Comparing the Recognition Accuracy of Humans and Deep Learning on a Simple Visual Inspection Task

Naoto Kato, Michiko Inoue, Masashi Nishiyama, and Yoshio Iwai

Graduate School of Engineering, Tottori University, Japan
nishiyama@tottori-u.ac.jp

Abstract. In this paper, we investigate the number of training samples required for deep learning techniques to achieve better accuracy of inspection than a human on a simple visual inspection task. We also examine whether there are differences in terms of finding anomalies when deep learning techniques outperform human subjects. To this end, we design a simple task that can be performed by non-experts. It required that participants distinguish between normal and anomalous symbols in images. We automatically generated a large number of training samples containing normal and anomalous symbols in the task. The results show that the deep learning techniques required several thousand training samples to detect the locations of the anomalous symbols and tens of thousands to divide these symbols into segments. We also confirmed that deep learning techniques have both advantages and disadvantages in the task of identifying anomalies compared with humans.

Keywords: Visual inspection · deep learning · people.

1 Introduction

In recent years, the automation of the manufacturing industry has been promoted to mitigate labor shortage [12, 1, 2, 6]. We focus here on visual inspection, of the various tasks performed manually by workers in a factory. Visual inspection is the task of finding such anomalies in products as scratches, dents, and deformations on a manufacturing line. Deep learning techniques [14, 9, 5, 13, 18] are widely used to automate visual inspection, and have achieved better performance than humans on various applications, such as object recognition [8, 4] and sketch search [19].

Network models of deep learning techniques to automatically predict anomalies are generated using training samples. To improve the accuracy of inspection using deep learning techniques, we need to prepare a large number of training samples containing stimulus images and labels indicating the presence or absence of anomalies. However, it is laborious to correctly assign anomaly-related labels to the stimulus images because the labels are manually assigned by experts through visual inspection. Furthermore, the number of these experts is small.

In this paper, we consider a simple task where non-experts can assign anomaly-related labels. We design the simple task based on properties of the visual inspection of manufacturing lines. A minority of the stimulus images contained anomalies while the majority were normal. Some of these anomalies were easy to find whereas others were more challenging. In experiments using the simple task, we investigated the number of training samples needed for deep learning techniques to deliver higher accuracy than humans. To this end, we automatically generated labels indicating anomalies in stimulus images used for the simple task. We compared the accuracy of inspection of human subjects with that of deep learning techniques. This study is the first step in investigating knowledge we can obtain from a comparison of accuracy between humans and the deep learning techniques. While the simple task cannot comprehensively represent visual inspection by experts, we think that this study can provide new guidelines on data collection for deep learning techniques, especially manually assigning labels through the interaction between people and information systems.

The remainder of this paper is organized as follows: Section 2 describes the detail of the simple task considered here, and Section 3 presents the accuracy of inspection of human subjects. Section 4 presents the accuracy of inspection of deep learning techniques, and Section 5 contains our concluding remarks.

2 Design of the simple task

2.1 Overview

To design the simple task of visual inspection that non-experts can perform, we consider the property of a general task of visual inspection at a factory. In this task, the worker manually determines the presence or absence of anomalies, such as scratches, dents, or deformations, by observing products on a manufacturing line. We assume that a certain surface of a product has many symbols arranged in a grid. The simple task is to check the symbols to determine whether there are anomalies. A label indicating an anomaly means that part of the symbol is defected. We call the symbol pattern on the grid the stimulus image. We assume that the position where the symbol is placed on the grid is fixed but rotation is not. Figure 1 shows an example of the simple task. Details of the stimulus images of the simple task are described below.

2.2 Generation of stimulus images

We set $8 \times 4 = 32$ symbols on a grid in a stimulus image. To generate an anomalous symbol, we altered part of a normal symbol. A stimulus image is generated by the following steps:

S1: We randomly set the maximum number of anomalous symbols to zero, two, four, and six in a stimulus image. Note that zero means that all symbols are normal. Six anomalies are determined by considering the relationship among the numbers 4 ± 1 [3] and 7 ± 2 [10] of the short-term memories of people.

