International Journal of Engineering & Technology, 7 (3.34) (2018) 221-225

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

How to Generate Image Dataset based on 3D Model and Deep Learning Method

Sooyoung Cho[1], Sang-Geun Choi[2], Daeyeol Kim[3], Kyounghak Lee[4], Chae-BongSohn[5*]

[1,2,3,5]Dept. of Electronics and Communications Engineering, Kwangwoon University, Seoul, Korea
[4]IACF, Kwangwoon University, Seoul, Korea
[*]Corresponding author E-mail: cbsohn@kw.ac.kr

Abstract

Performances of computer vision tasks have been drastically improved after applying deep learning. Such object recognition, object segmentation, object tracking, and others have been approached to the super-human level. Most of the algorithms were trained by using supervised learning. In general, the performance of computer vision is improved by increasing the size of the data. The collected data was labeled and used as a data set of the YOLO algorithm. In this paper, we propose a data set generation method using Unity which is one of the 3D engines. The proposed method makes it easy to obtain the data necessary for learning. We classify 2D polymorphic objects and test them against various data using a deep learning model. In the classification using CNN and VGG-16, 90% accuracy was achieved. And we used Tiny-YOLO of YOLO algorithm for object recognition and we achieved 78% accuracy. Finally, we compared in terms of virtual and real environments it showed a result of 97 to 99 percent for each accuracy.

Keywords: *CNN, VGG, YOLO, Virtual space, Deep learning, Dataset generation*

1. Introduction

It is generally accepted that object recognition is an important technology in Intelligent Robotics which can be applied to achieve 3D spatial information such as position, direction or size and format of the object by using learned data. In recent years, Deep-learning has been significantly improved in object recognition and classification. Among various architectures, CNN is in general use for object classification [1]. At first, objects were built on a layer-by-layer basis. Then, representative features learned the object for classification during the training process and it showed the result. In this experiment, we compared 15 types of objects using CNN and VGG-16. Using two converted layers in succession reduced the number of parameters and improved recognition performance. In addition, 10 object types were also recognized by using YOLO. The former CNN and VGG-16 showed poor performance in real time, although they were designed for classification purposes. However, the YOLO architecture achieved the probability of each region by dividing the image and predicting the bounding box [2]. The weights are applied to the bounding box of each region in accordance with the probability, and it can perform a fast performance in a single network. However, it has a difficult part to fine-tune or adjust in regard to the object recognition and classification. It is because a large amount of data is required for each data set regarding object recognition and classification. Object classification requires at least 1,000 or more data sets for each object in the dataset. Moreover, much more data are needed for the user so as to define for actual working environments, not a laboratory environment. More than 15,000 datasets were utilized using 15 different objects under the set environment in the laboratory. It is very useful for training if there are other objects apart from the main object of each image. Therefore, we generated a large number of data set for the experiment and it was carried out under the useful environment for the test. Furthermore, this study will suggest the way to create the data set which is needed by employing the virtual environment and FFMPEG in a simple and rapid method.

2. Materials and Methods

2.1. *Dataset*

The size of the data used in the classification was 15,000 pieces for each of 15 pieces of objects. In the recognition part, we have recorded 6 seconds and achieved 330 frames at 60 fps for each image with FFMPEG. We created and used 19,800 data sets with tagged information about 10 objects. Using the Unity, the 3D model shape can be expressed in 4 types, and the accuracy is expressed through CNN by constructing 4 datasets.

2.1.1. Dataset Type – Real Object and 3D Object

Dataset types are divided into various shapes, sizes, and colors to clearly distinguish them, and decisions have been made regarding the management, technology, and cost of each object [Table 1].

Table 1: Evaluation index

Evaluation	System performance
Administration	Object management
	Object manipulation
	Diversity of object
Physical	Object image processing
	Shooting
	Object process
Technical	Automatic detection
	Object recognition

	System throughput
Cost	Movement
	Production
	Installation

The objects are determined by using the condition of Table 1, and the objects to be used in the real environment are made into a database separately. It is divided into three types: Cylindrical, Rectangular pole, and polygon. Each type was subdivided into different types [Figure 1].

