Improving a Physical Search System that Detects Even Unknown Displaced Objects Using Image Differences

Shin Kajihara
Graduate School of Science and Engineering
Saga University
Saga, Japan
email: 20634002@edu.cc.saga-u.ac.jp

Masato Okazaki
Graduate School of Science and Engineering
Saga University
Saga, Japan
email: 22726005@edu.cc.saga-u.ac.jp

Chika Oshima
Faculty of Science and Engineering
Saga University
Saga, Japan
email: sj5872@edu.cc.saga-u.ac.jp

Koichi Nakayama
Faculty of Science and Engineering
Saga University
Saga, Japan
email: knakayama@is.saga-u.ac.jp

***Abstract*—This paper introduced a Physical Search System (PSS), which detects all objects displaced in a physical space with two cameras and a computer, based on image difference detection technology. A Linking Method (LM) improved the image clustering accuracy in the PSS. Using the PSS, displaced objects are detected by the differences between two images ordered in a time series photographed by two cameras each. In a Two-step Feature clustering Algorithm (TFA), features of the displaced object images are extracted from the deep residual network ResNet-50 pre-trained on ImageNet. Then, the displaced object images with features are grouped into clusters using a two-pass process. However, these clusters can include a few noisy images. In this paper, therefore, in parallel with the execution of TFA, the displaced object images were paired by the LM, as follows. Some cropped displaced object images were photographed by two cameras at the same time. By calculating the similarities between the displaced object images, pairs of images of the same displaced objects could be created. Namely, the noisy images do not pair with other images. Finally, the displaced object images in the clusters created via TFA that did not overlap with the images of pairs of groups created by LM were deleted. The results of the experiment showed that the method of combining the two algorithms indicated higher clustering accuracy than using TFA alone.**

***Index Terms*—Difference detection; deep learning; physical search system.**

I. Introduction

In the Cyber-Physical System [1] [2] proposed previously, many sensors, actuators, and control devices that acquire information are connected by a network. For example, if a sensor (e.g., an Apple Air Tag) is attached to each object, it will be easier to identify the location of the lost item; however, it is difficult to attach such devices to all objects.

In recent years, advances in deep learning technology have made it easier than ever to identify objects from camera images. A method of displaced object extraction based on differential detection [3] is proposed. Then, the images of any displaced object within the photographed area can be cropped and stored automatically without pre-training data [3] [4].

A Physical Search System (PSS) [5] can detect displaced objects by grouping images of the displaced objects into clusters and then searching for all displaced objects in a physical space using two cameras and a computer based on images difference detection technology [3] [4]. The cropped images leaving only the objects are entered into ResNet50 [6] [7], a pre-trained convolutional neural network. The features of the displaced object images are extracted using ResNet50 [6] [7] trained on ImageNet [8]. The displaced object images are then grouped into clusters based on the image features using a two-pass x-means clustering. This process is called “a Two-step Feature clustering Algorithm” (TFA) [5]. However, users of the PSS need to visually check and delete any noisy images that are generated by factors such as how the light hits.

In this paper, to improve the accuracy of the displaced object images clustering, a new algorithm, a Linking Method (LM), is proposed and combined with TFA. The LM creates pairs of displaced object images with high similarity based on the photographs taken by two cameras at the same time. Images obtained from only one camera cannot be paired, because they are likely to be images that failed to be photographed, captured, or cropped. The images that have undergone this process contribute to further improving the accuracy of images clustered using TFA.

The next section describes the TFA [5], which groups the displaced object images into clusters. Then, details of the LM are presented. Section III shows an experiment that compares the accuracy of clustering between the combining LM with TFA and using TFA alone. Finally, Section IV concludes the paper.

II. Physical Search System (PSS)

A. Overview

This section explains the overall structure of the proposed PSS [5]. Figure 1 shows the PSS' hardware configuration. The PSS consists of two cameras that constantly capture the target area and a computer that processes the images photographed by the cameras. The PSS' software configuration consists of a displaced object detection unit that extracts the displaced objects from images photographed by each camera, a displaced object image-clustering unit that creates clusters, and a search results display unit that retrieves and displays the displaced objects.

