IEEE Workshop on Nonlinear Circuit Networks
December 4-5, 2020
NCN 2020

Classification for Datasets Composed of Depth Prediction, Edge Extraction Images and 3D data from Overhead by Neural Network

Shu Sumimoto, Yoko Uwate and Yoshifumi Nishio
Dept. of Electrical and Electronic Engineering, Tokushima University
2-1 Minami-Josanjima, Tokushima 770–8506, Japan
Email:{sumimoto, uwate, nishio}@ee.tokushima-u.ac.jp

***Abstract*—In this study, we classify images and 3D data by using Convolutional Neural Network (CNN) and Spherical CNN. We aim at differentiating humans or cars taken from overhead with a camera far from an object. Datasets are composed of depth prediction images and edge extraction images. We investigate the prediction of the depth of some objects, such a human and cars, in overhead images with Fully Convolutional Residual Networks (FCRN). The other datasets have 3D data made from 2D images by using Pixel2Mesh. We investigate each train and test accuracy of image and 3D data classification with the datasets.**

I. 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 recognized objects quickly. However, YOLOv3 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.

In addition, we make 3D models, such a human and cars, in overhead images with Pixel2Mesh [5]. It is the neural network algorithm which can make 3D models from 2D images to extract features of object apexes and reshape an elliptical sphere to a 3D model. We classify the 3D models of a human and a car by using Spherical Convolutional Neural Networks on Unstructured Grids (UGSCNN) [6]. It has spherical convolutional layers which are icosahedral spherical meshes. It can classify 3D models to project 3D plots to a sphere and extract 3D features. We aim at differentiating human or other objects to make 3D models of them from 2D images.

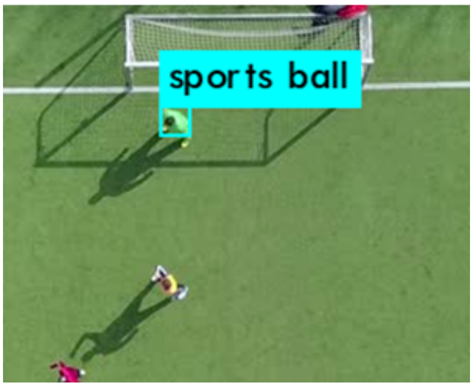

Fig. 1: Object detection by YOLOv3.

II. Basic techniques for image recognition

A. Convolutional Neural Network

In recent years, neural networks have been applied in various fields. In particular, the 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.

Such network having some layers is called deep learning, and many layers enable 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.

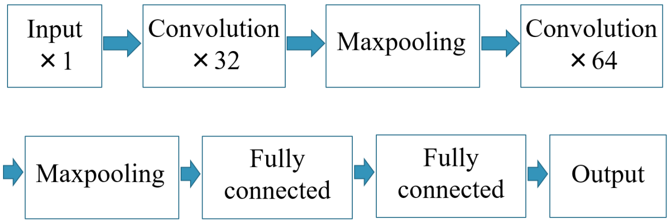

Fig. 2: Structure of CNN.

B. *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 a residual, and the residual network can obtain an output close to the input. 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.

C. *Pixel2Mesh*

Pixel2Mesh is used to make 3D data from 2D images. It is an end-to-end deep learning architecture that produces a 3D shape in triangular mesh from a single color image. This network represents 3D mesh in a graph-based CNN and produces correct geometry by progressively deforming an ellipsoid, leveraging perceptual features extracted from the input image. Extensive experiments show that this method not only qualitatively produces mesh model with better details, but also achieves higher 3D shape estimation accuracy compared to the state-of-the-art.

D. *Spherical Neural Network*

We use Spherical CNNs on Unstructured Grids (UGSCNN) for 3D object classification. It has convolution kernel for CNN on unstructured grids using parameterized differential operators focusing on spherical signals such as 3D point clouds or panorama images. It has 2 merits, one is a novel CNN approach on unstructured grids using parameterized differential operators for spherical signals, and the other is unique kernel parameterization. These allow model to achieve the same or higher accuracy with significantly fewer network parameters.

III. Proposed Method

We propose to classify a human and a 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. In addition, it is possible to improve recognition by shape, because the outline is easy to understand. 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. Thereupon, 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, 4) or other ratios. Six datasets are composed of these images, and all images have a car or a human. 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 I).

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 learning 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.

