2019 RISP International Workshop on Nonlinear Circuits, Communications and Signal Processing (NCSP2019)
Honolulu, Hawaii, USA, March 4-7, 2019

Design of Convolutional Neural Network for Classifying Depth Prediction Images from Overhead with Some Training Data Sets

Shu Sumimoto, Yuichi Miyata, Ryuta Yoshimura, Yoko Uwate and Yoshifumi Nishio

Dept. Electrical and Electronic Engineering,
Tokushima University
2-1 Minami-Josanjima, Tokushima 7708506, Japan
Email:{nara, nariai, uwate, nishio}@ee.tokushima-u.ac.jp

abstract

We classify RGB and depth prediction images by Convolutional Neural Network (CNN). We aim at differentiating person or other objects in images taken from overhead. We predict depth of some objects, such a person, chairs and car and so on, in overhead images with Fully Convolutional Residual Networks (FCRN). This networks can predict the depth of RGB images taken by monocular cameras. We use RGB and depth prediction images and investigate learning and test accuracies. We investigate learning and test accuracies of the image classification with some traing data sets. We make some training data sets. One is only 200 RGB images, another is only 200 depth prediction images, the other is both of RGB and depth prediction images which are 200 images, and we investigate the optimal proportion RGB images and depth prediction images. Furthermore, we investigate the optimal value of FCRN.

1. Introduction

Image recognition by deep learning is used in various fields. An automatic drive vehicle uses image recognition by deep learning when it avoids dangers. In this way, it is important for a machine to use image recognition by deep learning. Drones are also recently used in various fields. Then image recognition by deep learning is becoming important for that drones fly safely. Therefore, we used YOLOv3 [1]-[3] for drones to fly safely. It is a popular object detection algorithm. However, standing people in overhead images from the view of drones are not able to be recognized by using YOLOv3. So, the CNN needs more images from overhead, however it is difficult to get many images from overhead.

In this study, we investigate the prediction of the depth of some objects, such humans and cars in overhead images with Fully Convolutional Residual Networks (FCRN) [4]. This system can predict depth of images taken by a monocular camera. Depth prediction images from FCRN have 3D data, so Convolutional Neural Network (CNN) can obtain more data from them than 2D images. In this way, we set purpose that getting high accuracies with few images. We classify RGB images and depth prediction images taken by a monocular camera and we aim at differentiating human or other objects by using CNN. Also, the camera has two position, one is lower position and another is higher position.

2. Proposed System

We propose to classify depth prediction images with FCRN. First, we prepare four types images from overhead, images of a human and a car taken by a camera which is close to objects, and images of a human and a car taken by a camera which is far from objects in left side of Figs. 1 and 2. Second, we predict depth of overhead RGB images of a human and a car taken by monocular cameras in right side of Figs. 1 and 2. Also, we prepare test images in Fig. 3. Third, we classify the RGB images and the predicted depth images of a human or a car with a CNN trained by each training data. Forth, we classify the RGB images with a CNN trained by 50 RGB images and 50 depth prediction images. This CNN has 2 convolutional layers, 2 pooling layers and 2 fully connected layers in Fig. 4. When CNN learns, images are compressed. Therefore, training and test images are 28×28 and 32×32 pixels. The learning rate of this CNN is 0.00005.

We compare the learning and test accuracies when a camera is close to an object and when a camera is far from an object.

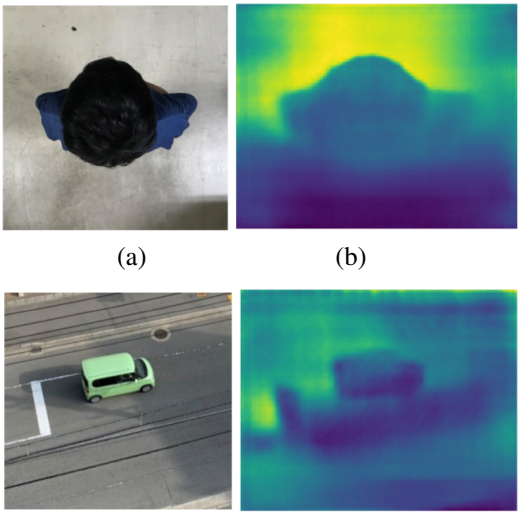


Figure 1: RGB and depth prediction images when a camera is close to objects (training data).
(a) A RGB image of a human.(b) A depth prediction image of a human.
(c) A RGB image of a car.(d) A depth prediction image of a car.

