ICIC Express Letters
Part B: Applications
Volume **13**, Number **12**, December **2022**

 ISSN 2185-2766
pp. **1277–1284**

ACCURACY IMPROVEMENT OF THERMOGRAPHY IMAGE RECOGNITION FOR CONTACTLESS INPUT DEVICE

Wataru Chidiwa[1], Noritaka Shigei[1,*] and Hiromi Miyajima[2]

[1]Graduate School of Science and Engineering
[2]Professor Emeritus of Kagoshima University
Kagoshima University
1-21-40 Korimoto, Kagoshima City, Kagoshima 890-0065, Japan
{ k8417159; k2356323 }@kadai.jp; *Corresponding author: shigei@ibe.kagoshima-u.ac.jp

Received April 2022; accepted June 2022

Abstract. *This work develops contactless input interfaces to identify the number of presented fingers from infrared thermography (IRT) images by using Convolutional Neural Network (CNN) and considers improving its recognition accuracy. The advantages of using IRT images over visible camera images are 1) not affected by lighting conditions, 2) better privacy, and 3) less computation due to the small number of pixels and channels. However, compared to the case of using camera images, the identification accuracy of input devices using IRT images is low because the IRT images are of low resolution. In order to improve the recognition accuracy, we investigate several types of CNN models and propose a method for removing objects such as fluorescent lights in the background of the target. Numerical experiments show the effectiveness of the proposed method and the effective CNN model among investigated ones.*
Keywords: Contactless input device, Infrared thermography image, Convolutional neural network, Image recognition

1. **Introduction.** In recent years, due to the prevalence of COVID-19, it is necessary to avoid contact with objects used by multiple people. For this reason, contactless human-machine interaction such as contactless person identification and contactless input devices is attracting attention. Contactless person identifications with artificial intelligence (AI) have progressed [1], and they become widespread in our lives. In general, the identification methods require accurate and detailed information on the identified person, and they use a high-definition image sensor such as camera. On the other hand, contactless input devices are not considered to be as widespread as for person identification. The reasons for this are considered to be ease of use, privacy, and cost issues. Today, information terminals used by many people in stores, public facilities, and workplaces are increasing, and the introduction of easy-to-use contactless interfaces in these areas will contribute not only to the prevention of infectious diseases but also to the elimination of the digital divide.

Studies, developments and manufacturing have been conducted on contactless input devices and related recognition techniques. The devices and techniques utilize voice and images as input signals. Voice-based contactless input is already widely used in smart speakers. However, this is not suitable for use in public places due to privacy and security issues. Input signals for images include depth images from depth sensors, and thermographic images from infrared (IR) sensors, in addition to visible light camera images (referred to as camera images) [2]. Camera images with/without depth information are probably the most common and widely studied [3, 4, 5]. Thanks to the low cost of cameras, it is now easy to obtain high-definition images that lead to high recognition accuracy, but the hardware to process them is relatively expensive. Depth cameras can improve

DOI: 10.24507/icicelb.13.12.1277

accuracy, but the devices and processing hardware are expensive. Furthermore, camera images would be difficult to use in areas where privacy considerations are necessary.

Infrared thermography (IRT) images have also been utilized as input signals for the device. Studies have been done on the recognition of hand gestures using images from a commercial device Leap Motion, which has two IR cameras and whose resolution is 640×240 [6, 7, 8]. The images from Leap Motion are clear comparable to visible camera images and the previous researches achieved more than 99% accuracy for 10 class classification. While this is an attractive device in terms of accuracy, it is more expensive than a regular camera and would require the same hardware processing power as usual camera images. Studies have been conducted on contactless input devices that use low-resolution information from infrared sensors [9, 10, 11]. Effective arrangement of low cost infrared distance sensors to recognize gestures of hand movement has been studied [9]. Although this device is low-cost, it is difficult to identify stationary gestures such as the number of fingers. In [10], the authors have developed a contactless human-machine interface that recognizes gestures based on the color obtained from infrared light emitted from a finger and demonstrated that real-time control of robotic vehicles is possible using the interface. While this interface has the advantage of not being influenced by the background, it is not considered easy for first-time users to use it immediately. In [11], an in-vehicle device control system by hand posture recognition with movement detection using low-cost IR array sensor of 32×24 resolution has been developed. In a 10-class classification, this work achieves an accuracy of about 95% by using a Convolutional Neural Network (CNN), pre-processing of background removal and static posture detection. In summary, the advantages of using IRT images over visible camera images are 1) not affected by lighting conditions, 2) better privacy, and 3) less computation due to the small number of pixels and channels. However, compared to the case of using camera images, the identification accuracy of input devices using IRT images is low because the IRT images are of low resolution.

