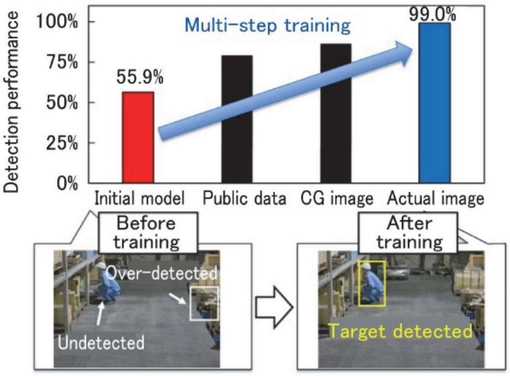
Object Detection Developed with Small Number of Acquired Images by Multi-step Training Method in Image Recognition Using Deep Neural Network



AMANE KOBAYASHI*1 TOMOHIRO MATSUMOTO*1

KIICHI SUGIMOTO*2 KENJI IWATA*3

An image recognition function is indispensable to detect objects in an image for human detection systems that support the safety of industrial vehicles. Image recognition using deep learning requires training with a large number of images of the detection target in the actual operational environment, and the process of acquiring images required a great deal of work. Therefore, in collaboration with the National Institute of Advanced Industrial Science and Technology (AIST), Mitsubishi Heavy Industries, Ltd. has developed a multi-step training method that augments the training data using public data and computer graphics which are close to the conditions of the actual operational environment and allows efficient training with these data in steps. It has been confirmed that this method can achieve the same detection performance with only 280 actual images as that achieved by the conventional method with approximately 5,000 actual images. We plan to widely apply the developed technology to our products that have image recognition functions using deep learning.

1. Introduction

Object detection, which estimates the position and type of a specific object such as a person or vehicle in an image, is often realized by image recognition technology using deep learning(1). In recent years, object detectors trained with public datasets can be easily obtained, but incorporating such a detector into our products as it is will not satisfy the detection performance required by our customers in their actual operational environments. This is because background objects, camera equipment, and other image capturing conditions differ between public datasets and actual operational environments. To eliminate this difference, it is necessary to perform re-training with a considerable number of actual images acquired in each operational environment. However, it is difficult to achieve high image recognition performance when image acquisition opportunities are limited, such as when the work of image acquisition requires much effort and when the customer's environment is not easily accessible or when the detection target is a special object that is not commonly seen. Therefore, we developed a multi-step training method that achieves high detection performance by augmenting the training data using public data and computer graphics that are close to the conditions of the actual operating environment and performing training with these data in steps.

2. Multi-step training

The multi-step training method developed in this study is based on “transfer learning” which has been used in the training of neural networks(2). In this study, we extended the transfer learning so as to apply it in steps to develop a method to perform training sequentially with Step 0: public dataset, Step 1: optimized public dataset, Step 2: background images superimposed with computer graphics, and Step 3: a small number of actual images of the detection target. The aim of this

*1 CIS Department, Digital Innovation Headquarters, Mitsubishi Heavy Industries, Ltd.

*2 Senior Researcher, CIS Department, Digital Innovation Headquarters, Mitsubishi Heavy Industries, Ltd.

*3 Senior Researcher, Information Technology and Human Factors, National Institute of Advanced Industrial Science and Technology

process is to reduce the number of required actual images of the detection target to a small number by sequentially tuning a general object detector to a specialized one for the actual operational environment through training up to Step 3. Hereafter, a model that has been trained with public datasets is regarded as Step 0, and Step 1 and the subsequent steps are described in detail.