In the displaced object detection unit [5], the photographed images are processed in the order in which they are photographed. The image at a certain time is then compared at the pixel level with the image photographed at a previous time. When a pixel with a difference of a certain standard or more is detected, it is determined that something has been displaced. The PSS can also detect people, and the photographs can define areas in which no one or nothing is present. In other words, the PSS does not yet detect objects that are moving/rotating around of the center of gravity, but it can detect displaced objects by comparing sets of images.

When objects overlap, we can obtain expected results, if they are displaced in order. For example, Object A is placed at a certain place. After the system has photographed an image near Object A, Object B is placed on top of Object A. If the PSS photographs the place again before each object is moved, both Objects A and B will be detected correctly. When two overlapping objects move together, Objects A and B are detected as a single object, so if Object A is pulled out from under Object B, Object A will not be detected. However, if Object A is placed elsewhere, it will be detected as a displaced object.

Figure 2 shows the differences between the two images in white. The target area is cropped as a rectangle [3]. The cropped image is called a "displaced object image." As described in Section II-B), in the displaced object images clustering unit, the displaced objects' images are grouped into clusters that contain the images of the same objects in each location and stored in the PSS. In the search result display unit, as shown in Figure 3, when a PSS user searches for a lost object (a displaced object), the search results are displayed in an application that displays augmented reality (AR) using an AR marker and an AR display terminal [5].

B. Object images clustering unit

This section describes the TFA [5], which groups the displaced object images into clusters. Next, the LM is proposed. Then, a method for combining the TFA and the LM is explained.

1) Two-step Feature clustering Algorithm (TFA): ResNet50 [6] [7] is applied to the displaced object images to quantify their features. ResNet50 is a convolutional neural network that is pre-trained on ImageNet [8], an image database. Therefore, the user does not need to prepare any image-learning data.

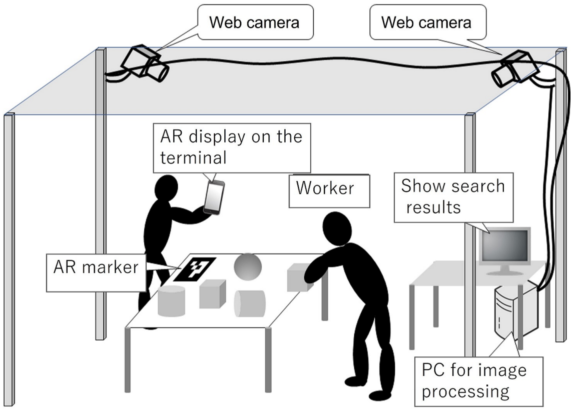

Fig. 1. Construction of the PSS hardware [5]

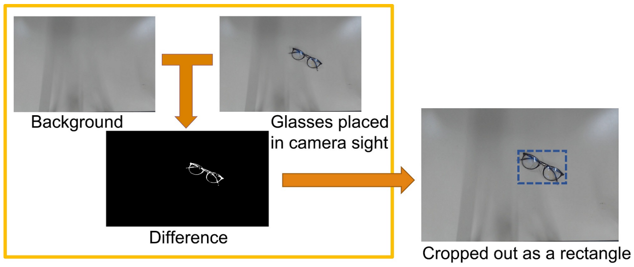

Fig. 2. How to crop out a displaced object [3].

ResNet is a residual network designed to alleviate the vanishing/exploding gradient problem caused by stacking residual blocks. In ResNet-50, one residual block consists of three convolution layers. The size of the convolution kernel, which is the element of convolution operation in the convolutional layer, should be smaller than the size of the input image [9]. The stacked layers in the residual blocks have 1×1, 3×3, and 1×1 convolution layers. The 1×1 convolution first reduces the dimensions. In the next layer, the bottleneck 3×3 layer, the features of the images are calculated. Then, the dimension of depth is again added in the next 1×1 layer (bottleneck) [10]. The final convolutional layer outputs 2048 feature maps of size 7×7.