In this research, we develop contactless input interfaces to identify the number of presented fingers from IRT images by using Convolutional Neural Network (CNN) and consider improving its recognition accuracy. In order to improve the recognition accuracy, we investigate several types of convolutional neural network models and propose a method for removing objects such as fluorescent lights in the background of the target. Numerical experiments show the effectiveness of the proposed method and the effective CNN model among investigated ones.

2. **Contactless Input Device.** The input device developed in this research can input six characters by presenting zero to five fingers, including a thumb, as shown in Figure 1. The system consists of an IR array sensor, a microcontroller and a host computer such as Single Board Computer (SBC), as shown in Figure 2(a). The temperature data acquired by the sensor is sent to SBC via the microcontroller ESP32. The SBC converts the data into a grayscale image, and the identification is performed by inputting the image into

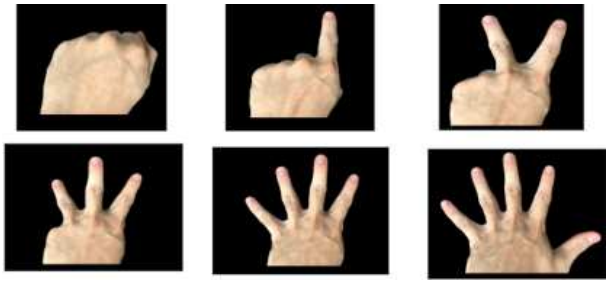
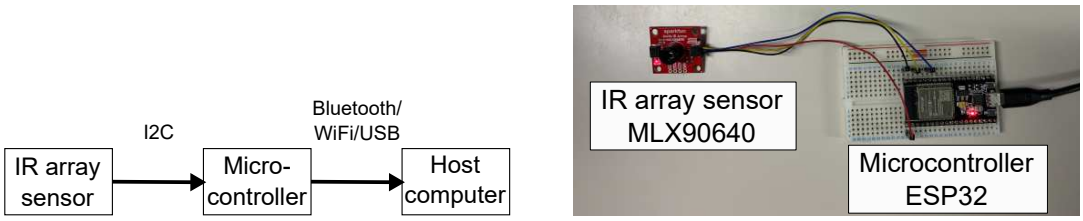


FIGURE 1. Six hand gestures used for inputting six types of characters

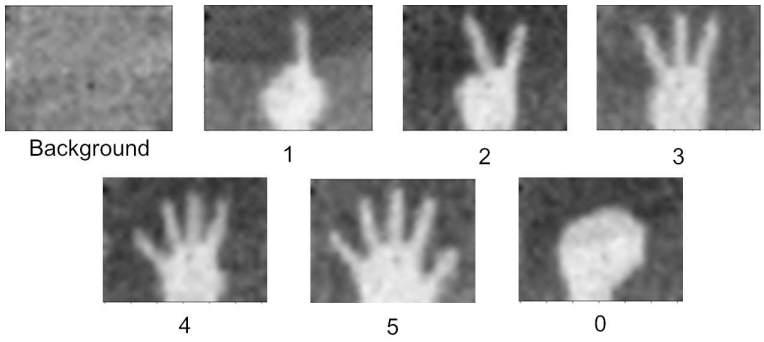


(a) System configuration (b) Appearance

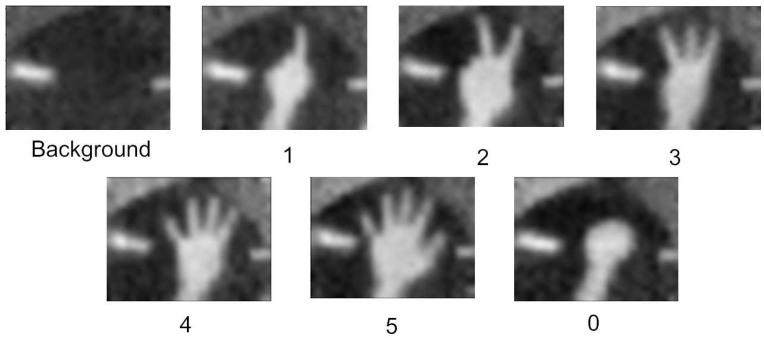
FIGURE 2. Contactless input device

CNN. Figure 2(b) shows the appearance of the input device. This input device is intended to be used with the sensor facing up and holding a hand.