Step 1: Training with optimized public datasets

Since the public datasets that are the training data for Step 0 generally cover a wide range of image variations, this step optimizes them so as to make them similar to the actual operational environment and performs re-training to build an object detector that is specialized for the actual operational environment. The optimization conditions include various factors depending on the application of object detection, such as the color of the detection target, objects in the background, and the time of the capture. However, it is not easy to determine these conditions for a huge amount of public data. The size of the detection target on the image, described below, can be easily determined from information in the bounding BOX (BBOX) used for annotation(*1). The size of the detection target on images captured in actual operational environments depends on the characteristics of the camera used, such as its mounting position and angle of view, as well as the relative positional relationship between the camera and the detection target. Therefore, images in which the ratio of the number of pixels in the height and width of the detection target to the whole size of the image is the same as that of the detection target captured in the actual operational environment are selected from the public dataset and used as the training data. The reason why the ratio to the whole size of the image is used as the condition is that any size of images is to be resized to a specified size when input to the object detector.

Step 2: Training with background images superimposed with computer graphics

This step creates training images by superimposing a computer graphics image of the detection target on background images of the actual operational environment for making up for a shortage of training images, and uses them in training. The computer graphics model is prepared so that its color, shape, and posture match the operational condition. To render the computer graphics into an image, the projection parameters are set to match the mounting position/angle and field of view of the camera used. A large number of training images with various arrangements of the computer graphics model relative to the projection grid and various orientations of it relative to the camera viewpoint are created.

Step 3: Training with small number of actual images

This step acquires a small number of actual images in which the detection target is present for a specific area in the actual operational environment and performs training with them. In this case, over-training may occur because due to the number of images to be used for the training is small. To avoid this, it is necessary to evaluate the trained object detector with evaluation images prepared in a way that does not impose a burden on the customer, such as by installing a fixed-point camera. The actual images are acquired from areas with conditions that are likely to cause undetected or over-detected cases, as identified by the background analysis method described in the next chapter.

(*1) Annotation : A task that imparts information necessary for training to image data by enclosing in a rectangle the location where the object to be detected appears on the image.

3. Method to build object detector using multi-step training

The multi-step training is a methodology to adapt the network parameters that the object detector trained previously to the detection target, which alone cannot reduce the number of actual images of the detection target required to achieve high performance and the cost of image collection. As such, we developed a method to build an object detector to effectively utilize multi-step training. **Figure 1** shows the workflow of this method. In this method, before multi-step training, the background of the actual operational environment is captured and analyzed, and a small number of actual images in which the detection target is present are captured. After multi-step training, the resulting object detector is evaluated, the achievement of the target performance is confirmed, and the object detector is incorporated into the product.

In the capturing of the background, background images are acquired over the entire operational area. Although the entire operational area has to be captured, the cost of image collection is much lower than the work of placing the detection target and capturing images under various conditions of its posture, orientation, and relative position to the camera. The background images obtained here are to be superimposed with computer graphics in Step 2. Then, areas that are likely to cause undetected or over-detected cases are identified by applying the two background analysis methods described below to the background images. By acquiring actual images of the detection target in such areas that are likely to cause undetected or over-detected cases, even when their number is small, conditions where the detection of the target is difficult, i.e., data that are essentially necessary for the training, can be applied in the final step of the multi-step training. Here, an undetected case refers to an event in which the position and type information of the detection target are not detected even though the target is present in the image, and an over-detected case refers to an event in which an object that is not the detection target is erroneously detected as such.