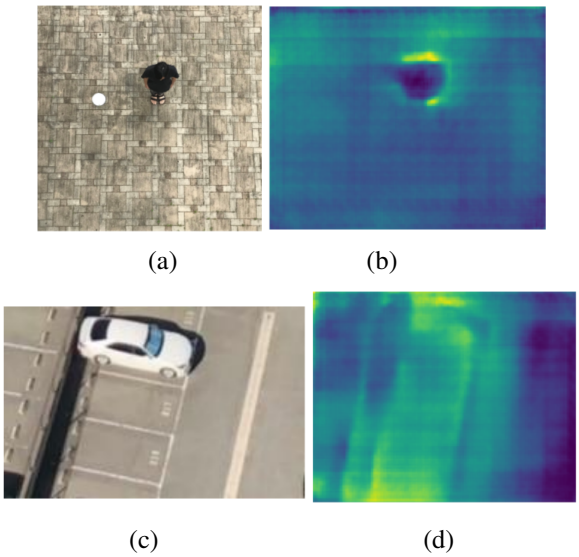


Figure 2: RGB and depth prediction images when a camera is far from objects (training data).
(a) A RGB image of a human. (b) A depth prediction image of a human.
(c) A RGB image of a car. (d) A depth prediction image of a car.

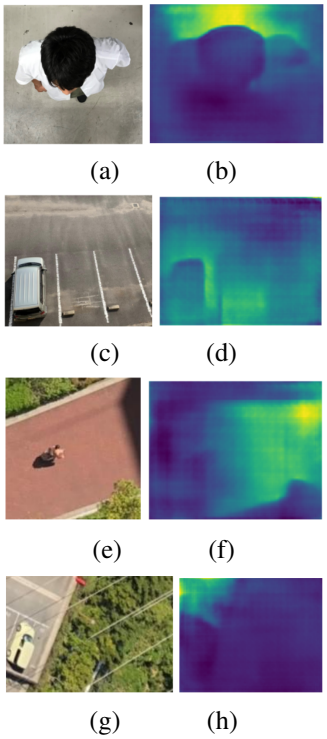


Figure 3: RGB and depth prediction images (test data).
(a)A RGB image of a human when a camera is close to a object.
(b)A depth prediction image of a human when a camera is close to a object.
(c) A RGB image of a car when a camera is close to a object.
(d)A depth prediction image of a car when a camera is close to a object.
(e)A RGB image of a human when a camera is far from a object.
(f)A depth prediction image of a human when a camera is far from a object.
(g)A RGB images of a car when a camera is far from a object.
(h)A depth prediction image of a car when a camera is far from a object.

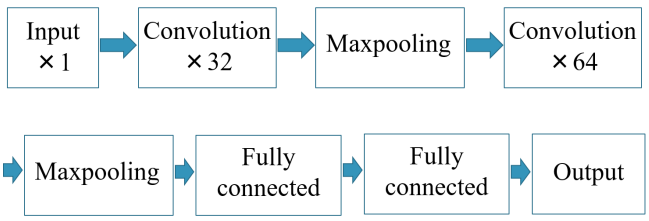


Figure 4: Structure of CNN.

3. Simulation results

Table 1 shows the average of learning and test accuracies when we classify only RGB images 10 times and we classify only depth prediction images 10 times and images are 28 × 28 pixels. We define as the learning 400 steps, the number of the training data sets 200 which are 100 overhead images of a human and 100 overhead images of a car and the number of the test data sets 10 which are 5 overhead images of a human

and 5 overhead images of a car. They are taken by a camera which is close to an object. Table 2 shows the average of learning and test accuracies with same datasets when we define as the learning 800 steps and they are taken by a camera which is far from an object. Table 3 shows the average of learning and test accuracies when we define as the learning 600 steps, the CNN learns same data and they are 32×32 pixels and they are taken by a camera which is far from an object.

Table 4 shows the average of learning and test accuracies when we classify only RGB images 10 times and images are 28×28 pixels. We define as the learning 700 steps, the number of the training data sets are 200 which are 100 overhead images which are 50 RGB images of a human, 50 RGB images of a car, 50 depth prediction images of a human and 50 depth prediction images of a car and the test date are same. They are taken by a camera which is far from an object. Table 5 shows the average of learning and test accuracies as well as Table 4, while data sets are 32×32 pixels.

From Tables 1 and 2, when a camera is close to an object, the test accuracy of classifying RGB images is higher than depth prediction images. On the other hand, when a camera is far from an object, the test accuracy of classifying predicted depth images is as high as RGB images. From Tables 2 and 3, when a camera is far from an object, the test accuracy of classifying 28×28 pixels images is as high as 32×32 pixels images. From Tables 4 and 5, when a camera is far from an object and tarining data are 50 RGB images of a human, 50 RGB images of a car, 50 depth prediction images of a human and 50 depth prediction images of a car, the test accuracy of classifying 28×28 pixels images is as high as 32×32 pixels images.

From Tables 3 and 5, when a camera is far from an object and images are 32×32 pixels, the test accuracy of classifying RGB images with CNN trained by 50 RGB images of a human, 50 RGB images of a car, 50 depth prediction images of a human and 50 depth prediction images of a car is higher than RGB and depth prediction images.

