

Image Classification for Some Data Sets Composed of Depth Prediction and Edge Extraction Images from Overhead by Convolutional Neural Network

Shu Sumimoto , Yuichi Miyata , Yoko Uwate and Yoshifumi Nishio

Tokushima University, Tokushima 770-8501, Japan

E-mail: sumimoto@ee.tokushima-u.ac.jp

1. Introduction

Image recognition by deep learning is used in various fields. Autonomous vehicles use image recognition by deep learning when they avoid dangers. Therefore, it is important to use image recognition by deep learning on various machines. In recent years, drones have been applied in various fields such as delivery, rescue and security. Therefore, the drones need to fly safely. Then, we thought the image recognition by deep learning would be important for drones to fly safely because it is necessary to grasp the dangers of landing early. We used YOLOv3 [1]~[3] to recognize objects in images from the drones' view. This is a general object detection algorithm, and is suitable for avoiding dangers because it can recognize objects quickly. However, YOLOv3 in Fig. 1 cannot recognize people in images taken from overhead like drones' view with a camera far from people. It is difficult for YOLOv3 to recognize a person whose body part is hidden in order to learn the shape of a person.

In this study, we investigate the depth prediction of objects such as people and cars in images taken from overhead using fully convolutional residual network (FCRN) [4]. This system can predict the depth of an image taken with a monocular camera so it is less expensive than any other system. The convolutional neural network (CNN) can acquire more data than the 2D image because the depth prediction images generated from the FCRN contains 3D data. Edge extraction images are generated using OpenCV in order to extract edge of a person or a car. However, the depth prediction images and the edge extraction images have less information about RGB than the RGB image. In order to solve this problem, blend images in which RGB images, depth prediction images, and edge extraction images were blended at a ratio of 1 : 1 are added to the datasets. It aims to classify datasets composed of RGB images taken with a monocular camera, depth prediction images, edge extraction images, and blended images, and to classify humans or other objects by using CNN.



Fig.1 Object detection by YOLOv3

2. Convolutional Neural Network

In recent years, neural networks have been applied in various fields. In particular, the convolutional neural network (CNN) has produced many results in the image field. In image recognition, results that exceed human ability are also seen. In this study, learning data is created by labeling images of the same type, and images are classified from the features extraction by CNN. In this way, we try to make learning datasets specialized for images taken from overhead. CNN has convolution layers, pooling layers, and fully connected layers. In a convolution layer, matrix calculation is performed using some filters, and feature quantities of the same number as the number of filters are extraction. Next, in a pooling layer, the feature map obtained in the convolution layer is reduced while leaving important information. In this study, the maximum value is obtained by max pooling, and the amount of calculation can be reduced. Finally, in a fully connected layer, a label is connected to each feature amount, and a label with high probability can be selected.

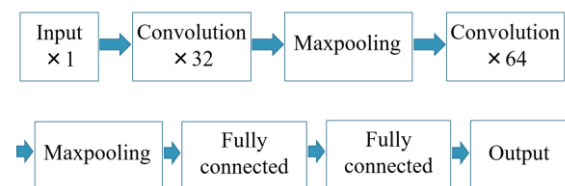


Fig.2 Structure of CNN.

Such network having some layers is called deep learning, and many layers enables accurate image classification. In this study, we use the CNN which has two convolution layers, two pooling layers, and two fully connected layers in Fig. 2.

3. Residual Network

It is possible to improve accuracy by deepening the network layer in image classification. However, it is known that in deeper networks, the accuracy decreases. A residual network uses a residual function to improve this problem. The residual function express a residual between from an output to an input by using a neural network and map it. The residual network can obtain an output close to the input due to this. In this way, gradient disappearance and gradient divergence are prevented by mapping the residuals.

The depth prediction image is generated by FCRN using this residual network.

4. Proposed method

We propose to classify human and car in images from overhead using CNN. First, a depth prediction image is created from an RGB image using FCRN. At this time, existing learning data learned in advance is used. The darker the color, the closer the distance, and CNN can obtain the feature value for the depth. Also, because the outline is easy to understand, it is possible to improve recognition by shape. Furthermore, an edge extraction image is created by using canny edge detection of OpenCV. Edge extraction makes the outline clearer than the depth prediction image. Then, a blend image is created by blending the RGB image and the depth prediction image, the RGB image and the edge extraction image at a ratio of 1 : 1 (Figs. 3) or other ratios.