2.1. **IR array sensor.** The used IR array sensor is Melexis MLX90640 [12], whose specification is as follows: view angle $55° \times 35°$ or $110° \times 75°$, output dimension 24×32, ADC resolution 18 bits, and measurement temperature range $-40°$C$\sim$80°C. The temperature data consisting of 768 elements is acquired at the frame rate 2$\sim$3 fps. Sample images captured by each sensor are shown in Figure 3.



(a) For $55° \times 35°$ sensor

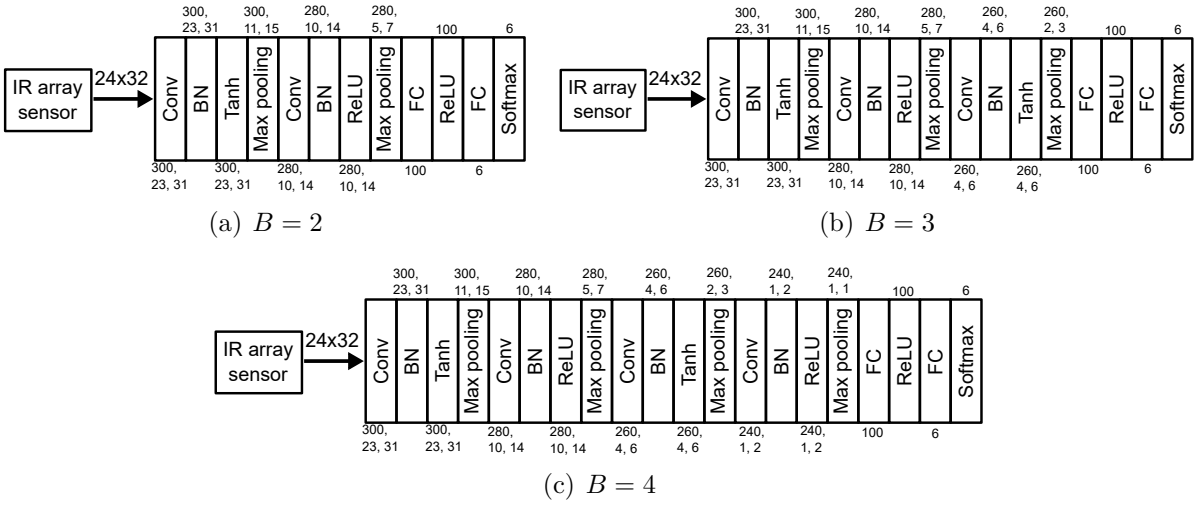


(b) For $110° \times 75°$ sensor

FIGURE 3. Images captured by sensors

For the narrow view angle version of $55° \times 35°$, its advantage is that it is unlikely to show a background that would interfere with identification, while its disadvantage is that it requires a relatively long distance (about 30 cm) from the object to be identified. For the wide view angle version of $110° \times 75°$, its advantage is that it can shorten the distance (about 15 cm) to the identification target, while its disadvantage is that it tends to reflect backgrounds that interfere with the identification.

2.2. **Convolutional neural network.** A convolutional neural network is used for the classifier. Simple CNN models are considered so that they can be run on non-powerful, low-cost hardware such as SBCs. Figure 4 shows three used CNN models for $B = 2$, $B = 3$ and $B = 4$ blocks, where Conv, BN, and FC represent convolution, batch normalization, and fully connected layers, respectively. In these models, consecutive B blocks consisting of Conv, BN, activation function (ReLU or Tanh), and Max pooling layers are connected to the first FC layer. In Figure 4, the numbers with each block are the number of outputs and the dimension of each output.