In addtion, we propose to classify 3D objects of a human and a car taken from overhead. First, we process RGB images from overhead into 3D models by using Pixel2Mesh in Fig. 5. It learned ShapeNet data set for 3D models and our data sets are made by transfer learning. One of our data sets have each 100 3D models of a human and a car were made from images with scenery. Another data set of 3D models made from images without scenery. Second, we classify 3D models of a human and a car with the UGSCNN. It has 7 spherical convolutional layers. They are made with 20 meshes, and 3D plots are projected to them. Convolutional processes work on sphere and features of 3D models are extracted. We compare the test accuracies of 2 data sets of images or 3D models of a human and a car. The first data set has 3D models, which are 100 human models and 100 car models with scenery made from the 2D RGB data set by using Pixel2Mesh. The second data set has 3D models as well as second data set, but they are made from first data set images without scenery.

TABLE I: Datasets.

	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

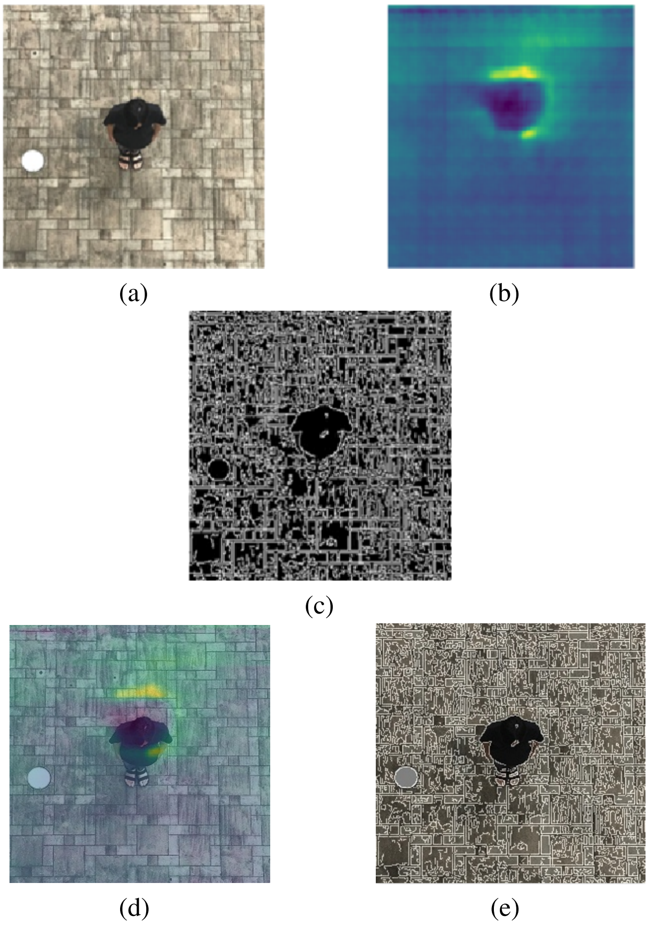

Fig. 3: Human images from overhead.
(a) RGB image. (b) Depth prediction image.
(c) Edge extraction image.
(d) Blended image with RGB image and Depth prediction image.
(e) Blended image with RGB image and Edge extraction image.

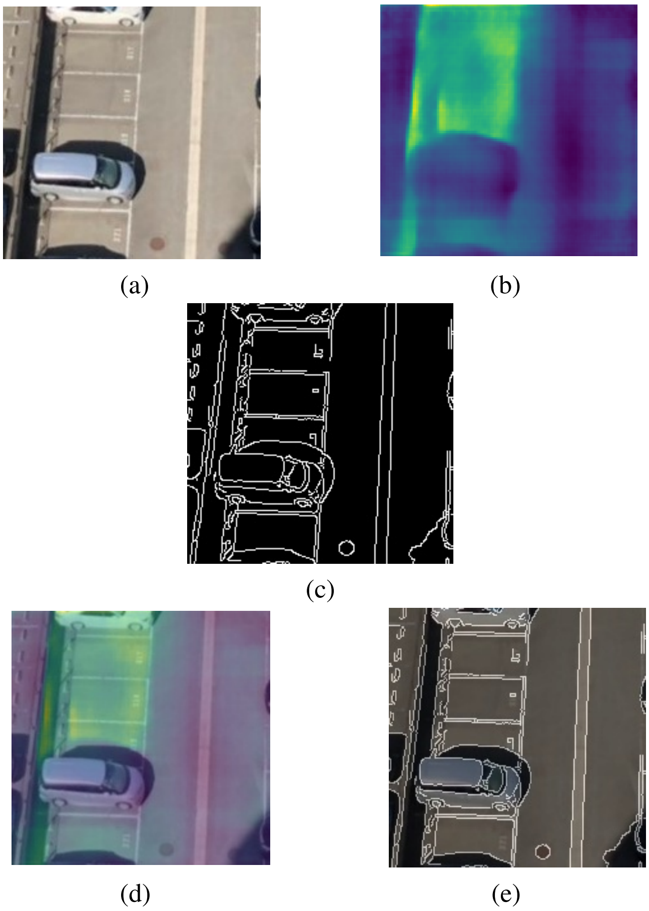

Fig. 4: Car images from overhead.
(a) RGB image. (b) Depth prediction image.
(c) Edge extraction image.
(d) Blended image with RGB image and Depth prediction image.
(e) Blended image with RGB image and Edge extraction image.

