection

Conventional object detection is an approach to raster scan a classifier. In deep learning, R-CNN extracts object candidate regions by Selective Search [15] and extract efficient feature using AlexNet to perform multi-class classification [14].

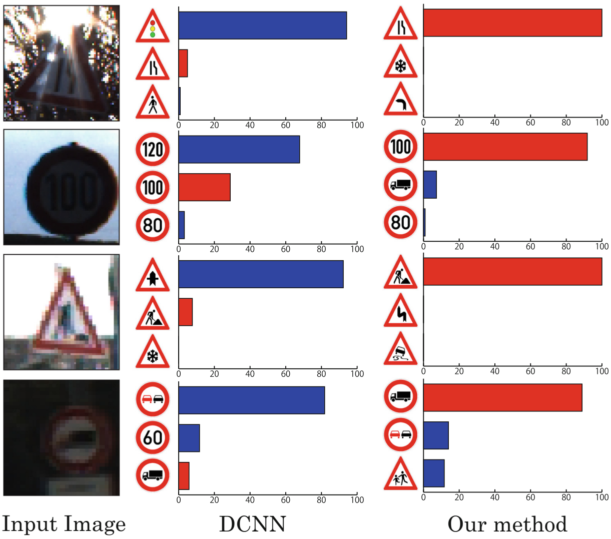


Fig. 4. Result of traffic sign recognition.

Selective search roughly segments an object candidate by repeatedly grouping regions with similar color and texture with various threshold values and detects object candidate regions. The point that object detection using CNN is realized and it is possible to detect multiclass is epochal, but since Selective Search integrates repeated regions when obtaining object candidate region, it is time consuming.

Faster R-CNN [16] introduces the Region Proposal Network (RPN) as shown in Fig. 5, and detects the object candidate region and recognizes the object class simultaneously. First, convolution processing is performed on the entire image to obtain a feature map. In RPN, an object is detected by sliding the detection window against the obtained feature map. RPN introduces detection method called anchor. The anchor is k number of detection windows that have

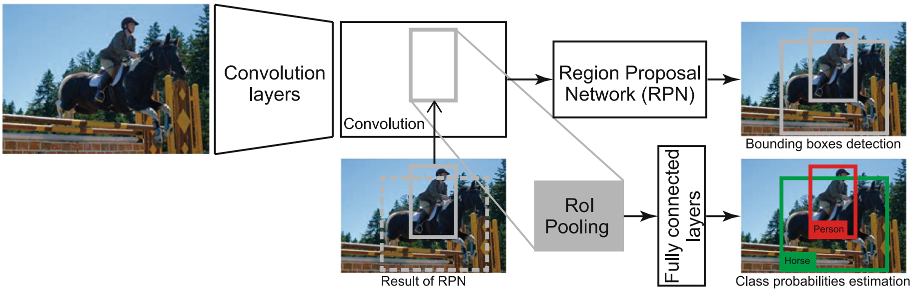


Fig. 5. Structure of Faster R-CNN.

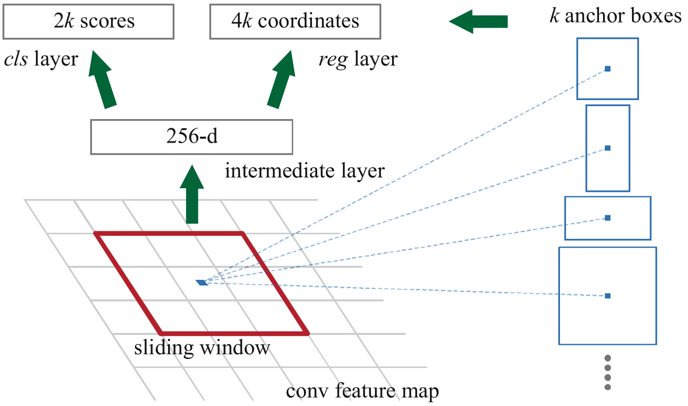


Fig. 6. Scanning using anchors [16].