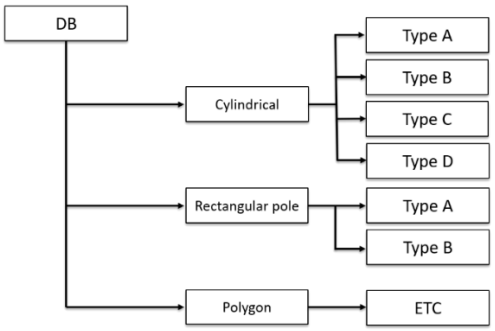

Figure 1: Object Database

Data sets used for learning were divided into classification and recognition. The classification was learned by CNN and VGG, and YOLO was used for recognition [Figure 2].

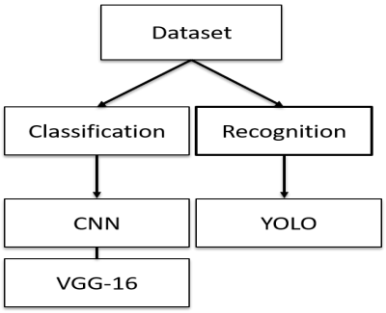

Figure 2: Object Database

Finally, we used 3D modeling to create four kinds of objects to be tested in the virtual environment and real environment. After creating four kinds of objects in a virtual environment, we created objects to be tested in a real environment using a 3D printer. The results were compared and analyzed according to the environment and conditions.

2.2. Classification Learning

Classification is a kind of supervised learning. The category of newly observed data is learned by learning the relationship between existing data and category. VGA was configured with GTX 1080. With vanilla CNN and VGG-16, 4 to 5 hours and 8 to 9 hours were taken respectively for training.

2.2.1. CNN

Although HOG [3] and SIFT [4] algorithms are widely used for detecting and classifying objects in images, there are limitations in performance. In order to improve the performance, many studies have attempted to solve the problem with other algorithms, but the improvement of the performance is insignificant [5]. In 2012, International Image Recognition Technology Competition named "ImageNet", AlexNet [6] showed up about 10% better performance than the algorithm that showed the best performance before. As a result, researchers related to image recognition are mainly based on CNN. In 1989, CNN was first applied to a study of handwritten postal code recognition [7], but it took too much time to learn and was limited in scope because of problems such as overfitting. However, with the development of hardware, the computation time has been shortened exponentially through a parallel operation using GPU, and improved algorithms such as Dropout [8] and ReLu [9] have been studied. Also, with the advent of big data, large amounts of data to be used for learning and verification can be obtained easily compared with the past, and an environment where CNN can be used has been created. CNN extracts and optimizes features for input images unlike existing techniques, and it has a structure to repeat the classification process several times. [Figure 3] shows the general structure of CNN.