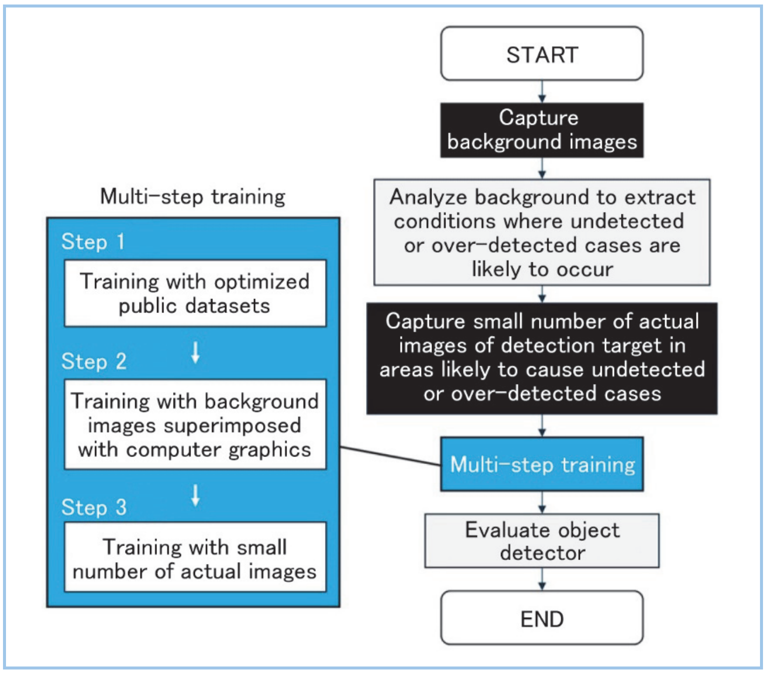

START
Capture background images
Analyze background to extract conditions where undetected or over−detected cases are likely to occur
Capture small number of actual images of detection target in areas likely to cause undetected or over−detected cases
Multi−step training
Evaluate object detector
END
Multi−step training
Step 1
Training with optimized public datasets
Step 2
Training with background images superimposed with computer graphics
Step 3
Training with small number of actual images

Figure 1 Flow of building object detector

Undetected cases are caused when the “confidence score” calculated by the object detector, which indicates the probability that the object in the image is the detection target, is low. One of the reasons why the confidence score decreases is that the feature extraction of the detection target is weakened by the texture of the background area near it. Therefore, it is possible to identify areas where undetected cases are likely to occur by determining which textured background areas are likely to cause undetected cases. In this study, to identify areas where undetected cases are likely to occur, we visualized the difficulty in detecting the detection target in background images as heat maps as shown in **Figure 2**(a). This visualization with detection difficulty maps was made by using a feature extractor based on a convolutional network. We superimposed images of the detection target sequentially for all of its arrangements on images from the background image dataset Places205[(3)], which covers a wide range of texture variations, and performed detection for each to train the feature extractor with the confidence score as the difficulty in detecting the detection target for that superimposing position. By applying the feature extractor established in this manner to background images of the actual operational environment and analyzing the detection difficulty map, it is possible to identify areas in which objects with high difficulty are present as areas where undetected cases are likely to occur.

On the other hand, an over-detected case is an event caused by a background area with a specific texture where the confidence score is high. In this study, we conducted object detection for the background image dataset Places205 in which no detection target was present to prepare an object detector that was trained with textured areas with high confidence scores as over-detected objects. As shown in Figure 2(b), it is possible to identify areas where over-detected cases are likely to occur by applying this object detector to background images of the actual operational environment and detecting texture regions with a high possibility of causing over-detected cases.