different sizes and aspect ratios as shown in the Fig. 6. The output layer outputs the object score and the coordinates for the region specified by the anchor. In addition, the region also input to another fully connection layer, to recognize object class. By using these Region Proposal methods, it is possible to detect objects of multiple classes with different aspect ratios. Single Shot based method is attracting attention as a novel multi-class object detection approach. This is a method of detecting multiple objects by only giving the entire image to CNN without sliding window. YOLO (You Only Look Once) [17] is a representative method and detects multiple objects on each local region that is divided by a grid of 7 × 7 as shown in Fig. 7. First, a feature map is generated through convolution and pooling of the input image. The position (i, j) of each channel of the obtained feature map $(7 \times 7 \times 1024)$ is a region feature corresponding to the grid (i, j) of the input image, Enter the map in all tie layers. The output layer have units corresponding to object score, coordinates, object size and score of each category for each grid position.

Since YOLO detects objects along roughly defined grids, it is not robust to changes in scale, especially small object. The Single Shot Multi-Box Detector (SSD) [18] outputs scores of object coordinates and categories from several convolution layers as shown in Fig. 8. In SSD, small objects are detected in layers closer to the input layer, and large objects are detected in layers closer to the output layer. The feature map closer to the input layer is less affected by the reduction of the feature map due to pooling. The SSD outputs object coordinates and object categories for each position of the feature map. Therefore, it is unnecessary to estimate the object category via another network, so it is possible to perform fast object detection. Figure 9 shows the result of pedestrian detection by SSD.

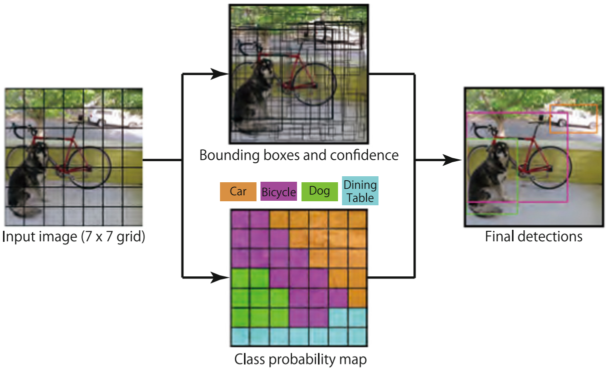


Fig. 7. Structure of YOLO

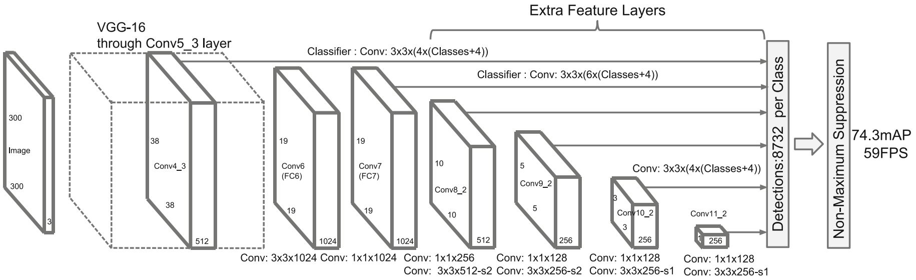


Fig. 8. Structure of SSD [18].

5 Semantic Segmentation

In the computer vision field, semantic segmentation is a task with high difficulty and has been studied for many years. And it was thought that it would take time to realize highly accurate semantic segmentation. However, similar to other tasks, a method based on deep learning has been proposed and achieves performance that exceeds the conventional method.

Fully Convolutional Network (FCN) [20] is a method capable of learning and labeling end-to-end only using CNN. The structure of FCN is shown in Fig. 10. The FCN is a network structure that does not have fully connection layer. By repeating the convolution layer and the pooling layer, the size of the feature map becomes smaller. In order to make it the same size as the original image, the feature map is upsampled 32 times in the final layer and convolution processing is performed. This is called deconvolution. The final layer outputs a probability map of each class to be labeled. When upsampling the feature map in this manner, coarse segmentation results are obtained. To obtain fine segmentation result, it integrates the intermediate feature maps. Generally, the feature maps that are closer to the input layer extract detailed information. These detail information

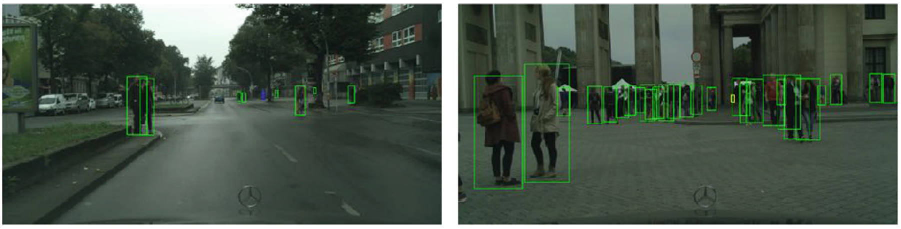


Fig. 9. Result of pedestrian detection.