In the PSS, the images of the displaced object are resized to 224 $\times$ 224 pixels and entered into ResNet50. Then, each

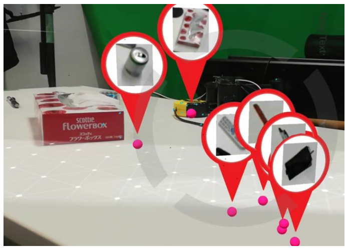

Fig. 3. Icons that display displaced objects in the AR application [5].

resized image is flattened into the 100352-dimentional vector ($1 \times 7 \times 7 \times 2048$, which is a tensor: $depth(none) \times width \times height \times channel$) [11]. We call these "image features" in this paper.

Then, the displaced object images with the feature are processed with the x-means clustering algorithm [12], which is a method of clustering while automatically estimating the number of clusters k of k-means [13]. Therefore, the x-means clustering is a type of unsupervised learning like the k-means, wherein the data points (the features of the displaced object images) are grouped into different clusters based on their degree of similarity. A cluster number is assigned to each cluster, as determined by the x-means method. All displaced object images are stored in folders according to their cluster number.

There are a few folders (clusters) in which only noisy images are included. The PSS user manually deletes these. Then, the same processing protocol as used in the first stage is performed again for all images in the remaining clusters (the second step). Some noisy images will remain in a few folders [5], so there is room for improvement in accuracy.

2) Linking method: We propose a linking method in this section to improve the accuracy of the TFA clustering. In the PSS [5], two cameras (Cameras A and B) usually take pictures of the same displaced object at the same time from different angles. However, noisy images are photographed by only one of the two cameras, because noisy images are a result of misrecognition due to light rays or mistakes in cropping out the object parts of the images. Therefore, as shown in Figure 4, in the LM, pairs of the displaced object images with high degrees of similarity are created from the displaced object images derived from the photographs taken simultaneously by Cameras A and B. This process is called "linking" in this paper. In other words, a pair combination is created with a displaced object image derived from Camera A's photograph and another displaced object image derived from Camera B's photograph. Displaced object images obtained from only one of the cameras cannot be paired.

Next, the method for calculating the similarity between the displaced object images is explained. "imgsim [14]" is a library for computing perceptual hashes of images. The "distance" between images can be calculated using the imgsim library. The distances between the displaced object image, "a1," derived from Camera A's photograph and the displaced object images, "b1 to bx," derived from Camera B's photograph, are calculated. The higher the degree of image similarity, the smaller the distance between them. The distance between identical images is 0.

As shown in Figure 5, pairs are created in order, starting from those with the smallest distance value (the highest degree of similarity) between two images. For example, when two displaced object images are obtained from Cameras A and B's photographs taken at a certain time, there are four possible pair combinations. The image that is paired with another image is excluded from the candidate images for the other pairs. In addition, combinations with distance values exceeding 23 are not considered pairs. A gathering of the pairs is called a "pair group."

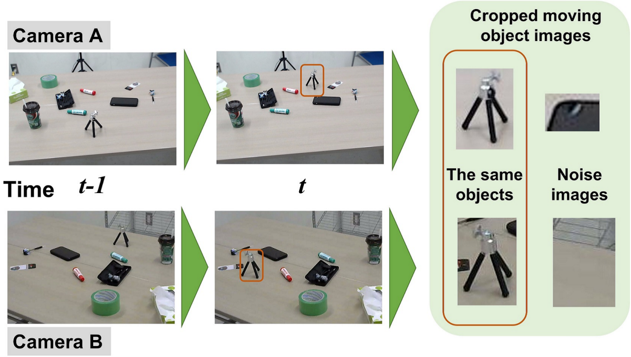

Fig. 4. Noisy images cannot be paired with another image.

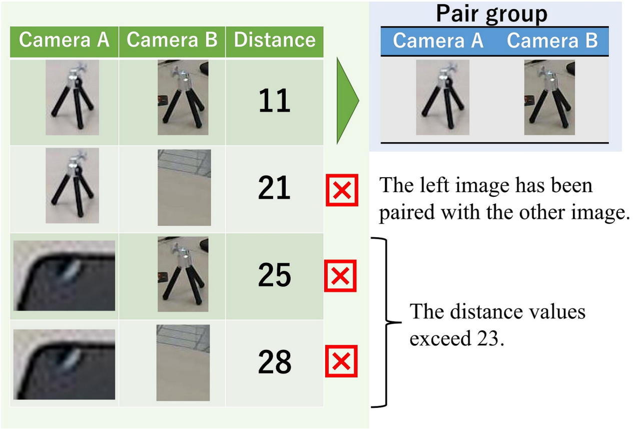
Camera A
Camera B
Distance
11
21
25
28
Pair group
Camera A
Camera B
The left image has been paired with the other image.
The distance values exceed 23.