(a) $B = 2$

(b) $B = 3$

(c) $B = 4$

FIGURE 4. Convolutional neural network models

Each output p_c ($c \in \{0, 1, 2, 3, 4, 5\}$) of Softmax corresponds to the probability of input being class c, where c corresponds to the number of fingers as input. The CNN model determines the input c^* by the following equation.

$$c^* = \arg\max_{c' \in \{0,\ldots,5\}} p_{c'} \tag{1}$$

3. **Background Removal Technique.** In thermography images captured by IR sensors, the objects with high temperatures other than the objects to be identified may appear as noise. In Figure 3(b), two fluorescent lights are in the background. These background objects degrade the accuracy of the identification. In order to solve this problem, we propose a background removal technique for removing obtrusive objects in the background, and apply it to the image before being input to the CNN.

The proposed method utilizes background-only images with obtrusive objects to delete the obtrusive ones. Basically, background removal would be sufficient to generate an image differing from the background image. However, one of the difficulties for a lower-resolution thermographic image than a usual camera image is that thermographic images change rapidly between successive frames even though there is no actual change. For effective background removal, a suitable background image should be determined to take the difference. The proposed method determines the background image to be used based on the average absolute error of pixels against the target image. The detailed algorithm is shown below.

Background Removal Algorithm

Input:

Target image: $\boldsymbol{x} = (x_1, \ldots, x_{IJ})$, where $I \times J$ is the image size.

Set of background images: $B = \left\{ \boldsymbol{b}^{(n)} = \left(b_1^{(n)}, \ldots, b_{IJ}^{(n)} \right) \middle| \, n \in 1, \ldots, N \right\}$, where N is the number of background images.

Output:

Background removal image: $\hat{\boldsymbol{x}} = (\hat{x}_1, \ldots, \hat{x}_{IJ})$

Step 1: For each $n \in \{1, 2, \ldots, N\}$, calculate the average absolute difference $\bar{d}_n$ of pixels between the n-th background image $\boldsymbol{b}^{(n)}$ and the target image $\boldsymbol{x}$ as follows.

$$\bar{d}_n = \frac{\sum_{i=1}^{IJ} \left| x_i - b_i^{(n)} \right|}{I \times J} \tag{2}$$

Step 2: Determine the n^*-th background to use for background removal.

$$n^* = \underset{n \in \{1,\ldots,N\}}{\arg\min}\, \bar{d}_n \tag{3}$$

Step 3: Generate a background removal image $\hat{\boldsymbol{x}}$ by taking the difference between the target image $\boldsymbol{x}$ and the selected background image $\boldsymbol{b}^{(n^*)}$ as follows.

$$\hat{x}_i = \left| x_i - b_i^{(n^*)} \right|, \ i \in \{1, 2, \ldots, IJ\} \tag{4}$$

□

Figure 5 shows a background removed image example. It can be observed that objects in the background are clearly removed. The reason why Equation (3) is effective could be because the area of the object of recognition is small relative to the rest of the area, so the smaller the difference is on average, the more the background is matched.

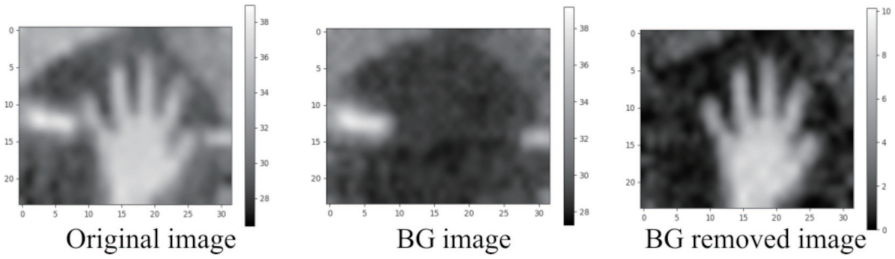


FIGURE 5. Background removal image example

4. **Experimental Evaluation.** This experiment examines the effective combination of sensors and CNN models and the effectiveness of the proposed background removal method.

The training data consists of 6,000 images collected over two days, containing 1,000 images for each finger class from 0 to 5. For background removal during training, $N =$ 1,000 background-only images are used. When collecting data, the sensor was placed, so that background objects such as fluorescent lights do not overlap with the presented subject. For test data, the two types of test data, Test data 1 and Test data 2, are used. For Test data 1, as for the training data, the presented subject is not allowed to overlap with the background object, and for Test data 2, the presented subject is shown as large as possible so that it is allowed to overlap with background objects. The purpose of using Test data 2 is to verify the robustness against the size of the object and background objects. The examples of training data and test data are shown in Figure 6. Both Test data 1 and 2 consist of 1,800 images collected over three days, which are different from the two days for training, containing 300 images for each of the six classes.