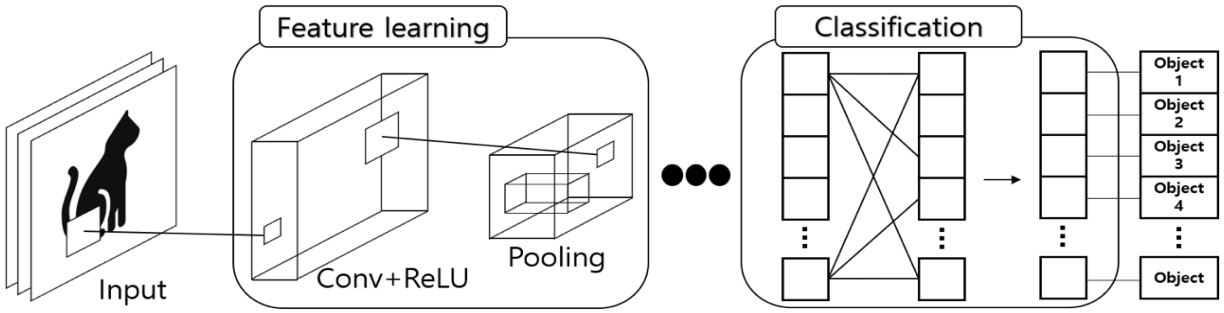

Figure 3: General Architecture of CNN

The convolution layer is responsible for extracting meaningful features from the image. Convolution operations are applied to a certain mask to perform a kind of filtering, and a convolution operation is performed by applying a filter to obtain the desired feature, and a feature map is finally generated. Then, the pooling layer extracts image feature values in the feature map created in the convolution layer. The method of extracting feature values includes Max pooling method of selecting the maximum value and Average pooling method of selecting the average value. The feature amount is reduced by reducing the number of dimensions by selectively extracting necessary features. Finally, the Fully Connected Layer classifies and predicts objects using features extracted as a result of the iteration of the Convolution Layer and the Pooling Layer. Although CNN is applied to image classification, the field of image classification has developed remarkably. However, CNN is not suitable to solve the problem of object detection, as CNN can only know the existence of the object to be searched in the input image but cannot know the position of the object.

2.2.2. VGG

VGG is a limited convolution neural network model in "Very Deep Convolutional Networks for Large-Scale Image Recognition" [10]. VGG showed high accuracy due to image storage such as ImageNet and high-performance computing systems such as GPU or Large Scale distributed clusters, in particular, performing in-depth and visual perception. In the field

of computer vision, there has been much technology development and many attempts to improve lots of architectures as it requires the use of ConvNets.

ConvNets has become more and more used as peer annotation, and many techniques have been developed and attempted to improve a number of architectures as the VGG architecture [Figure 4] is an image with a fixed size of 244 x 244 at the input of ConvNets. Modify the RGB values through preprocessing, modify other parameters, and use very small (3 x 3) convolution filters at all layers[11]. Spatial pooling is performed by five max-pooling layers, followed by some convolution. In addition, it can increase the depth of network by adding more convolution layers. The result is an accurate ConvNet architecture that not only achieves the latest accuracy for ILSVRC classification and localization work, but also for other image-aware data sets that achieve superior image performance.

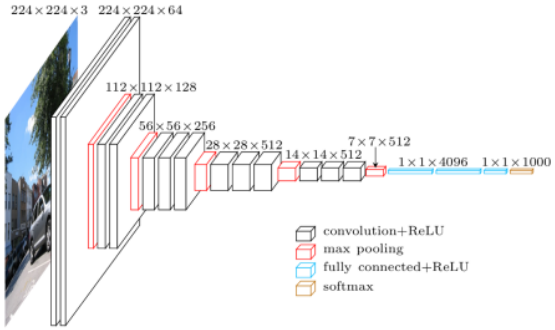

Figure 4: Architecture of VGG-16

2.3. Recognition learning

Recognition using YOLO is a structure in which the network finds the position of the bounding box at the final post process and class classification is interrupted at the same time. It is a structure in which only one network extracts feature at a time and classifies classes by creating boxes. With GTX 1080, it takes 24 hours to train using Tiny-YOLO.

2.3.1. YOLO

YOLO [4] solved the slow detection rate of the existing proposal method by applying the Grid method to the object detection process. Grid method has significantly shortened the bounding box prediction time, and the distribution probability of the class can be calculated at the same time as the detection, as it is possible to detect the object at high speed. In addition, YOLO uses the sliding window approach to search the entire image, and the window is moved at regular intervals in the image to determine the presence of an object in each image region within the window. Unlike the existing method, the YOLO algorithm applies a single network to the entire image. This single network divides the image into defined areas and creates a bounding box that indicates the type and location of the object in each area. At the final stage of the YOLO, bounding box prediction and classification are processed simultaneously. Then draw a bounding box of unequal size for each region that is partitioned into the image. The bounding box contains information about the recognized object and consists of x, y, w, h, confidence. Where x and y are the object center coordinates, and w and h are the width and height of the box. Confidence is a value indicating whether the box contains and object. The larger the confidence, the thicker the bounding box. And the box with a value less than the threshold is removed. At the same time, the class of the recognized object is classified by the bounding box of different color according to the class [Figure 5].