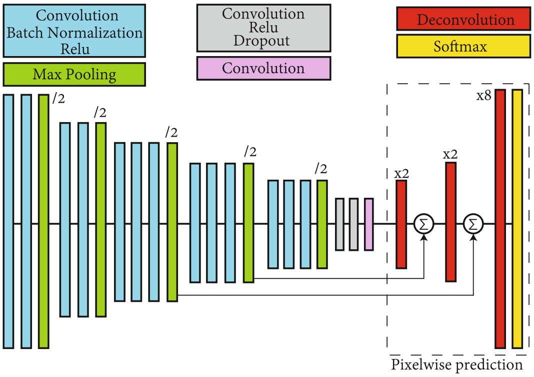
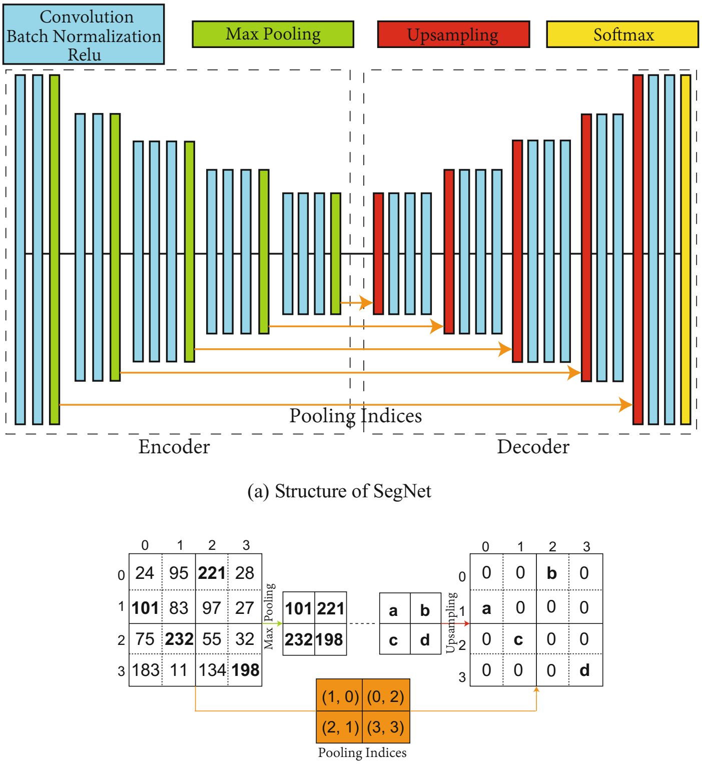


Fig. 10. Structure of Fully Convolutional Network

are missing by pooling processing. In object recognition, these detailed information are unnecessary, but are important information in semantic segmentation tasks. Therefore, the FCN performs processing of integrating feature maps in the intermediate layer before the output layer. FCN have several types, FCN-32s, FCN-16s and FCN-8s, depending on the size of the feature map used for this integration. In FCN-8s, it integrates the feature maps after the third and fourth pooling process to the input of the final layer. At this time, in order to obtain same size feature maps, the feature maps of fourth pooling is upsampled by a factor of 2, and the feature map before the last layer is upsampled by a factor of 4. The segmentation result is obtained from these feature maps.

The FCN needs to store the feature map of the intermediate layer, and the memory usage is large. SegNet [21,22] employ an encoder - decoder architecture which does not need to store the feature map of the intermediate layer. The encoder of SegNet repeatedly performs convolution processing and pooling processing as shown in Fig. 11(a). On the decoder, the feature maps are upsampled by deconvolution processing and the segmentation result of the original image size is outputted. In these processes, as shown in Fig. 11(b), the selected position by pooling is stored, and it refers when upsampling the feature map on the decoder. As a result, it is possible to restore detailed information without using the feature map of the intermediate layer.



