Deep-Learning-Assisted Single-Pixel Imaging for Gesture Recognition Considering Privacy

Naoya Mukojima[1], Masaki Yasugi[1,2], Yasuhiro Mizutani[3], Takeshi Yasui[4], Hirotsugu Yamamoto[1,2]

hirotsugu@yamamotolab.science
[1]Utsunomiya Univ., 7-1-2 Yoto, Utsunomiya, Tochigi 321-0904, Japan
[2]JST, ACCEL, 7-1-2 Yoto, Utsunomiya, Tochigi 321-0904, Japan
[3]Osaka Univ., 2-1 Yamadaoka, Suita, Osaka 565-0871, Japan
[4]Tokushima Univ., 2-1 Minamijosanjimacho, Tokushima, Tokushima 770-8506, Japan
Keywords: Single-Pixel Imaging, Deep Learning, Gesture Recognition

ABSTRACT

We report difference of structural similarity of restored images in terms of different irradiation methods for single-pixel imaging. Images of hand shadows taken with different irradiation methods are processed by single-pixel imaging and restores by U-Net for output. The proposed method is used for considering privacy gesture recognition.

1 INTRODUCTION

In recent years, the number of pixels of images captured by common cameras has been increasing, and high pixel images are required. The cost of manufacturing image sensor and the number of image measurements are also problems. single-pixel imaging is one of the solutions to these problems [1]. In this imaging method, the illumination is repeatedly modulated with respect to the subject and the image information is acquired with a single-pixel detector. However, in order to obtain data similar to images taken with common cameras, a large number of measurements are required to obtain images with a large number of pixels. In order to solve this problem, there is an increasing number of researches that use deep learning to restore images. By using deep learning, it is possible to restored images from a small amount of measured data by extracting feature of measured signal patterns. By using this method, it is possible to capture images as a shadowgraph without photographing the privacy of such as the face, and thus the purpose of this method is to perform gesture recognition while protecting the privacy of the person. We consider using a high frame rate LED display to modulate the illumination in objective gesture recognition [2]. The application to a large-scale LED display is also expected to protect the privacy of a large number of viewers and count the number of people in the audience [3].

In this paper, we compare illumination modulation from behind and in front of the gesture to protect privacy and test the effectiveness of using single-pixel imaging to capture the shadow of the gesture without capturing the person's privacy.

2 PRINCIPLE

2.1 SINGLE PIXEL IMAGING

The basic principle of single-pixel imaging used in this paper is shown in Fig.1. A randomly generated mask is used to modulate the illumination for the subject at an arbitrary number of times, and the measurement is made with a single-pixel detector. The information obtained from the measurements is converted into a matrix and calculated by an intensity correlation function. The calculation gives the restored image as a result of a floating-point operation that takes a range from -1 to 1. In order to obtain a clear restored image, a large number of measurements are required, and the restored image with a small number of measurements contains a lot of noise. The intensity correlation function can be expressed as:

$$\begin{aligned} G(x,y,n) &= \langle \Delta I(x,y,n)\Delta A(n)\rangle \\ &= \langle [I(x,y,n) - \langle I(x,y,n)\rangle][A(n) - \langle \Delta A(n)\rangle]\rangle \\ &= \langle I(x,y,n)A(n)\rangle - \langle I(x,y,n)\rangle\langle A(n)\rangle \end{aligned} \quad (1)$$

where $\Delta I(x,y,n)$ is the deviation between the light intensity $I(x,y,n)$ and the mean $\langle I(x,y,n)\rangle$ of the nth randomly patterned mask in the coordinates (x,y). $\Delta A(n)$ is the deviation of the average value of the light intensity $A(n)$ and its mean $\langle \Delta A(n)\rangle$ measured by the nth single-pixel detector. $A(n)$ can also be given as:

$$A(n) = \iint T(x,y)I(x,y,n)dxdy \quad (2)$$

by using the transmission function $\mathrm{T(x.y)}$ [4,5]. For example, reconstructed images with 100, 250, and 500 modulation counts of the illumination are shown in Fig.2.

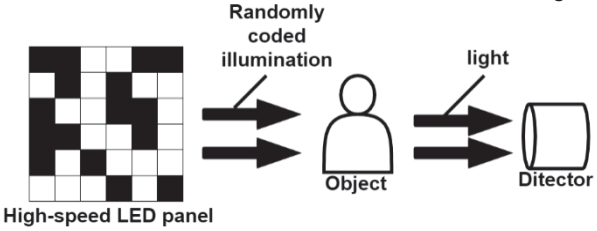


Fig.1 Schematic diagram showing a single-pixel imaging by use of mask patterns.

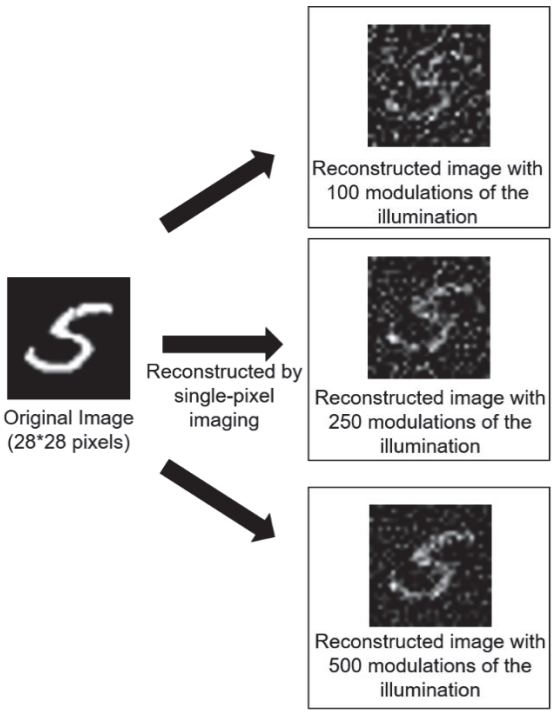


Fig.2 Reconstructed images for single-pixel imaging with illumination modulated 100, 250, 500 times.