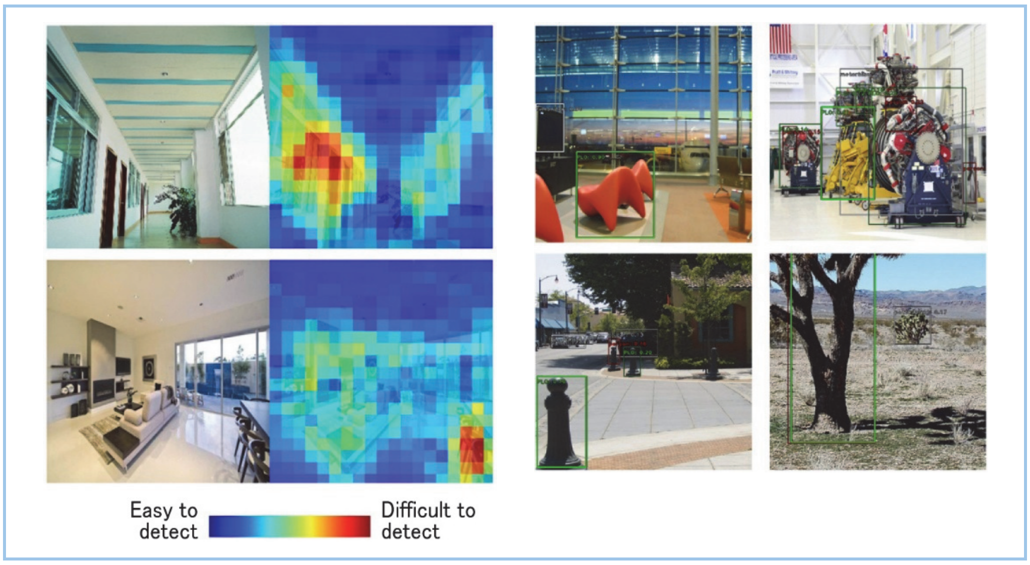


Figure 2 Example of extracting conditions where undetected or over-detected cases are likely to occur by using background analysis
(a) Background image (left) and detection difficulty map (right)
(b) Over-detected objects in background images (enclosed in green frames)

4. Evaluation assuming actual operation

We verified the developed method with a person detection system for safety support of industrial vehicles as an example, according to the workflow shown in Figure 1. The test area was a warehouse.

4.1 Capturing of background images and images in areas likely to cause undetected or over-detected cases

First, we captured the background images using the same camera height and viewing angle as in the operational conditions while moving throughout the test area. As a result of the background analysis, areas that were likely to cause undetected or over-detected cases were identified at six camera arrangements in the test area covered in this verification. **Table 1** shows a summary of the acquired training data. Considering changes in natural light, three scenes (morning, afternoon, and evening) of background images were acquired. In the areas that were likely to cause undetected or over-detected cases, the posture of the detection target was set to a walking posture only so that the image capturing could be completed in a short period of time.

Table 1 Training data and evaluation data of actual image

	Training data		Evaluation data	
Acquired data	Background image	Detection target image in areas likely to cause undetected or over-detected cases	Detection target image in areas likely to cause undetected or over-detected cases	Detection target image in areas other than the areas likely to cause undetected or over-detected cases
Capturing method	Continuous capturing while moving throughout the entire test area	Capturing images of walking persons	Capturing images of persons in upright, mid-rise, and crouching postures and looking forward, backward, left, and right	Capturing images of walking persons
Data volume	Morning: 4,437 images Afternoon: 5,448 images Evening: 5,131 images	Image: 280 Detection target: 280 persons	Image: 1,245 Detection target: 1,245 persons	Image: 193 Detection target: 193 persons

The entire training data acquisition work, including background analysis, was completed in two days. For the background, several thousand images were acquired in a single capture, but the image collection cost was low since the collection time is only the moving time. In addition, for the detection target, the number of collected images was as few as 280 since the capturing was not performed over the entire test area, while in the past it had been several thousand as is the case with the background image.

4.2 Application of multi-step training

Table 2 shows the training data used in the multi-step training. The initial model adopted an object detector that was pre-trained with MS-COCO[(4)]. First, training with public datasets selected so that only detection targets of the same size as in the operational conditions were included was performed. The size of the detection target in the previously acquired images of the actual operational environment ranged from 30 to 290 pixels wide by from 100 to 690 pixels high, while the overall image was 1,150 pixels wide by 870 pixels high. Then, training with 200,000 images, roughly the same as the case for MS-COCO, extracted from FCDB (Fashion Culture Data Base)[(5)] public datasets, in which only detection targets of the above size were present, was performed.