(a) Structure of SegNet

	0	1	2	3
0	24	95	**221**	28
1	**101**	83	97	27
2	75	**232**	55	32
3	183	11	134	**198**

101	**221**
232	**198**

a	**b**
c	**d**

	0	1	2	3
0	0	0	**b**	0
1	**a**	0	0	0
2	0	**c**	0	0
3	0	0	0	**d**

(1, 0)	(0, 2)
(2, 1)	(3, 3)

(b) Upsampling using pooling indices

Fig. 11. Structure of SegNet

PSPNet [23] obtains rich information of different scale by the Pyramid Pooling Module which extracts efficient feature at multiple scales as shown in Fig. 12. The Pyramid Pooling Module downsamples the feature map to 1 × 1, 2 × 2, 3 × 3, 6 × 6 and performs convolution process. After that, the feature map is upsampled to the original size and concatenate them. The convolution processing is performed them and a probability map of each class is output.

In addition, Cityscapes Dataset [24] photographed with an in-vehicle camera achieves high accuracy. We also propose scale aware semantic segmentation method especially small objects as shown in Fig. 13. The contributions of the method are (1) to feed the features of small region by multiple skip connections, (2) to extract context from multiple receptive field by multiple dilated convolution blocks. The proposed method has achieved high accuracy in the Cityscapes dataset as shown in Fig. 14. The comparison with state-of-the-art methods, it has achieved the comparative performance at category IoU and iIoU metrics.

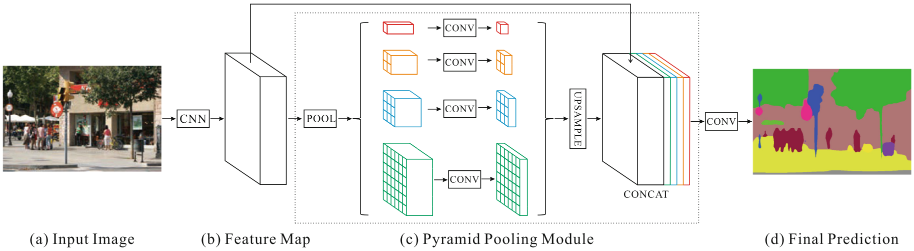


Fig. 12. Structure of PSPNet [23]