The experiment is performed by using an open source machine learning framework, PyTorch. The CNN models are trained using the following parameters: batch size 60, the number of epochs 100 and Adam with step size 0.001. A total of 12 combinations of sensors $55° \times 35°$ or $110° \times 75°$, CNN models for $B = 2$, 3 or 4, with or without background

(a) Training data (b) Test data 1 (c) Test data 2

FIGURE 6. Example of training data and test data

removal are evaluated by averaging ten trials of the following accuracy.

$$\text{Accuracy} = \frac{\#\text{ of correct identifications}}{\#\text{ of total identifications}} \tag{5}$$

The evaluation results are shown in Tables 1 and 2. Table 1 shows the accuracy for each sensor, each CNN model and with/without background removal. "Average" is the average of the accuracy of Test data 1 and 2. Each bold number indicates the best accuracy for each sensor and each of "Test data 1", "Test data 2" and "Average". From the result, the following tendencies are observed.

(1) The accuracy of Test data 2 is lower than that of Test data 1. Their differences are 0.143 points for $55° \times 35°$ and 0.428 points for $110° \times 75°$.
(2) The sensor of $55° \times 35°$ achieves better accuracy than the one of $110° \times 75°$. In particular, for Test data 2, the difference is 0.323 points in the best accuracy.
(3) The best CNN model depends on the combination of sensor type, test data type and with or without background removal. In particular, when using background removal, according to the average accuracy, the best model is $B = 2$ for $55° \times 35°$ and $B = 4$ for $110° \times 75°$.
(4) Although background removal is not effective in some cases for Test data 1, for Test data 2 and average accuracy, background removal improves accuracy in all cases. In particular, when the best accuracy is achieved, the improvements for $55° \times 35°$ are 0.171 points for Test data 2 and 0.094 points for average accuracy, and the improvements for $110° \times 75°$ are 0.192 points for Test data 2 and 0.139 points for average accuracy.

From the tendencies (1) and (2), it can be said that the sensor of $55° \times 35°$ is more suited to the contactless input device than the one of $110° \times 75°$ in terms of accuracy.

TABLE 1. Accuracy for each sensor, each CNN model and with/without background removal

Sensor	B	BG removal	Test data 1	Test data 2	Average
$55° \times 35°$	2	No	0.972	0.676	0.824
		Yes	0.989	**0.847**	**0.918**
	3	No	0.987	0.724	0.856
		Yes	0.988	0.833	0.911
	4	No	**0.990**	0.711	0.850
		Yes	0.984	0.791	0.888
$110° \times 75°$	2	No	0.838	0.321	0.579
		Yes	0.921	0.329	0.625
	3	No	0.947	0.367	0.657
		Yes	0.902	0.522	0.712
	4	No	0.865	0.332	0.599
		Yes	**0.952**	**0.524**	**0.738**

TABLE 2. Confusion matrix for a trial with $55° \times 35°$, background removal and $B = 2$

(a) For Test data 1

		Estimated					
		0	1	2	3	4	5
Actual	0	300	0	0	0	0	0
	1	0	300	0	0	0	0
	2	0	0	298	2	0	0
	3	0	0	0	299	1	0
	4	0	0	0	1	299	0
	5	0	0	0	0	0	300

(b) For Test data 2

		Estimated					
		0	1	2	3	4	5
Actual	0	300	0	0	0	0	0
	1	0	294	6	0	0	0
	2	0	17	269	14	0	0
	3	0	0	0	299	1	0
	4	0	20	0	45	218	17
	5	0	2	0	0	102	196

TABLE 3. The numbers of correct and incorrect patterns with and without background removal for a trial of Test data 2: $55° \times 35°$, $B = 2$ for with BG removal and $B = 3$ for without BG removal

		With BG removal	
		True	False
Without BG removal	True	1267	105
	False	309	119

The tendency (3) also suggests that the sensor of $55° \times 35°$ would benefit from low-cost implementation since a smaller CNN model would yield better accuracy. Further, the tendency (4) demonstrates that the proposed background removal method is effective.