Six datasets are composed of these images. The first dataset is composed of 200 RGB images. The second dataset is composed of 100 RGB images and 100 depth prediction images. The third dataset is composed of 100 RGB images and 100 edge extraction images. The fourth dataset is composed of 100 RGB images and 100 images blended with RGB images and depth prediction images. The fifth dataset is composed of 100 RGB images and 100 images obtained by blending the RGB image and the edge extraction image. The sixth dataset is composed of 100 images blended with RGB images and depth prediction images, and 100 images blended with RGB images and edge extraction images (Table 1).

Next, we classify human and car in images by using a CNN with two convolution layers, two pooling layers, and two fully connected layers.

Finally, the Train accuracy and test accuracy of each dataset are compared. When CNN learns, training images and test images are compressed to 32×32 pixels. The learning rate of this CNN is 0.000009. Each learning is from 600 to 700 steps, and the training is conducted until the training accuracy converges to about 1.00.

Table.1 Data sets

	RGB images	Depth prediction	Edge extraction	Blend (RGB:Depth)	Blend (RGB:Edge)
1	200	0	0	0	0
2	100	100	0	0	0
3	100	0	100	0	0
4	100	0	0	100	0
5	100	0	0	0	100
6	0	0	0	100	100

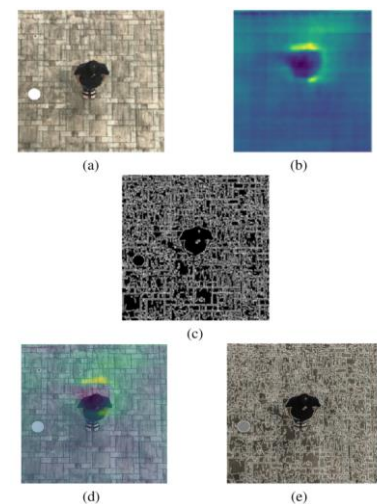


Fig.3 Human images from overhead.
 (a) RGB image. (b) Depth prediction image.
 (c) Edge extraction image.
 (d) Blended image with RGB image and Depth prediction image.
 (e) Blended image with RGB image and Edge extraction image

5. Simulation results

In Table 2, data set 2 is higher than that in data set 1, it is considered that CNN learning the depth prediction image is effective in classifying images taken from overhead.

Further, when the CNN learns the blended image, the test accuracy of data set 2, 5 and 6 is higher than that of data set 1. From these results, it can be said that the combination with RGB information is important.

Table 3 shows the test accuracies when the CNN learns 100 RGB images and 100 blended images with RGB and depth prediction images with other ratios from 1 : 9 to 9 : 1. From this result, when the blend ratio of RGB images and depth prediction images is 6 : 4, the test accuracy is the highest. Similarly, the blend ratio of

the RGB image and the edge extraction image is changed. When the blend ratio of RGB images and Edge extraction images is 6 : 4, the test accuracy is the highest. It can be said that a little more RGB information is easier to recognize objects.

Table.2 Train and test accuracies of data sets 1~6

Data sets	Train accuracy	Test accuracy
1	1.00	0.50
2	1.00	0.80
3	1.00	0.67
4	1.00	0.70
5	1.00	0.67
6	1.00	0.73

Table.3 Train and test accuracies of classification for blended

	Test accuracies	
	Blend RGB : Depth	Blend RGB : Edge
1 : 9	1.00	0.80
2 : 8	1.00	0.73
3 : 7	1.00	0.70
4 : 6	1.00	0.77
5 : 5	1.00	0.70
6 : 4	1.00	0.83
7 : 3	1.00	0.73
8 : 2	1.00	0.63
9 : 1	1.00	0.70

6. Conclusions

In this study, depth prediction images, edge extraction images, and blended images were used as learning data in addition to RGB image, and we tried making to improve the accuracy of image recognition by CNN. From these simulation results, it is considered effective to classify images taken by a camera far from the object and predict the depth of the image. The test accuracy of CNN learned with RGB and depth prediction images is higher than with only RGB images. However, the test accuracy is still low, and it is necessary to improve the test accuracy of image classification.

In the future, we will examine new combinations with RGB images and combinations with different image processing. It will also improve the CNN network itself

References

- [1] Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," arXiv preprint arXiv:1506.02640, 2015.
- [2] Joseph Redmon, Ali Farhadi, "YOLO9000: Better, Faster, Stronger," arXiv preprint arXiv:1612.08242, 2016.
- [3] Joseph Redmon, Ali Farhadi, "YOLOv3: An Incremental Improvement," arXiv preprint arXiv:1804.02767, 2018.
- [4] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, Nassir Navab, "Deeper Depth Prediction with Fully Convolutional Residual Networks," 2016 Fourth International Conference on 3D Vision (3DV), pp. 239248, 2016.