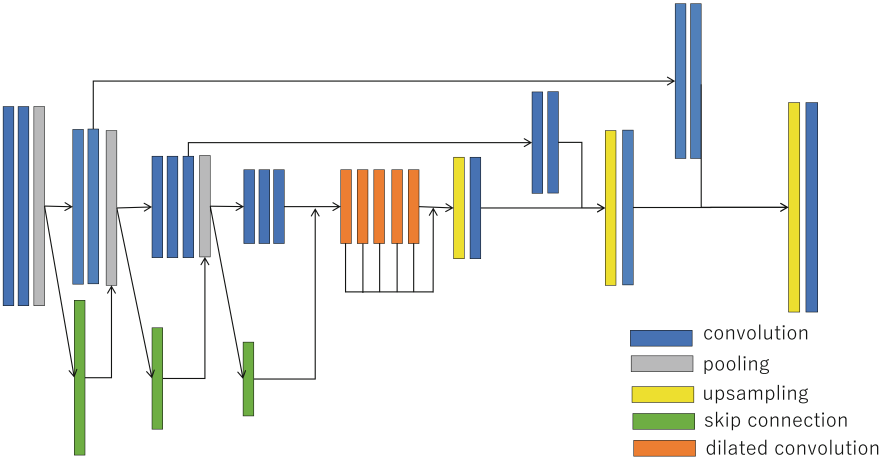


Fig. 13. Structure of our network

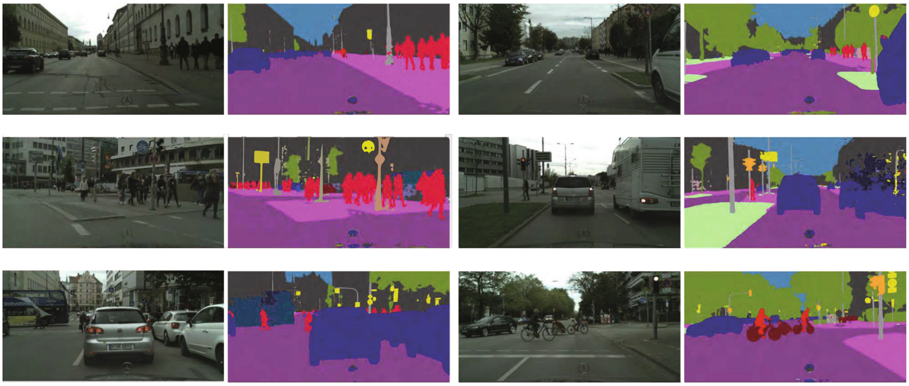


Fig. 14. Result of semantic segmentation

6 Conclusion

In this paper, I explained how deep layer learning is applied in image recognition task. Although the network model and application method are different depending on the task, the problem solvable by deep learning is to find a mapping function from a large amount of data and an accurate teacher label. In the future, learning from a small amount of data, realization of semi-teaching learning with a small amount of teacher-labeled data and a large amount of unlabeled data is a problem in deep learning. Furthermore, we hope to achieve end-to-end learning including reinforcement learning so that deep learning simultaneously acquires the recognition process required to generate better motion and motion of robot.

References

1. Cheng, D., Gong, Y., Zhou, S., Wang, J., Zheng, N.: Person re-identification by multi-channel parts-based CNN with improved triplet loss function. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1335–1344 (2016)
2. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: IEEE Computer Society Computer Vision and Pattern Recognition (2001)
3. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 886–893 (2005)
4. Csurka, G., Dance, C.R., Fan, L., Willamowski, J., Bray, C.: Visual categorization with bags of Keypoints. In: ECCV Workshop on Statistical Learning in Computer Vision (2004)
5. Perronnin, F., Dance, C.: Fisher kernels on visual vocabularies for image categorization. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2007)
6. Jégou, H., Perronnin, F., Douze, M., Sánchez, J., Pérez, P., Schmid, C.: Aggregating local image descriptors into compact codes. IEEE Trans. Pattern Anal. Mach. Intell. **34**(9), 1704–1716 (2012)
7. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**, 91–110 (2004)
8. Yi, K.M., Trulls, E., Lepetit, V., Fua, P.: LIFT: learned invariant feature transform. In: European Conference on Computer Vision (2016)
9. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998)
10. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems (2012)
11. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition (2014). arXiv preprint arXiv:1409.1556
12. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: IEEE Conference on Computer Vision and Pattern Recognition (2015)

13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (2016)
14. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587 (2014)
15. Uijlings, J.R.R., Van De Sande, K.E.A., Gevers, T., Smeulders, A.W.M.: Selective search for object recognition. Int. J. Comput. Vis. **104**(2), 154–171 (2013)
16. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, pp. 91–99 (2015)
17. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 779–788 (2016)
18. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, C., Fu, C.-Y., Berg, A.C.: SSD: single shot multibox detector. In: European Conference on Computer Vision, pp. 21–37 (2016)
19. Farabet, C., Couprie, C., Najman, L., LeCun, Y.: Learning hierarchical features for scene labeling. IEEE Trans. Pattern Anal. Mach. Intell. **35**(8), 1915–1929 (2013)
20. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition (2015)
21. Badrinarayanan, V., Kendall, A., Cipolla, R.: SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation (2015). arXiv preprint arXiv:1511.00561
22. Kendall, A., Badrinarayanan, V., Cipolla, R.: Bayesian SegNet: Model Uncertainty in Deep Convolutional Encoder-Decoder Architectures for Scene Understanding (2015). arXiv preprint arXiv:1511.02680
23. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid Scene Parsing Network (2016). arXiv preprint arXiv:1612.01105
24. The Cityscapes Dataset. https://www.cityscapes-dataset.com
25. Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D., Torr, P.H.: Conditional random fields as recurrent neural networks. In: IEEE International Conference on Computer Vision, pp. 1529–1537 (2015)
26. Dai, J., He, K., Sun, J.: Instance-aware semantic segmentation via multi-task network cascades. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3150–3158 (2016)