Table 2 shows confusion matrices obtained by one trial of the best sensor and the best CNN model for Test data 1 and 2. For Test data 1, classification is very few and limited to a close number of classes. For Test data 2, misclassifications are more frequent in the 5-finger, 4-finger, and 2-finger classes, in that order. In particular, about 30% of the 5-finger classes are misclassified as 4-finger classes, and about 15% of the 4-finger classes are misclassified as 3-finger classes. In the images of Test data 2, the background and the identification target are overlapped, and the target appears larger than in training data. Table 3 shows the numbers of correct and incorrect patterns with and without background removal for a trial of Test data 2. According to the result, the background removal reduces the number of misclassified patterns by about half. The remaining misclassification patterns are expected to be mainly due to differences in the scale of the target objects.

5. **Conclusions.** In this paper, we developed contactless input interfaces to identify the number of presented fingers from IRT images by using CNN and improved its recognition accuracy. It is shown that the sensor of $55° \times 35°$ is more suited to the contactless input device than the one of $110° \times 75°$ in terms of accuracy and the proposed background removal method is effective. One of the future works is to improve the accuracy for the different scales of the target objects.

Acknowledgment. This work is supported by JSPS KAKENHI Grant Number 20K11994.

REFERENCES

[1] S. Matsuda and H. Yoshimura, Personal identification with artificial intelligence under COVID-19 crisis: A scoping review, *Systematic Reviews*, vol.11, DOI: 10.1186/s13643-021-01879-z, 2022.

[2] M. Oudah, A. Al-Naji and J. Chahl, Hand gesture recognition based on computer vision: A review of techniques, *Journal of Imaging*, vol.6, DOI: 10.3390/jimaging6080073, 2020.

[3] P. Molchanov, X. Yang, S. Gupta, K. Kim, S. Tyree and J. Kautz, Online detection and classification of dynamic hand gestures with recurrent 3D convolutional neural networks, *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.4207-4215, 2016.

[4] K. Akehi, S. Matuno, N. Itakura, T. Mizuno and K. Mito, Study of non-contact eye glance input interface with video camera, *IEEJ Trans. Electronics, Information and Systems*, vol.137, no.4, pp.628-633, 2017.

[5] M. S. Kabisha, K. A. Rahim, M. Khaliluzzaman and S. I. Khan, Face and hand gesture recognition based person identification system using convolutional neural network, *International Journal of Intelligent Systems and Applications in Engineering*, vol.10, no.1, pp.105-115, 2022.

[6] T. Mantecon, C. R. del-Blanco, F. Jaureguizar and N. Garcia, Hand gesture recognition using infrared imagery provided by leap motion controller, *Int. Conf. on Advanced Concepts for Intelligent Vision Systems (ACIVS2016)*, pp.47-57, 2016.

[7] T. R. Gadekallu, G. Srivastava, M. Liyanage, M. Iyapparaja, C. L. Chowdhary, S. Koppu and P. K. R. Maddikunta, Hand gesture recognition based on a Harris Hawks optimized convolution neural network, *Computers and Electrical Engineering*, vol.100, DOI: 10.1016/j.compeleceng.2022.107836, 2022.

[8] A. Sen, T. K. Mishra and R. Dash, A novel hand gesture detection and recognition system based on ensemble-based convolutional neural network, *Multimedia Tools and Applications*, 2022.

[9] C. Xia, A. Saito and Y. Sugiura, Using the virtual data-driven measurement to support the prototyping of hand gesture recognition interface with distance sensor, *Sensors and Actuators A: Physical*, vol.338, DOI: 10.1016/j.sna.2022.113463, 2022.

[10] S. An, H. Zhu, C. Guo, B. Fu, C. Song, P. Tao, W. Shang and T. Deng, Noncontact human-machine interaction based on hand-responsive infrared structural color, *Nature Communications*, vol.13, DOI: 10.1038/s41467-022-29197-5, 2022.

[11] S. Tateno, Y. Zhu and F. Meng, In-vehicle device control system by hand posture recognition with movement detection using infrared array sensor, *SICE Journal of Control, Measurement, and System Integration*, vol.13, no.3, pp.148-156, 2020.

[12] Melexis, *MLX90640 32x24 IR Array Datasheet*, Rev.12, 2019.