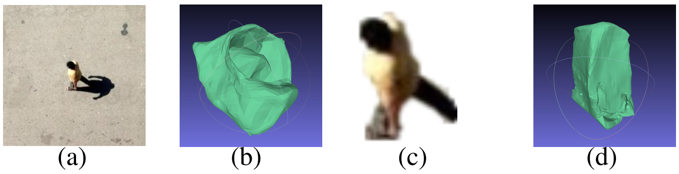

Fig.5: RGB images and 3D models.
(a) A RGB image with scenery.
(b) A 3D model with scenery.
(c) A RGB image without scenery.
(d) A 3D model without scenery.

IV. Simulation Results

The 2D data sets are classified by using CNN, and these results are shown in Tabale Ⅱ. The test accuracy is 0.50, when the CNN learns only RGB images shown in Table Ⅱ (1). In other words, CNN cannot classify people and cars in images taken from far overhead. However, it is considered that CNN learning the depth prediction image is effective in classifying images taken from overhead, because the test accuracy in Table Ⅱ (2) is higher than that in Table Ⅱ (3). Further, when the CNN learns the edge extraction images, the test accuracy is

0.67 in Tables Ⅱ (3). It is higher than Table Ⅱ (1), but lower than Table Ⅱ (2). From these results, it can be said that the depth prediction information is more effective than the edge extraction information. Table Ⅱ (4) shows the test accuracy 0.83 when the CNN learns 100 RGB images and 100 blended images with RGB and depth prediction images with ratio 6 : 4. The highest score is got with this ratio in Table Ⅲ. From Table Ⅳ, when the blend ratio of RGB images and Edge extraction images is 6 : 4, the test accuracy is the highest 0.80 in Table Ⅱ (5). It can be said that a little more RGB information is easier to recognize objects.

In addition, the 3D data sets are classified by using UGSCNN, and these results are shown in Table V. The test accuracy of the classification for the first data set is 0.50. This low score is contrasted by higher classification accuracy values for the second and third dataset, which respectively contain 3D models data with and without scenery. It is considered that 2D images have only RGB information of a head or a shoulder. On the other hands, 3D models have more detailed information than 2D images.

TABLE II: Test accuracies with 2D datasets.

	Test accuracy
(1)	0.50
(2)	0.80
(3)	0.67
(4)	0.83
(5)	0.80
(6)	0.73

TABLE III: Test accuraccies when the CNN learns 100 RGB images and 100 blended images with RGB and depth prediction images with other ratios.

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

TABLE IV: Test accuraccies when the CNN learns 100 RGB images and 100 blended images with RGB and edge extraction images with other ratios.

Blend RGB : Edge	Test accuracy
1 : 9	0.73
2 : 8	0.70
3 : 7	0.70
4 : 6	0.73
5 : 5	0.67
6 : 4	0.80
7 : 3	0.70
8 : 2	0.70
9 : 1	0.70

TABLE V: Test accuracies with 3D datasets.

Data set	Test accuracy
3D models with scenery	0.60
3D models without scenery	0.70

V. 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 to make to improve the accuracy of image recognition by CNN. From these simulation results, it is effective for classifying images taken by a camera far from the object to predict the depth of the images. 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. We consider that low quality depth prediction images cause the low test accuracies.

In addition, we have investigated comparison of the 3D objects classification into a human or a car made from 2D RGB images with the 2 data sets. From these simulation results, we consider it is effective for classifying objects taken from overhead to make 3D models from 2D images. However, these results are still low as well.

In the future, we will let UGSCNN learn another 3D data made by the sensors.

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), pp. 239248, 2016.

[5] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang, “Pixel2Mesh: Generating 3D Mesh Models from Single RGB Images”, ECCV 2018.

[6] Chiyu“ Max ”Jiang, Jingwei Huang, Karthik Kashinath, Prabhat, Philip Marcus, Matthias NieBner, “Spherical CNNs on Unstructured Grids”, ICLR 2019.