Fig. 5. Create a pair based on the similarity between two images.

3) Brush up TA clustering results with pair groups created using LM: Figure 6 shows that the displaced object images are updated by comparing the TFA results with that of LM; then, the clusters (folders) are reorganized. The displaced object images in the clusters that do not overlap with the displaced object images of the pair group are deleted. In this process, the noisy images and the images for which one camera has failed to detect a displaced object can be deleted from the folders.

Finally, the clusters created by TFA are reorganized. If two displaced object images that are paired belong to different clusters, they are processed as follows: the similarity (distance) between each of two images and other images that belong to the same cluster of each of two images is calculated using the imgsim library. Next, the averages of the distances in each cluster are calculated. The one with the larger average value moves to the cluster that includes the other displaced object image with the smaller average value. For example, Image_p, which belongs to Cluster_P, is paired with Image_q, which belongs to Cluster_Q. The distance values are calculated between Image_p and each of the other images in Cluster_P, and between Image_q and each of the other images in Cluster_Q.

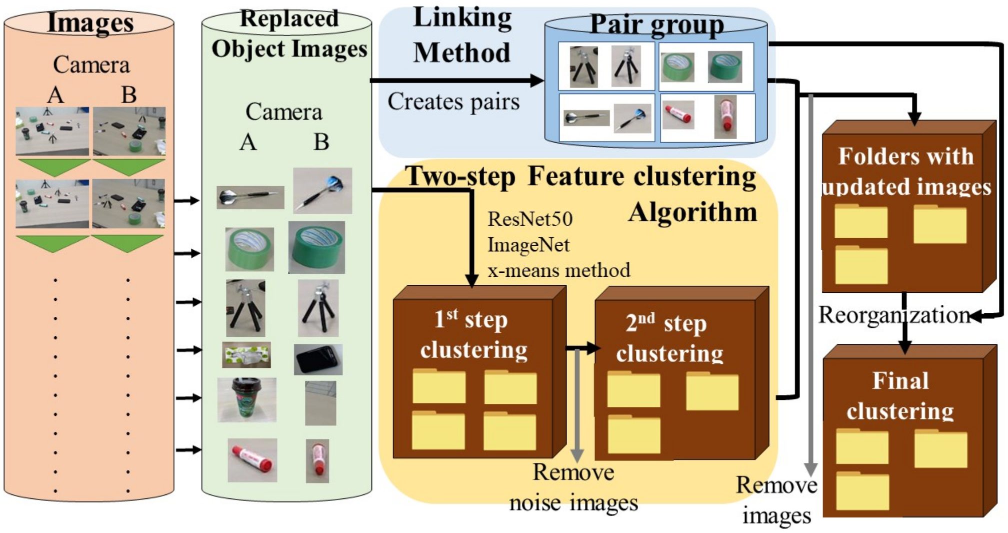

Fig. 6. Combine the results of the linking method with the two-step feature clustering algorithm to update the images; then, reorganize the clustering.

Then, the averages of the distance values are calculated for both Cluster_P and Cluster_Q. If the average of distances between Image_p and the other images in Cluster_P is larger than that of Cluster_Q, Image_p is moved to Cluster_Q.

III. Experiment

A. Aim

In this section, we detail an experiment conducted to compare the accuracy of clustering between the combination of LM with TFA and TFA alone.

B. Method

Figure 7 shows ten objects on a table. The objects were a red pen, a green pen, a smartphone tripod, a box of tissues, a cup of coffee, a black smartphone, a box of darts, a dart, a plastic bag of replacement dart feathers, and gum tape. Two cameras were located so that the entire table could be photographed from two different directions. Even if the PSS is running, when there are people present, nothing will not be photographed.