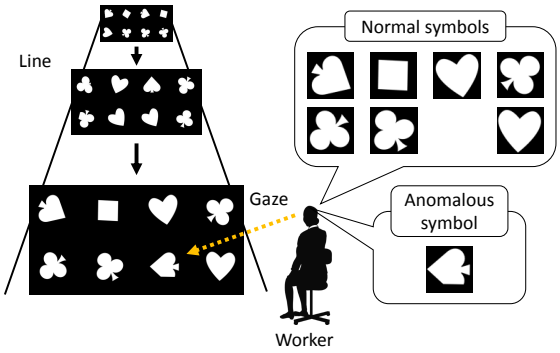

Fig. 1. Overview of the simple task of visual inspection by non-experts.

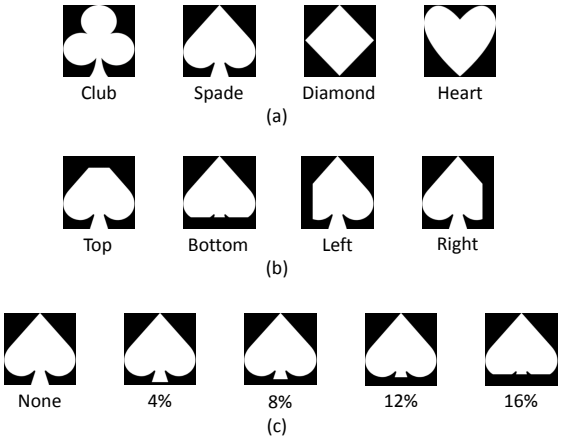

Fig. 2. Examples of the parameters used to generate anomalous symbols. We show the suits in (a), the defective positions in (b), and the rate of defect in (c).

S2: We determine the parameters of a given suit, rate of defect, defective position, and angle of rotation to generate an anomalous or a normal symbol.
- We randomly select a suit from among club, spade, diamond, and heart. Figure 2(a) shows examples of the suits.
- We randomly determine whether the given symbol is defected. Note that a symbol is not defective when the given number of anomalous symbols is the maximum determined in S1.
- We set the rate of defect and the defective position to generate the anomalous symbol.
  - We randomly set the position from among top, bottom, left, and right. Figure 2(b) shows examples of defective positions.
  - We randomly set the rate of defect to 4%, 8%, 12%, or 16%. Figure 2(c) shows examples of the rate of defect.