Table 2 Training data for multi-step training

Step	Training data	Number of images
0	MS-COCO	120,000
1	Optimized FCDB	200,000
2	Images superimposed with computer graphics + MS-COCO	5,000 + 10,000
3	Images of persons in areas likely to cause undetected or over-detected cases	280

Next, training with acquired background images superimposed with computer graphics of persons was performed. For the images of persons, computer graphics models of figures with different body shapes and clothing and in various postures for walking and other movements were created, and rendered using a camera with the angle of view, viewpoint, and lens curvature adjusted to match the camera conditions at the location of operation. The computer graphics thus created were superimposed on the background image, as shown in Figure 3. The number of images superimposed with the computer graphics was set equal to the number of actual images required by the conventional method. In addition to these superimposed images, MS-COCO data was mixed in the training data to prevent over-training. Finally, training with images of persons walking in areas that are likely to cause undetected or over-detected cases was performed. Since the computational cost of training with a large number of images in Step 1 is high, ABCI, an AI Bridging Cloud Infrastructure provided by AIST, was used.

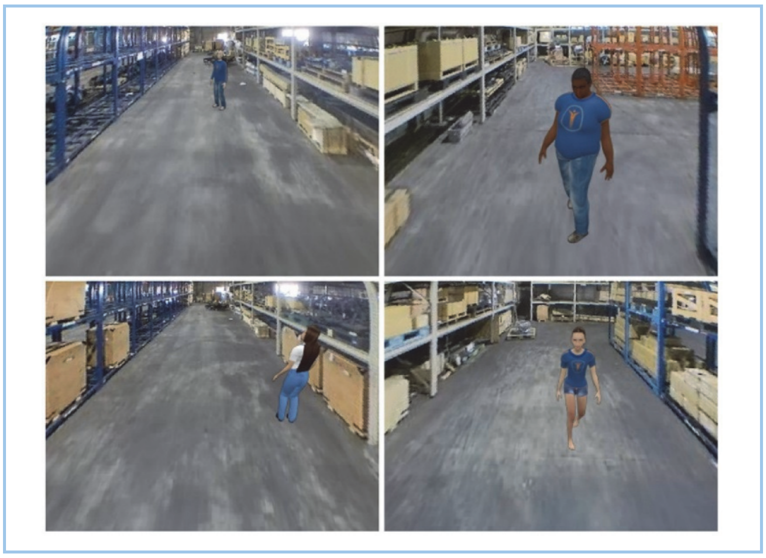


Figure 3 Examples of background image superimposed with computer graphics

4.3 Evaluation of object detector

Images of persons in various postures and orientations in areas that are likely to cause undetected or over-detected cases (Table 1) were used as the evaluation data for the object detector established by the multi-step training. These data were in the conditions with high detection difficulty, so the object detector was evaluated to be capable of detection under other conditions in the actual operational environment if it could detect the detection target successfully under these conditions. However, assuming that over-training may have occurred in multi-step training, images of persons in areas other than the areas that are likely to cause undetected or over-detected cases were added to the evaluation.

The following recall rate and over-detection rates were used as evaluation indices.

$$\text{Recall rate} = \frac{\text{Number of correctly detected cases}}{\text{Number of detection targets}}$$

$$\text{False positive rate} = \frac{\text{Number of over-detected cases}}{\text{Number of detections}}$$

A correctly detected case was defined as a detection where the resulting IoU = A∩B/A∪B, which evaluates the overlap between BBOX A, the correct answer, and BBOX B, the predicted location of the identification result, is greater than 0.25, and the confidence score is greater than an arbitrary threshold, and the other detections were regarded as over-detected cases. In this study, the threshold of the score at which the false positive rate becomes 0.5% or less was adopted, and the reproduction rate at that time was adopted as the detection performance of the object detector.

As shown in **Figure 4**, the multi-step training improved the reproduction rate from 55.9% to 99.0%. It was confirmed that undetected or over-detected cases in the initial model were accurately detected after this training (**Figure 5**). As shown in **Figure 6**, we classified the detection targets into those that appeared larger than 200 to 290 pixels wide by 500 to 690 pixels high (57 to 79% of the total image height) in the image, those that appeared smaller than the above size, and those hidden behind objects, i.e., those in an occlusion(*2) condition, and then counted the number of undetected cases for each of these classifications in each step. **Table 3** shows the results of this attempt. By analyzing the trends indicated by these results, the following findings were obtained.