During the experiment, one of the authors moved one of the objects on the table and then moved beyond the ranges of the cameras. After confirming that the PSS recognized the displaced object, he moved the next object on the table. This method was applied to the ten objects. He moved each object 10 times in two conditions, “LM with TFA” and “TFA.”

This process was repeated twice on two different tables in two different rooms, Rooms C and D. In Room C, the ten objects were on a desk, as shown in Figure 8. This is a dimly lit space because three displays are lined up, and the fluorescent lamp is not directly overhead. In Room D, a large table was placed in the center of the room, with fluorescent lights directly above it. There was nothing around it to block the light, as shown in Figure 9.

The PSS created clusters in both conditions. The accuracy of the clustering is indicated in precision values, recall values, and F-measures. All displaced object images showing one of the ten objects are regarded as an “actual positive.” The cluster in which the most images is included is considered to be a correct cluster, and the displaced object images of the correct cluster are regarded as a “predicted positive.” In the predicted positive images, the actual positive images are considered to be true positives (TP), and the others are considered false positives (FP). In the actual positive images, the images that are not in the correct cluster are considered false negatives (FN).

The recall is calculated using the following formula.

$$Recall = \frac{TP}{TP + FN} \tag{1}$$

The precision is calculated using the following formula.

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

The F-measure is calculated using the following formula. The F-measure represents the harmonic mean of precision and recall.

$$F - measure = \frac{2Precision * Recall}{Precision + Recall} \tag{3}$$

C. Results

1) Comparison of the results of TFA and LM with TFA conditions: The PSS created 11 clusters in both conditions. Each displaced object image was grouped into one to five clusters, depending on the type of displaced object. For example, all images of the smartphone tripod were grouped into Cluster_3

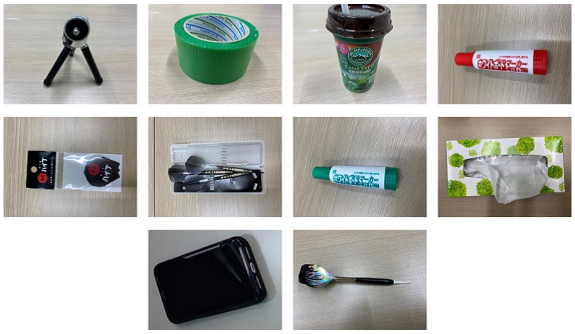

Fig. 7. Ten kinds of objects for the experiment.

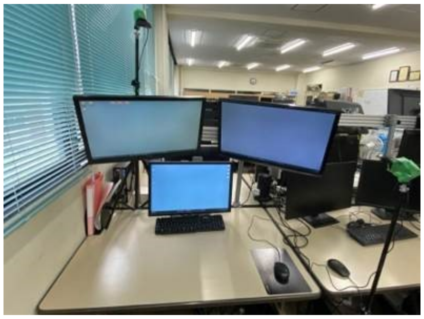

Fig. 8. Desk in Room C.

in the LM with TFA condition. The images of the green pen were grouped into five kinds of clusters in the TFA condition.

Table I shows the recall values, precision values, and F-measures to compare the results of the two conditions in both Room C and D. For eight out of ten objects, the recall values were higher in the LM with TFA condition than in the TFA alone condition. In particular, the recall value of the gum tape under the LM with TFA condition was improved by 17%, compared to the TFA alone condition, and it became even closer to 100%.

For all objects, the F-measures were higher in the LM with TFA condition than in the TFA alone condition. However, the precision values of the LM with TFA condition were almost the same as that of the TFA alone condition. The displaced object images, the smartphone, the box of darts, and the dart remained around 30%.

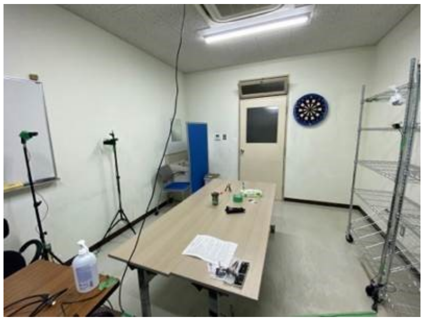

Fig. 9. Table in Room D.