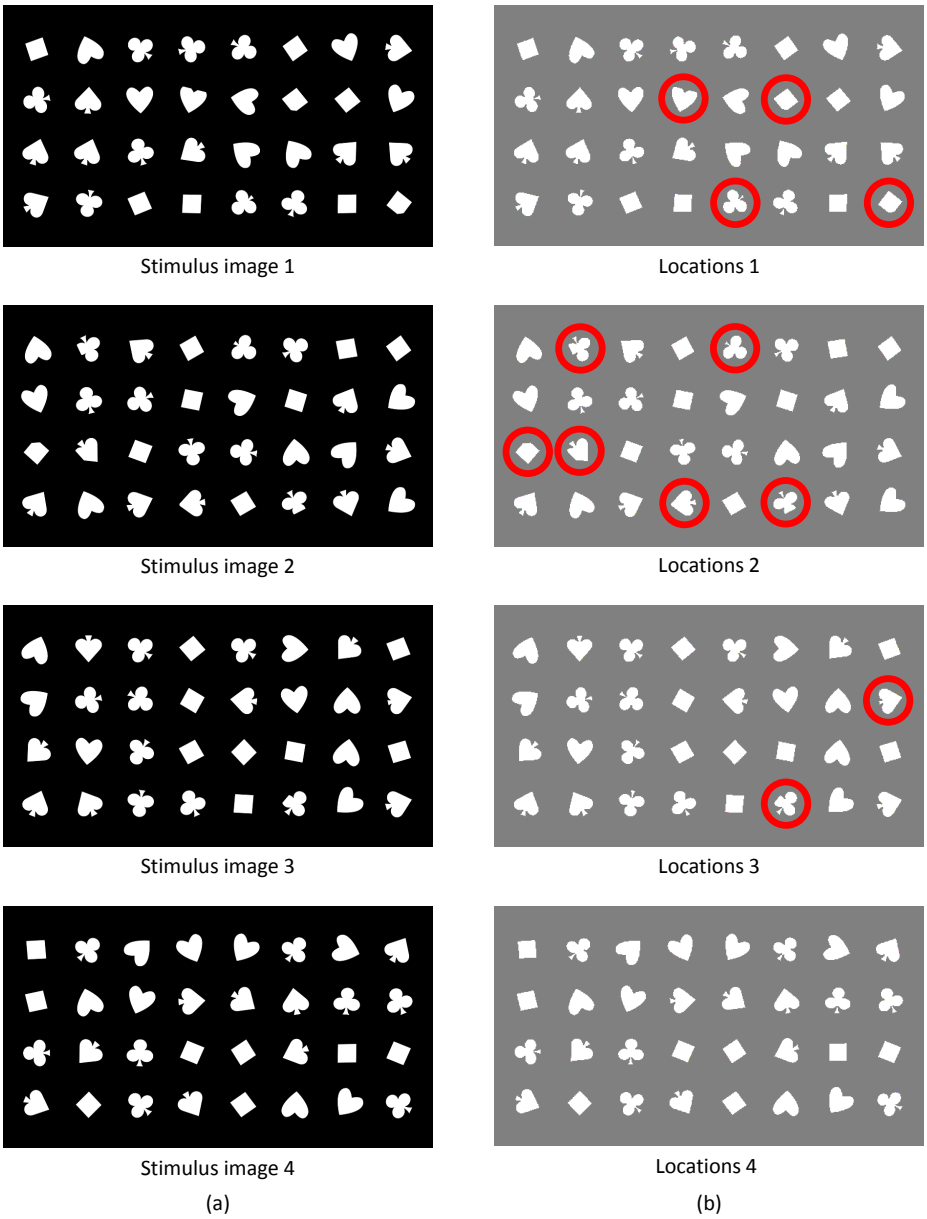

Fig. 3. Examples of stimulus images. The red circles indicate the locations of anomalous symbols.

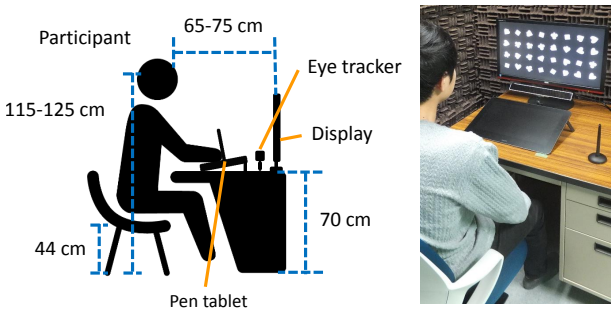

Fig. 4. Setting of the simple task for the participants.

– We randomly set the angle of rotation in the range −180–180 degrees in steps of 1 degree.

S3: We generate an anomalous or a normal symbol using the parameters determined in the above steps.

S4: We place the generated symbol at an intersection on the grid.

S5: We repeat S2, S3, and S4 until the number of symbols in the stimulus image is smaller than or equal to $4 \times 8 = 32$.

Figure 3(a) shows examples of the generated stimulus images and (b) shows locations of anomalous symbols in them. There is a small possibility that the same stimulus images reappear because the total number of variations in stimulus images is 2.3×10^{21}. We calculated inspection accuracy on the simple task using the generated stimulus images. Section 3 describes the accuracy of inspection of human subjects and Section 4 describes that of the deep learning techniques used.

3 Inspection accuracy of human participants

3.1 Setting

We investigated the accuracy of visual inspection of non-experts on the simple task. Twenty people (15 males, five females, average age, 22.2 ± 1.0 years, graduate school students) participated in the study. Figure 4 shows the settings of the simple task performed by the participants in a dark room. The intensity of light in the dark room was 360 ± 5 lx. A participant sat in a chair in a comfortable posture. We used 24-inch display (AOC G2460PF 24) with a resolution off 1920×1080 pixels to show the stimulus image. We measured the gaze locations of the participant using a standing eye tracker (Gazepoint GP3 Eye Tracker, sampling rate 60 Hz)because gaze has a potential capability to increase the accuracy of various recognition tasks[16, 7, 11, 17]. To record anomalous symbols found by the participant, a pen tablet (Wacom Cintiq Pro 16) was used.

3.2 Experimental procedures

We asked the participants to perform the simple visual inspection task using the following procedure:

P1: We randomly selected a participant.
P2: We explained the experiment to the participant.
P3: The participant performed visual inspection using a stimulus image on the display. We simultaneously measured the gaze locations of the participant.
P4: The participant recorded locations of anomalous symbols on the pen tablet by marking them.
P5: We repeated P3 and P4 until all 12 stimulus images had been examined by the participant.
P6: We repeated P1 to P5 until all 20 participants had finished the simple task.