Table 3 Number of undetected cases in each step

Step	Model	Number of undetected cases			Total
		Detection target appeared smaller	Detection target appeared larger	Detection target in occlusion condition	
0	Initial mode Step 1	209	368	58	635
1	Trained with optimized public datasets	114 55%	148	52	314
2	Trained with images superimposed with computer graphics	109 Step 2	51 66%	33	193
3	Trained with small number of actual images	1	2	11	14

Step 1: Optimization of public datasets significantly improved the detection performance for detection targets appeared in all sizes.

Step 2: Training with images superimposed with computer graphics improved the detection performance for detection targets that appeared larger. On the other hand, no effect was observed for detection targets that appeared smaller.

Step 3: Training with a small number of actual images decreased the number of undetected cases for detection targets in all conditions. However, a certain number of undetected cases remained, mainly for detection targets in an occlusion condition.

(*2) Occlusion : A condition in which the object in front of the camera hides the object behind it.

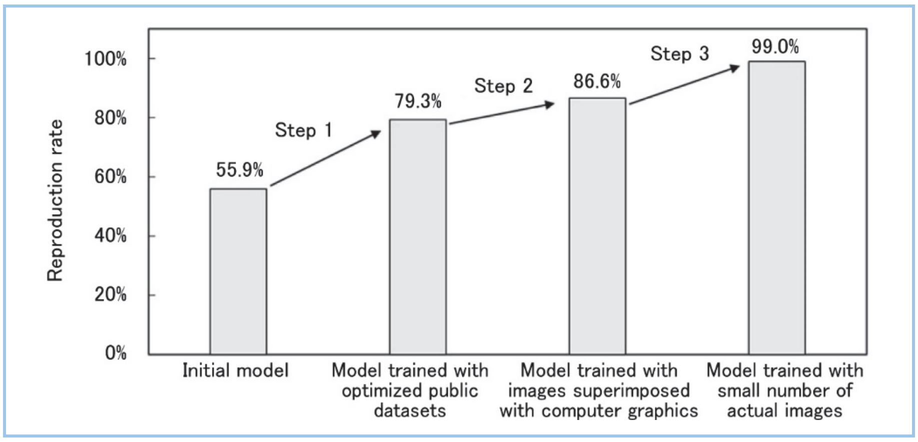


Figure 4 Detection performance in each step of multi-step training

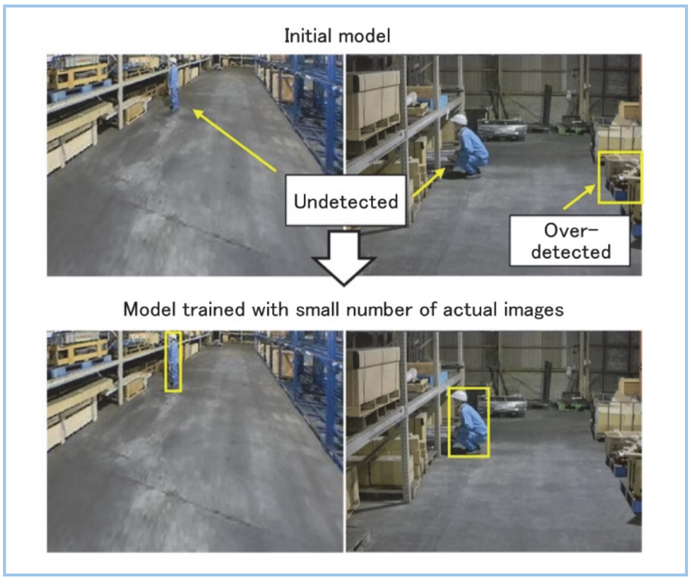


Figure 5 Example detections before and after training

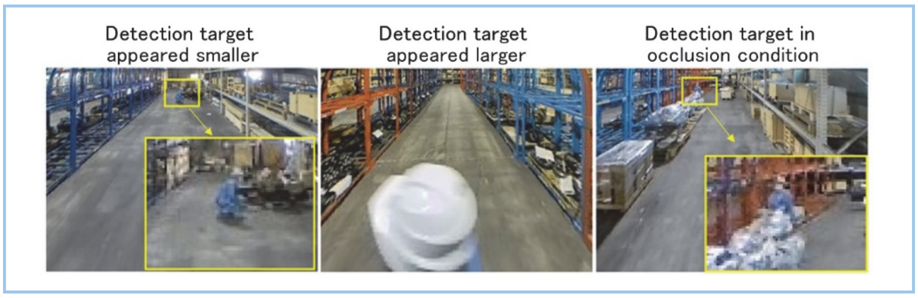

Detection target appeared smaller
Detection target appeared larger
Detection target in occlusion condition

Figure 6 Typical examples of evaluation data

5. Consideration

Based on the results in Section 4.3, it can be inferred that optimization of the public datasets improves the detection performance for detection targets in the relevant size, and that training with images superimposed with computer graphics has the effect of reducing undetected cases for detection targets in postures that can be represented by computer graphics. On the other hand, no significant improvement resulted from the training with images superimposed with computer graphics in the case of detection targets that appeared smaller even though many computer graphics under the same conditions were superimposed, which suggests that the computer graphics representation of detection targets that appeared smaller was far from the features of the actual images. It was also confirmed that detection targets in an occlusion condition are difficult to be detected unless they are reflected in the training data. This condition can be expected to be improved by introducing a superimposed image generation logic that reproduces an occlusion condition by determining the front-back positional relationship in the depth direction between objects in the background and the computer graphics superimposed thereon, using deep learning-based monocular depth estimation[6] or other methods that have been improving in accuracy in recent years.