2) Comparison of the results of Rooms C and D: Table II shows the precision values, recall values, and F-measures to compare the results of Rooms C and D. The number of images is less than Table I because of the results for each room. Therefore, the number of clusters was different, and these evaluated values between Table I and II are different. The F-measures of all displaced object images in Room D were higher than that of Room C.

D. Discussion

The results showed that LM with TFA improved the clustering over TFA alone. As an example of improvement, the images of the gum tape were grouped into four clusters in the TFA alone condition because there were some images that missed a part of the gum tape, as shown in Figure 10. However, in the LM with TFA condition, most images of gum tape could be grouped into one cluster.

The F-measures for the red and green pens were not high, even in the LM with TFA condition. Since the shapes of these pens are similar, it was difficult to group them into one cluster using these algorithms. An algorithm using color features should be applied to such objects [5].

For the smartphone tripod and the cup of coffee, the precision values in the LM with TFA condition were lower than in the TFA alone condition. In the LM process, contrary to our expectations, the images of the smartphone tripod and the cup of coffee were paired with noisy images. Something similar to these objects was photographed as noisy images.

There were differences in accuracy depending on the room. The table in Room D was brighter than the desk in Room C. It can be suggested that the brighter space led to higher accuracy in detecting the displaced objects.

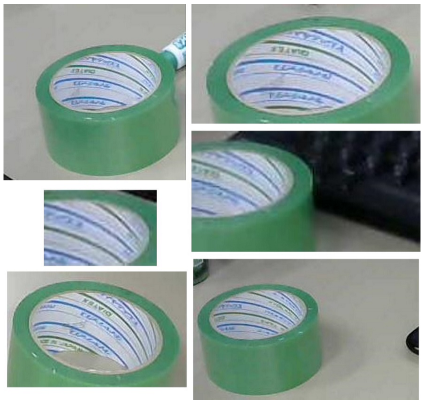

Fig. 10. Images that missed a part of the gum tape.

IV. CONCLUSION

In a Physical Search System (PSS), the features of images were quantified, and the images were grouped into clusters using the x-means method (TFA: Two-step Feature clustering Algorithm). However, some noisy images remained, and the

TABLE I
ACCURACY OF CLUSTERING IN THE CONDITIONS TFA ALONE AND LM WITH TFA.

	Recall		**Precision**		**F-measure**	
Displaced objects	*TFA*	*LM with TFA*	*TA*	*LM with TFA*	*TFA*	*LM with TFA*
Red pen	0.50	0.53	0.63	0.64	0.56	0.58
Green pen	0.38	0.43	0.47	0.43	0.42	0.43
Tripod	0.93	1.00	1.00	0.95	0.96	0.98
Box of tissues	0.49	0.63	1.00	1.00	0.65	0.77
Cup of coffee	0.91	1.00	1.00	0.94	0.95	0.97
Smartphone	0.42	0.58	0.23	0.29	0.30	0.39
Box of darts	0.44	0.69	0.30	0.30	0.36	0.42
Dart	0.75	0.75	0.24	0.27	0.36	0.40
Plastic bag	0.66	0.63	0.48	0.57	0.56	0.60
Gum tape	0.76	0.93	1.00	1.00	0.86	0.96
Average	0.62	0.72	0.64	0.64	0.60	0.65

TABLE II
ACCURACY OF CLUSTERING IN THE CONDITIONS OF ROOMS C AND D.