The details of P2, P3, and P4 are described blow.

P2: Explanation of the instruction We explained to the participants the rule of the simple task, the procedure of gaze measurement, and how to use the pen tablet for marking the symbols. In this procedure, the participant was allowed to practice simple visual inspection task. Once the participant had completed the example task, we informed them of the correct answers for the locations of the anomalous symbols. Note that we did not provide the answers to the participants in the procedures below.

P3: Visual inspection for measuring gaze locations The participants performed the simple visual inspection task by viewing a stimulus image on the display. Figure 5 illustrates the procedure of P3. We guided the initial gaze of the participant before he/she viewed the stimulus image by inserting blank image 1, containing a fixation point, on the center of the image. The blank image 1 was shown for two seconds. We then showed a stimulus image on the display for 30 seconds, which was considered sufficient time for the participant to check all symbols. We measured the gaze locations of the participants while they observed the stimulus image. We showed blank image 2 to the participant for five seconds after he/she had completed the task. While the participant was viewing the stimulus image on the display, we turned off the pen tablet.

P4: Marking anomalous symbols Each participant indicated the locations of the anomalous symbols by marking symbols on the pen tablet. We showed the same stimulus image displayed in P3 on the pen tablet. The participant circled anomalous symbols using a pen. Figure 6 shows examples of the circled symbols. While the participant was marking symbols in the stimulus image, we turned off the display. We embedded an eraser function in case the participant accidentally circled something and wanted to remove it. The maximum time allowed for marking anomalous symbols was 30 seconds. As soon as the marking had been finished, we moved to the next procedure.

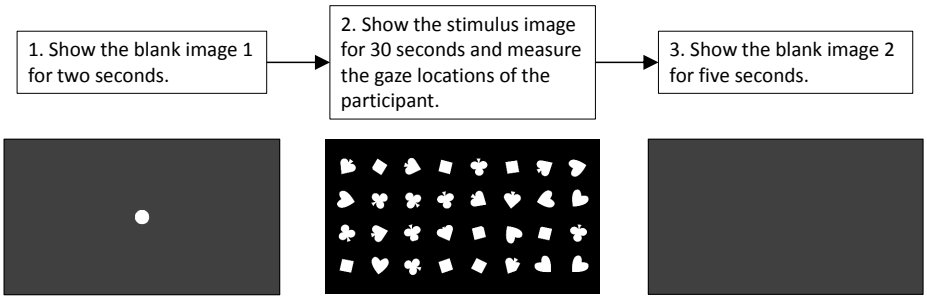

Fig. 5. Procedure of the participant in P3 viewing the stimulus image.

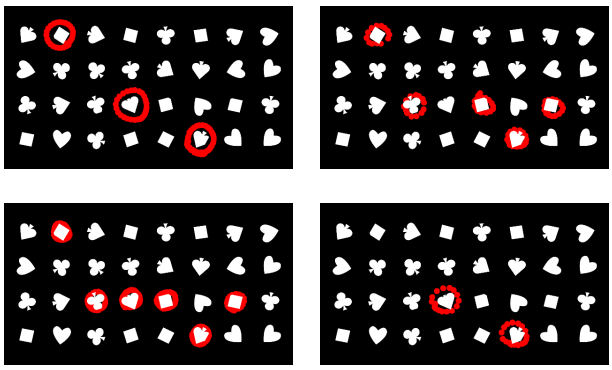

Fig. 6. Examples of symbols circled by the participants in the marking procedure.

3.3 Results

Table 1 shows the F-measure, precision, recall, and accuracy of the simple task performed by the participants. Precision was high at 0.97, and the participants rarely made a mistake in identifying a normal symbol as an anomalous symbol. On the contrary, the recall rate was 0.85, indicating that the participants had missed a large number of anomalous symbols when identifying them.