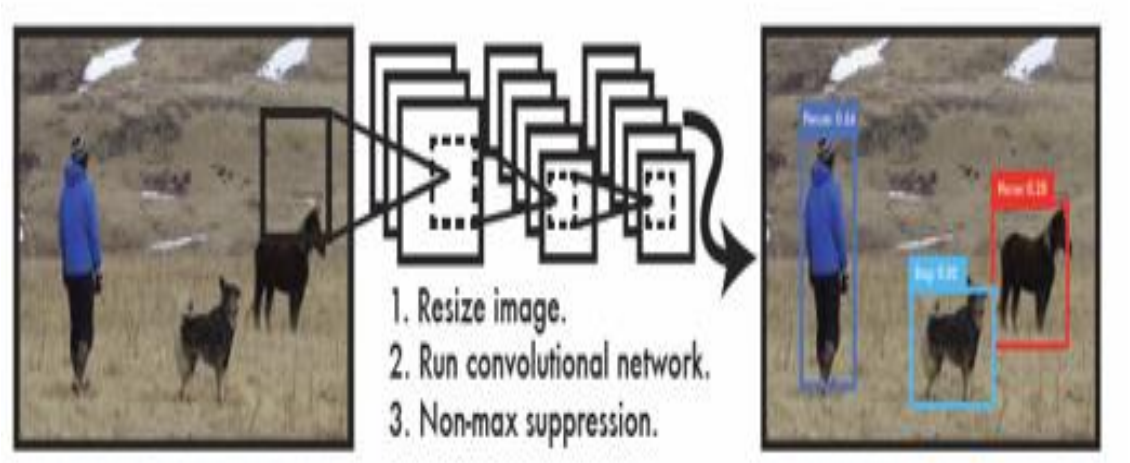

Figure 5: The YOLO Detection System

YOLO has three features different from existing object recognition algorithms. First, YOLO is very fast. Based on the Titan X GPU, real-time image computation is possible at 45 FPS operation speed, and it shows better accuracy than DPM algorithm [12], which is another real-time object recognition algorithm. Feature extraction, region classification, and bounding box prediction work individually in the DPM algorithm using the sliding window method. The mean average precision (mAP) is used as an index for evaluating the object recognition algorithm by calculating the recognition accuracy for all classes of average precision (AP) and then averaging the values. The DPM algorithm shows 30 FPS and the mAP shows 26.1, but Fast YOLO has a fast processing speed of 155 FPS while maintaining an accuracy of 52.7, which is twice the mAP value of the DPM algorithm. Second, YOLO uses the whole image when it processes learning, so it can grasp the flow of the image and it is strong against errors caused by background change. Finally, YOLO learns about the general characteristics of objects, so performance does not fall far behind new variables [13, 14].

2.4. Test Environment

The test was conducted in the laboratory environment. As a necessary condition for the environment, the background color was made monochromatic and the experiment was conducted in order to reduce the probability of misrecognition of the background as an object. In the real environment, we set the object as shown in [Figure 6] and proceeded to capture the data through recording. On the contrary, in the virtual environment, the object was created and recorded using the program as described in

[Figure 7]. The recorded video was divided by FFMPEG for the learning process.

Figure 6: Practical test environment

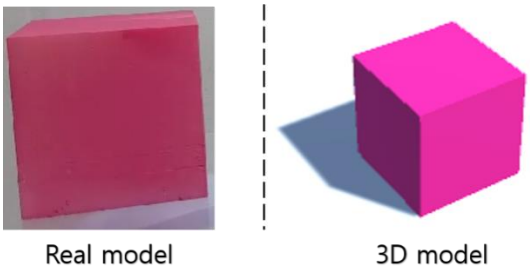

Figure 7: Real and 3D object

3. Results and Discussion

In the classification, the accuracy is expressed through a data set generated by another environment. The results are presented in two directions. We tested whether the network such as CNN, VGG, and YOLO can be correctly identified by using data sets created in a test environment. We confirmed the accuracy of 90% or more in 10 types of objects recognition and 78% of the Mean IU in recognition using Tiny-YOLO [Table 2].