Figure 5 shows accuracies and epochs. From Fig. 5, rising accuracies with the CNN trained 32×32 pixels images are faster than 28×28 pixels at every situation.

Table 6 shows the relationship between the learning rate and the test accuracies. From Table 6, the test accuracy is 0.8 which is the highest value when the learning rate is 0.000009.

Table 7 shows the relationship between the test accuracies and the ratio between RGB images and depth prediction images which are training data. From Table 7, the test acuracies are about 0.8 when the training data ratio between RGB images and depth prediction images have between $40 : 60$ and $50 : 50$. The test accuracies are around 0.5 when the ratios are $100 : 0$ and $0 : 100$. In this way, the test accuracies are higher when the ratio is around $50 : 50$ than when the ratio is not around $50 : 50$.

Table 1: Average of learning and test accuracies when a camera is close to an object and images are 28×28 pixels

	RGB images	Depth Prediction images
Learning accuracy	1.00	0.81
Test accuracy	0.97	0.48

Table 2: Average of learning and test accuracies when a camera is far from an object and images are 28×28 pixels

	RGB images	Depth Prediction images
Learning accuracy	1.00	1.00
Test accuracy	0.44	0.40

Table 3: Average of learning and test accuracies when a camera is far from an object and images are 32×32 pixels

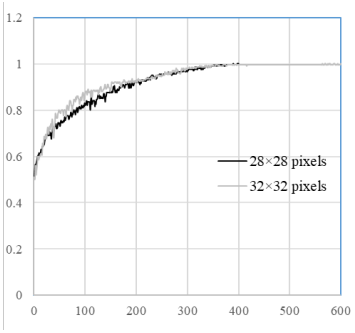
	RGB images	Depth Prediction images
Learning accuracy	1.00	1.00
Test accuracy	0.50	0.48

Table 4: Average of learning and test accuracies when a camera is far from an object and images are 28×28 pixels and training data are 50 RGB images and 50 Depth prediction images

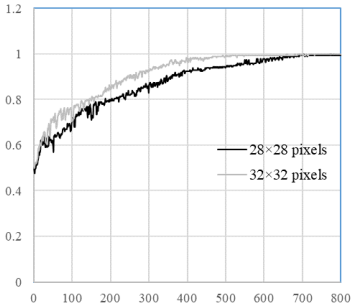
	RGB images
Learning accuracy	1.00
Test accuracy	0.76

Table 5: Average of learning and test accuracies when a camera is far from an object and images are 32×32 pixels and training data are 50 RGB images and 50 Depth prediction images

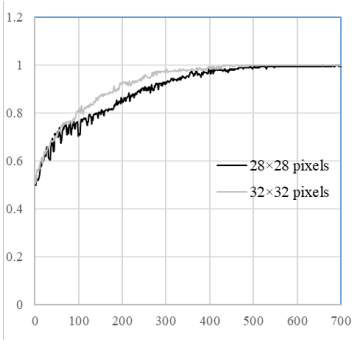
	RGB images
Learning accuracy	1.00
Test accuracy	0.77



(a)



(b)



(c)

Figure 5: Accuracies and epoch
(a) Accuracies with the CNN trained by RGB images when a camera is far from an object.
(b) Accuracies with the CNN trained by depth prediction images when a camera is far from an object.
(c) Accuracies with the CNN trained by 50 RGB images and 50 depth prediction images

Table 6: Relationship between the learning rate and the test accuracies

Learning rate	Learning accuracies	Test accuracies
0.000007	1.00	0.66
0.000008	1.00	0.74
0.000009	1.00	0.80
0.000010	1.00	0.77
0.000011	1.00	0.68

Table 7: Relatioship between training data ratio between RGB images and depth prediction images and the test accuracies

Training data RGB : Depth prediction	Learning accuracies	Test accuracies
100 : 0	1.00	0.50
20 : 80	1.00	0.66
25 : 75	1.00	0.70
40 : 60	1.00	0.80
45 : 55	1.00	0.72
50 : 50	1.00	0.80
55 : 45	1.00	0.70
60 : 40	1.00	0.70
75 : 25	1.00	0.54
80 : 20	1.00	0.52
0 : 100	1.00	0.48

4. Conclusions

From these simulation results, we consider it is effective for classifying images taken by a camera which is far from an object to predict the depth of images. The test accuracies with the CNN trained by RGB and depth prediction images are higher than by only RGB images and only depth prediction images. Also, we found the optimal learning rate and the optimal training data ratio between RGB images and depth prediction images.

However, the test accuracies are still low. Then we will try to raise test accuracies of classifying images.

We will try three methods. First, we will change pixels of images. 32×32 pixels is too small, so we will increase pixels. Second, we will let CNN learn RGB and depth prediction images which are same situation at the same time. Also, we will make edge extraction images and let CNN learn RGB and depth prediction and edge extraction images.

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: 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), 2016, pp. 239248.