To check for anomalous symbols that had been incorrectly identified as normal symbols, we calculated the recall rate of each parameter used to generate the symbols. Figure 7(a) shows the recall rate of each suit, (b) shows that of each defective position, and (c) shows the recall rate for each rate of defect. We used Bonferroni's method as a multiple comparison test. There was no significant difference among suits in (a). We thus cannot claim that the suit influenced the inspection accuracy of the participants. However, there was a significant difference between the results of defective positions for the bottom and top, bottom and left, and bottom and right in (b). Significant differences were also observed between the results of rates of defect for 4% and 8%, 4% and 12%, 4% and 16%, and 8% and 16% in (c). Thus, the participants frequently missed the anomalous symbols in the downward defective position and those with a low rate of defects.

Table 1. The inspection accuracy of participants on the simple task.

F-measure	Precision	Recall	Accuracy
0.91	0.97	0.85	0.97

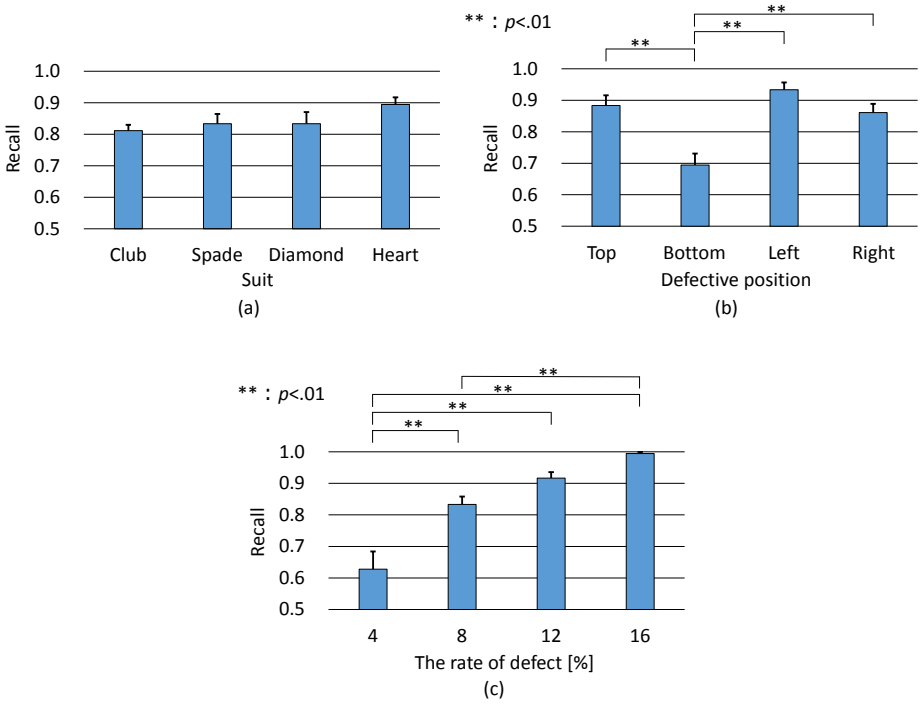

Fig. 7. Recall rates of the participants by suit, defective position, and the rate of defects.

3.4 Analysis of gaze locations of the participants

We analyzed gaze locations while the participants performed the simple visual inspection task by recording the duration of gaze fixation of all participants. Figure 8 shows heatmaps of the duration of fixation and inspection accuracy on stimulus images of the simple task. In the heatmap, a deeper red means longer gaze fixation. In the heatmap of inspection accuracy, a deeper red indicates better performance. The results are summarized below.

- Gazes of the participants lingered around anomalous symbols for a longer time than on normal symbols.
- Certain anomalous symbols on which the participants recorded low accuracy featured long durations when their gazes were fixed. We think that the participants could simply not decide whether the given symbol contained anomalies.

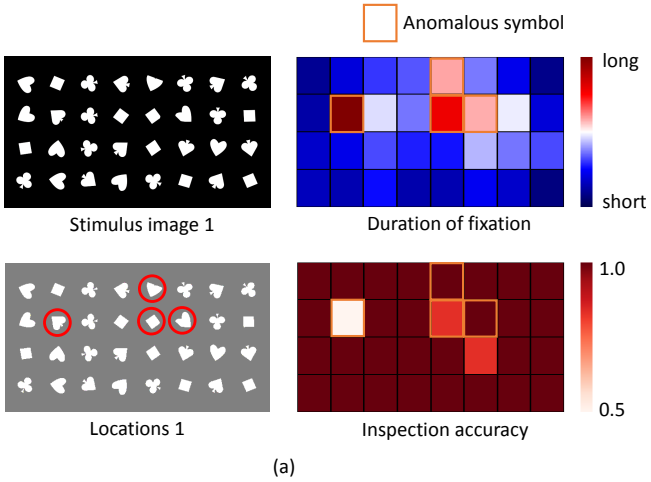

Fig. 8. Heatmaps of the duration of fixation and inspection accuracy on the stimulus images of the simple task.