Table 2: Evaluation result

No.	Evaluation items	measures
1	Object recognition rate	100 %
2	Object discrimination accuracy	90 %
3	Matching accuracy	97 %
4	Mean IU	78 %

Mean IU can be obtained from the equation (1).

$$\text{Mean IU} = \frac{tp}{(tp+fn+fp)} \quad (1)$$

In equation (1), tp is truly positive, fn is false negative and fp is a false positive. It is confirmed that it is recognized even in other environments [Figure 8].

Figure 8: Experiment using image media

Secondly, we obtained the accuracy through CNN and VGG with the data set for the virtual object created in [Figure 7] and the object made in the real world. The test was conducted to make the virtual environment and the real environment similar. The real environment model was created by using a 3D printer. When we created a dataset using virtual and real objects and learned it, we could confirm 98% accuracy on the average [Table 3].

Table 3: Model environment test

No.	Model	Model Measures
1	Cube	99%
2	Cylinder	98%
3	Circle	99%
4	polygon	97%

4. Conclusion

We experimented with various data sets in the proposed way. We could figure out that there is no significant difference observed when the conditions in the laboratory and the virtual were compared. When we checked the above experimental results, we could confirm the high accuracy in the recognition field and 78% in the Mean IU. These results can be seen in the limited environment of laboratories and lack of data sets. The lack of data sets is expected to be improved by developing data sets in a virtual environment. Among the various reasons, the values measured in [Table 3] were high in the accuracy of 3D models and real objects. These results show that if we have information on various objects, we can use it to solve the problem of creating the data set of the desired network, and it is useful in many other environments as well. But in this paper, the experimental environment may have different results because the experiment was conducted except for external factors such as the complexity of the background and the light reflection of an object by an external light source. The difference in results due to external factors is a problem that can be solved by creating an environment similar to the actual environment when creating virtual data. Therefore, in computer vision, the more various data sets, the higher the result. By using this, it is possible to obtain high accuracy in classification. In addition, if the information is entered through the labeling of the created dataset, good results can be expected in the field of recognition.

Acknowledgment

This material is based upon work supported by the Ministry of Trade, Industry & Energy(MOTIE, Korea)under Industrial

Technology Innovation Program. No.10077659, 'Development of artificial intelligence based mobile manipulator for automation of logistics in manufacturing line and logistics center'

References

[1] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105).

[2] Redmon, J., Divvala, S., Girshick, R., &Farhadi, A. (2016). You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 779-788).

[3] Lu, W. L., & Little, J. J. (2006, June). Simultaneous tracking and action recognition using the pca-hog descriptor. In Computer and Robot Vision, 2006. The 3rd Canadian Conference on (pp. 6-6). IEEE.

[4] Zhou, H., Yuan, Y., & Shi, C. (2009). Object tracking using SIFT features and mean shift. Computer vision and image understanding, 113(3), 345-352.

[5] Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., ... & Darrell, T. (2014, November). Caffe: Convolutional architecture for fast feature embedding. In Proceedings of the 22nd ACM international conference on Multimedia (pp. 675-678). ACM.

[6] Nguyen, A., Yosinski, J., & Clune, J. (2015). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 427-436).

[7] LeCun, Y., Jackel, L. D., Bottou, L., Cortes, C., Denker, J. S., Drucker, H., ... &Vapnik, V. (1995). Learning algorithms for classification: A comparison on handwritten digit recognition. Neural networks: the statistical mechanics perspective, 261, 276.

[8] Lin, M., Chen, Q., & Yan, S. (2013). Network in network. arXiv preprint arXiv:1312.4400.

[9] Xu, Z., Yang, Y., & Hauptmann, A. G. (2015, June). A discriminative CNN video representation for event detection. In Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on (pp. 1798-1807). IEEE.

[10] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

[11] Blog.heuritech.com. (2018). A brief report of the Heuritech Deep Learning Meetup #5. [online] Available at: https://blog.heuritech.com/2016/02/29/a-brief-report-of-the-heuritech-deep-learning-meetup-5/.

[12] Haykin, S. S. (Ed.). (2001). Kalman filtering and neural networks(pp. 221-269). New York: Wiley.

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

[14] Dalal, N., &Triggs, B. (2005, June). Histograms of oriented gradients for human detection. In Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on (Vol. 1, pp. 886-893). IEEE.