It was confirmed that an object detector having a high detection performance with a reproduction rate of 99% can be built with only 280 actual images of the detection target when the workflow (Figure 1) introducing multi-step training is used. As shown in **Table 4**, this method was found to be effective in reducing the data acquisition time to one-fourth of the conventional method. Furthermore, because the time required for annotation can also be significantly reduced, we evaluated that this flow improves the efficiency of all the work.

Table 4 Comparison with conventional method

Item compared	Conventional method	Developed method
Number of actual images of detection target	4,647	280
Data collection work period	8 days	2 days
Annotation period	1 month	10 hours

6. Conclusion

In this study, we augmented the training data using public datasets and computer graphics, and developed a technology that can build an object detector having extremely high detection performance even with only a small number of actual images. This achievement is largely due to the establishment of a flow for strategic training using multi-step training, which allows the image acquisition work to be systematically integrated into the object detector building process without repeating data acquisition over and over. This enables the reduction of man-hours and minimizes the impact on customer’s on-site production work. We plan to further advance the method of building object detectors in the future by expanding this technology to various applications and identifying issues.

References

(1) L. Jiao et al., A Survey of Deep Learning-Based Object Detection, IEEE Access, vol. 7, pp. 128837-128868, 2019.

(2) Atsushi Ikeda et al., Cystoscopic diagnosis of bladder cancer using AI, Precision Medicine, 2, 230-233, (2019).

(3) B. Zhou et al., Learning Deep Features for Scene Recognition using Places Database, NIPS 2014, Vol. 1, pp. 487–495, 2014

(4) Lin, Microsoft COCO: Common Objects in Context, arXiv:1405.0312 (2015).

(5) Abe et al., Fashion Culture Database: Construction of Database for World-wide Fashion Analysis., ICARCV, page 1721-1726. IEEE, (2018).

(6) René Ranftl et al., Vision Transformers for Dense Prediction, Proc. ICCV 2021, pp. 12179-12188.