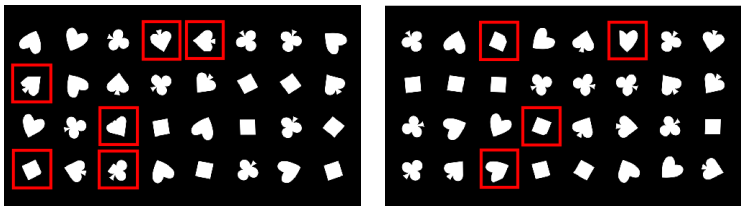

Fig. 9. Examples of outputs predicted using the SSD.

– Normal symbols with high accuracy of recognition featured a short duration of fixed gazes. We think that in these cases, the participants quickly judged that there were no anomalies.

4 Inspection accuracy of deep learning techniques

4.1 Overview

We investigated the inspection accuracy of representative deep learning techniques—the single-shot multibox detector (SSD) [9] and U-Net [14]—on the simple task. SSD was designed for localization tasks and U-Net for segmentation task. To prepare a large number of training samples for the deep learning techniques, we used the stimulus images and labels generated by the steps described in Section 2.2. The results of the SSD are described in Section 4.2 and those of U- Net in Section 4.3.

4.2 Visual inspection using SSD

Experimental conditions We used bounding boxes of the anomalous symbols as labels to train the SSD model to predict their locations in the stimulus image. Figure 9 shows examples of the outputs predicted by the SSD. We used the VGG16 model [15] for the base network of the SSD. A total of 2,500 to 10,000 samples were prepared. For the test samples, we generated 1,000 stimulus images that were not used for training samples of the SSD. We repeated these procedures three times to evaluate inspection accuracy.

Inspection accuracy of SSD Figure 10(a) shows the F-measure of the simple task using the SSD for different numbers of training samples. Bonferroni's method was used as a multiple test. There was a significant difference in the results between 2,500 and 5000, 2,500 and 7,500, and 2,500 and 10,000 samples. We confirmed that the inspection accuracy using the SSD was satisfactory when the number of training samples was more than or equal to 5,000. We also confirmed that the accuracy of the SSD of 0.95 was superior than that of the human participants of 0.91 when using 5,000 training samples.

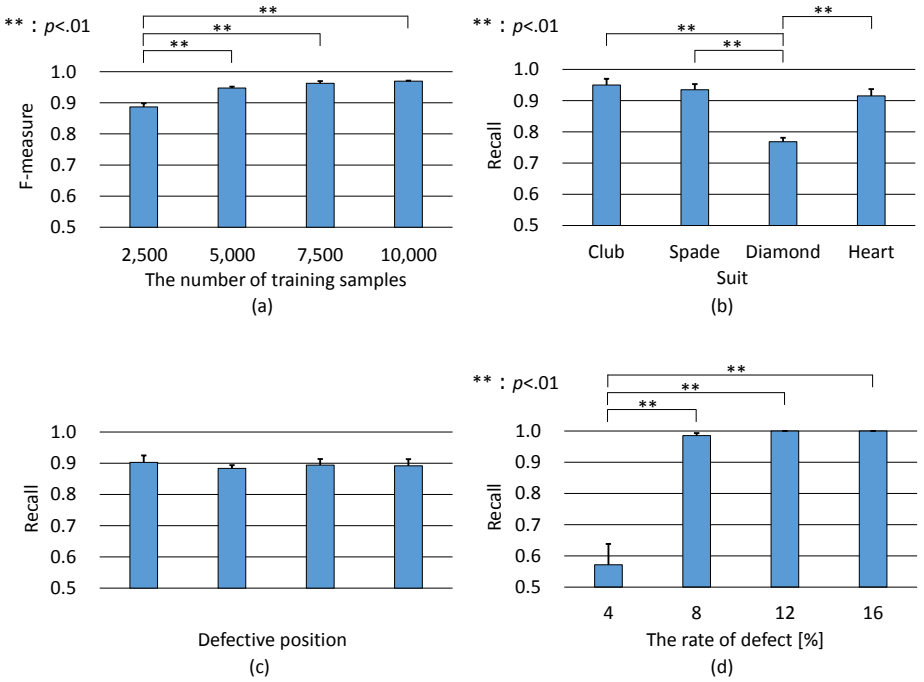

Fig. 10. Inspection accuracy using the SSD on the simple task.