	Recall				**Precision**				**F-measure**			
	TFA		*LM with TFA*		*TFA*		*LM with TFA*		*TFA*		*LM with TFA*	
Displaced objects	C	D	C	D	C	D	C	D	C	D	C	D
Red pen	1.00	0.88	1.00	0.75	0.43	0.45	0.43	0.55	0.60	0.60	0.60	0.63
Green pen	0.87	0.81	0.95	0.63	0.48	0.81	0.50	0.45	0.62	0.55	0.66	0.53
Tripod	0.76	1.00	0.90	1.00	0.90	0.90	1.00	0.90	0.83	0.95	0.95	0.95
Box of tissues	0.50	1.00	0.50	1.00	0.50	0.80	1.00	1.00	0.50	0.89	0.67	1.00
Cup of coffee	0.93	1.00	0.71	1.00	0.87	0.86	1.00	1.00	0.90	0.92	0.83	1.00
Smartphone	0.64	0.81	0.76	0.63	0.20	0.39	1.00	0.50	0.31	0.53	0.86	0.56
Box of darts	0.53	0.81	0.47	0.63	0.22	0.39	0.57	0.48	0.31	0.53	0.52	0.63
Dart	0.69	0.56	1.00	1.00	0.09	0.90	0.12	0.95	0.17	0.69	0.22	0.97
Plastic bag	1.00	0.80	0.95	0.93	0.82	0.92	1.00	0.93	0.90	0.86	0.98	0.93
Gum tape	0.78	1.00	0.78	1.00	1.00	0.95	1.00	1.00	0.88	0.97	0.88	1.00
Average	0.77	0.87	0.80	0.86	0.55	0.74	0.76	0.78	0.60	0.75	0.72	0.82

accuracy of the clustering was not so high. Therefore, in this paper, we propose a Linking Method (LM). The images photographed simultaneously by two cameras before displaced objects were cropped out were paired according to the images' similarity. This method cannot allow most noisy images to be paired with other images. Finally, among the images in the cluster created by TFA alone, only the images paired using LM were left. As a result, the clustering accuracy was improved.

In the future, we will continue to make improvements to achieve greater accuracy.

ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI Grant Number 20H04470.

REFERENCES

[1] R. Baheti and H. Gill, "Cyber-physical systems," The impact of control technology, IEEE Control Systems Society, vol. 12, no. 1, pp. 161–166, 2011.

[2] A. Ahmad, A. Paul, M. M. Rathore, and H. Chang, "Smart cyber society: Integration of capillary devices with high usability based on Cyber–Physical System," Future Generation Computer Systems, Elsevier, vol. 56, pp. 493–503, 2016.

[3] R. Hamasaki and K. Nakayama, "A deep learning system that learns a discriminative model autonomously using difference images," Proceedings of the Genetic and Evolutionary Computation Conference Companion, ACM, pp. 1683–1685, 2019.

[4] K. Nakamura, R. Hamasaki, C. Oshima, and K. Nakayama, "Optimizing Combinations of Teaching Image Data for Detecting Objects in Images," Lecture Notes in Computer Science, Springer, vol. 12185, pp. 491-–505, 2020.

[5] S. Kajihara, M. Okazaki, K. Kawabata, H. Furukawa, C. Oshima, O. Fukuda, and K. Nakayama, "Proposal and verification of a physical search system that does not require pre-learning data and sensors other than cameras," IPSJ Transactions on digital practices, vol. 3, no. 2, pp. 76–92, 2022. (in Japanese)

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," In Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, p. 770–778.

[7] "ResNet50," https://jp.mathworks.com/help/deeplearning/ref/resnet50.html;jsessionid=c2d1fcfb1eb58ff18ab9a8beff0c. [retrieved: 08, 2022]

[8] "ImageNet," https://www.image-net.org/. [retrieved: 08, 2022]

[9] B. Li and D. Lima, "Facial expression recognition via ResNet-50," International Journal of Cognitive Computing in Engineering, vol. 2, pp. 57-64, 2021.

[10] S. Bhattacharyya, "Understand and Implement ResNet-50 with TensorFlow 2.0," Towards Data Science, https://towardsdatascience.com/understand-and-implement-resnet-50-with-tensorflow-2-0-1190b9b52691 [retrieved: 08, 2022]

[11] C. Chen, W. Zhu, J. Steibel, J. Siegford, J. Han, and T. Norton, "Classification of drinking and drinker-playing in pigs by a video-based deep learning method," Biosystems Engineering, vol. 196, pp. 1-14, 2020.

[12] D. Pelleg and A. W. Moore, "X-means: Extending k-means with efficient estimation of the number of clusters," Proceedings of ICML 2000, vol. 1, pp. 727–734, 2000.

[13] P. S. Bradley and U. M. Fayyad, "Refining initial points for k-means clustering," Proceedings of ICML 1998, vol. 98, pp. 91–99, 1998.

[14] "imgsim," https://github.com/Nr90/imgsim. [retrieved: 08, 2022]