Figures 10(b), (c), and (d) show the recall rates for the suits, defective positions, and the rate of defects. There was a significant difference between diamond and club, diamond and spade, and diamond and heart in (b). The SSD did not perform well on diamond. On the contrary, there was no significant difference among the defective positions in (c). Furthermore, there was a significant difference between 4% and 8%, 4% and 12%, and 4% and 16% in (d). The SSD performed poorly at a rate of defect of 4%.

4.3 Visual inspection using U-Net

Experimental conditions To train the U-Net model, we used label images in which only anomalous symbols appeared. The model was trained to predict segments of the only anomaly regions in the stimulus image. Figure 11 shows examples of the outputs predicted by U-Net. We used nine downsampling layers and nine upsampling layers, and used 5,000 to 20,000 training samples at intervals of 5,000. For the test samples, we generated 1,000 stimulus images not used as training samples of the U-Net model. We repeated these procedures three times to evaluate inspection accuracies.

Inspection accuracy of U-Net Figure 12(a) shows the F-measure of U-Net on the simple task for different numbers of training samples. Bonferroni's method

Fig. 11. Examples of outputs predicted using U-Net.

was used as multiple test. There was a significant difference in the results between 5,000 and 10,000, 5,000 and 15,000, and 5,000 and 20,000 samples. We confirmed that inspection accuracy using U-Net was good when the number of training samples was more than or equal to 10,000. We also confirmed that the accuracy of U-Net of 0.94 was superior than that of the human participants of 0.91 when using 15,000 training samples.

Figures 12(b), (c), and (d) show the recall rates by suit, defective position, and the rate of defects. There was a significant difference between diamond and club, and diamond and spade in (b). U-Net performed poorly on diamond. On the contrary, there was no significant difference by defective position in (c), but there was a significant difference between results for 4% and 8%, 4% and 12%, and 4% and 16% in (d). U-Net performed poorly at a rate of defect of 4%.

4.4 Comparison with human participants

The participants as well as the deep learning techniques delivered poor results at a small rate of defects. The participants were good at identifying defects in diamond suit but the deep learning techniques were not. The participants performed poorly at the downward defective position but the deep learning techniques were good at this. To increase the inspection accuracy of the deep learning techniques, several thousand stimulus images and labels are needed for the localization task and tens of thousands for the segmentation task.

5 Conclusions

In this paper, we investigated and compared the inspection accuracy of humans and deep learning techniques on a simple visual inspection task. The task consisted of checking whether a symbolic pattern contained anomalies. Experimental results revealed the number of training samples needed by deep learning techniques to match or exceed the accuracy of human subjects. They also revealed the differences in accuracies between humans and deep learning techniques. In future work, we will expand our assessment to a long, more complex task to represent the various practical applications of visual inspection.

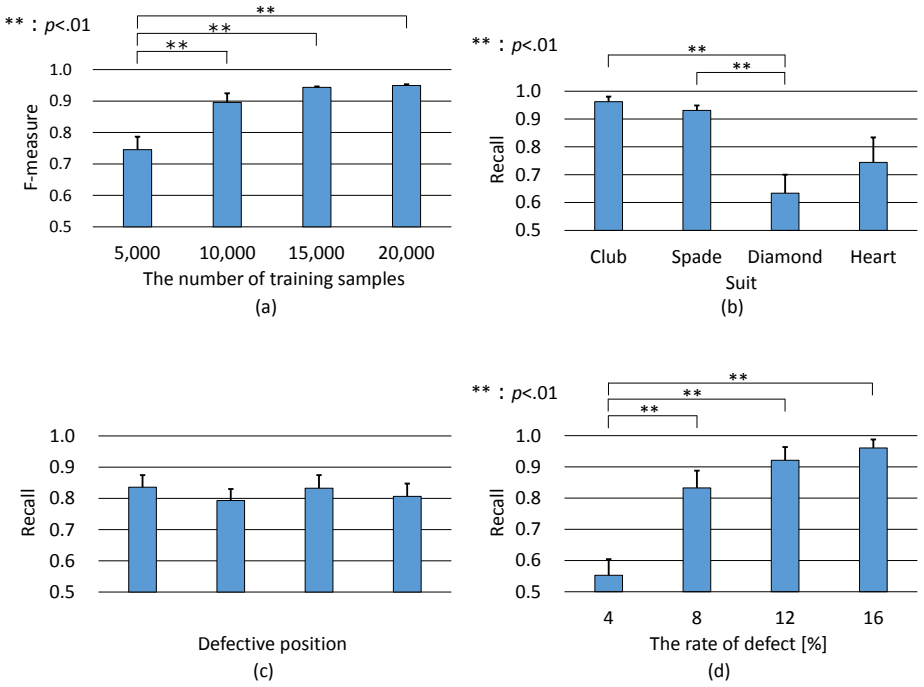

Fig. 12. Inspection accuracy using U-Net on the simple task.

References

1. Beyerer, J., F. P. Leon, C.F.: Machine vision: Automated visual inspection: Theory, practice and applications. Springer (2015)
2. Cha, Y.J., Choi, W., Suh, G., Mahmoudkhani, S., Buyukozturk, O.: Autonomous structural visual inspection using region-based deep learning for detecting multiple damage types. Computer-Aided Civil and Infrastructure Engineering **33**(9), 731–747 (2018)
3. Cowan, N.: The magical number 4 in short-term memory: a reconsideration of mental storage capacity. Behavioral and Brain Sciences **24**(1), 87–114 (2001)
4. Dodge, S., Karam, L.: A study and comparison of human and deep learning recognition performance under visual distortions. In: Proceedings of the 26th International Conference on Computer Communication and Networks. pp. 1–7. ICCCN (2017)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778. CVPR (2016)
6. Huang, S.H., Pan, Y.C.: Automated visual inspection in the semiconductor industry: A survey. Computers in Industry **66**, 1 – 10 (2015)
7. Jiaxin Wu, S.h.Z., Ma, Z., Heinen, S.J., Jiang, J.: Gaze aware deep learning model for video summarization. In: Proceeding of Pacific Rim Conference on Multimedia. pp. 285–295 (2018)

8. Kheradpisheh, S.R., Ghodrati, M., Ganjtabesh, M., Masquelier, T.: Deep networks can resemble human feed-forward vision in invariant object recognition. Scientific Reports **6**, 32672:1–24 (2016)
9. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: Proceedings of the European Conference on Computer Vision. pp. 21–37. ECCV (2016)
10. Miller, G.A.: The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychological review **63**(2), 81 (1956)
11. Murrugarra-Llerena, N., Kovashka, A.: Learning attributes from human gaze. In: Proceeding of Winter Conference on Applications of Computer Vision. pp. 510–519 (2017)
12. Newman, T.S., Jain, A.K.: A survey of automated visual inspection. Computer Vision and Image Understanding **61**(2), 231 – 262 (1995)
13. Proceeding of Rekabdar, B., Mousas, C.: Dilated convolutional neural network for predicting driver's activity. In: Proceeding of International Conference on Intelligent Transportation Systems. pp. 3245–3250. IEEE (2018)
14. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Proceedings of the Medical Image Computing and Computer-Assisted Intervention. pp. 234–241. MICCAI (2015)
15. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: Proceedings of the International Conference on Learning Representations (2015)
16. Tavakoli, H.R., Rahtu, E., Kannala, J., Borji, A.: Digging deeper into egocentric gaze prediction. In: Proceedings of Winter Conference on Applications of Computer Vision. pp. 273–282. IEEE (2019)
17. Tingting Qiao, J.D., Xu, D.: Exploring human-like attention supervision in visual question answering (2018)
18. Wang, X., Gao, L., Song, J., Zhen, X., Sebe, N., Shen, H.T.: Deep appearance and motion learning for egocentric activity recognition. Neurocomputing **275**, 438–447 (2018)
19. Yu, Q., Yang, Y., Liu, F., Song, Y.Z., Xiang, T., Hospedales, T.M.: Sketch-a-net: A deep neural network that beats humans. International Journal of Computer Vision **122**(3), 411–425 (2017)