2.2 NETWORK MODEL FOR DEEP LEARNING

The network model for deep learning used in this research (U-net) is shown in Fig.3. At first, this model repeats convolution and MAX Pooling of the input data. After that, convolution and UN Pooling are repeated the same number of times. In addition, it uses the data used in the encoding process to decode the input data, so it is good at extracting features from the input data [6].

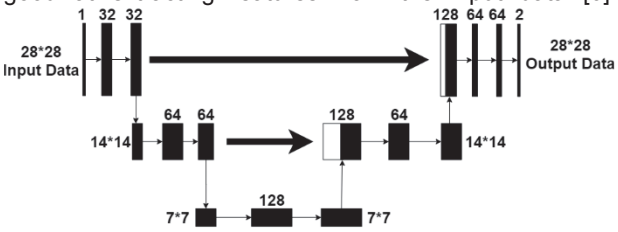


Fig.3 U-Net Structure.

3 TRAINING DATA SET FOR HAND GESTURE

We present a method for creating the training data used in this study for deep learning. Three different hand gestures were performed in front of a lighted display. We took movies of the gestures in various angles. Fig.4 shows the scene. Each frame was cropped as an image from the video data taken and resized to 28*28 pixel images. From these images, 9000 images were used in this study (PNG format). These images were binarized to create the images when the illumination was applied from behind. Furthermore, these images of the binarized images were negative-positive inverted and used as these when the illumination was irradiated from the front. These two types of data were modulated 250 and 500 times in single-pixel imaging to obtain 6800 training data and 1700 validation data and 500 test data, respectively. As the number of modulations at the time of reconstruction increases, the array format has better quality than the image format in deep learning. Therefore, each data generated is in array format, which is reconstructed by single-pixel imaging[7]. Fig. 5 shows an example of the data for training together with the original image. For evalution of restored image, we use SSIM which indicates the structural similarity of the restored images in this paper. The restored images with a range of SSIM values from 0 to 1 and a value close to 1 is more similar to the original images.

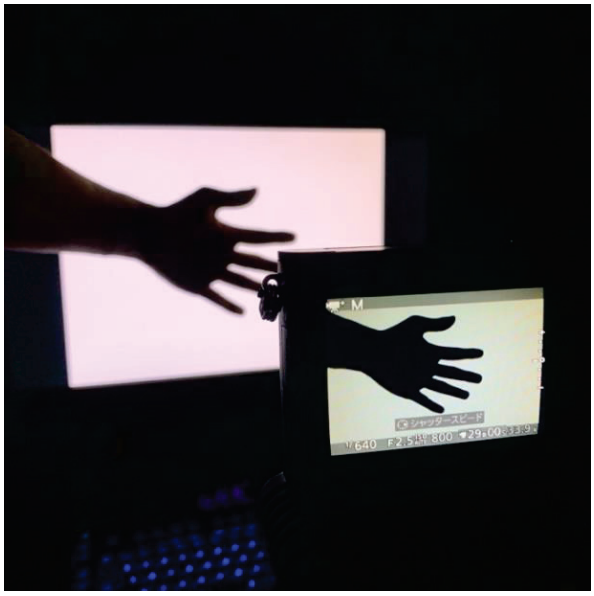


Fig.4 Scene of taking movie for training data.

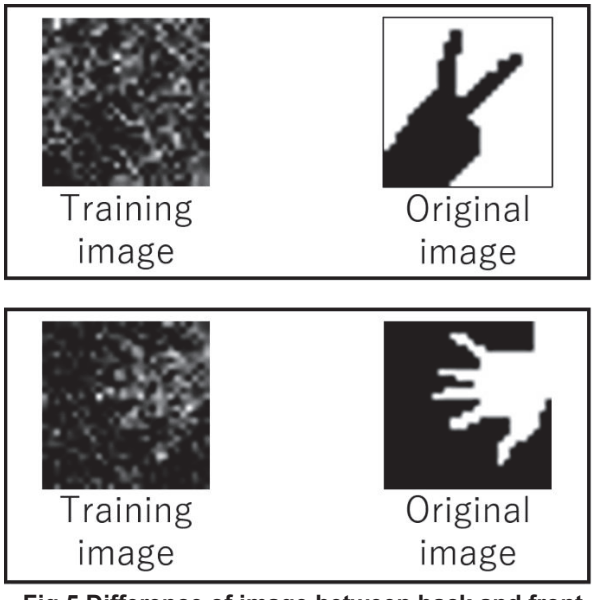


Fig.5 Difference of image between back and front illumination.

4 RESULTS

We used four types of training data to reconstruct the images on U-Net. An example of the images at each stage of processing is shown in Fig. 6. Fig. 6 shows three types of images for each calculation process. The image recovered by single-pixel imaging contains a lot of noise and can't be recovered, while the image recovered by U-Net is very close to the original image because the noise disappears. The values of SSIM for each modulation frequency of the illuminating light are shown in Table 1. Fig.7 show scatterplots of the calculation results when the front illumination is modulated 500 and, 250.Next, Fig.8 show scatterplots of the calculation results when the illumination is modulated 500 and, 250. From Table 1, it can be seen that the SSIM is closer to 1 for both results when the illumination is from behind than when the illumination is from in front. These results showed that the irradiation from behind the gesture was effective. In addition, Fig.9 and Fig.10 show the back and front illumination of the images restored by deep learning, respectively.

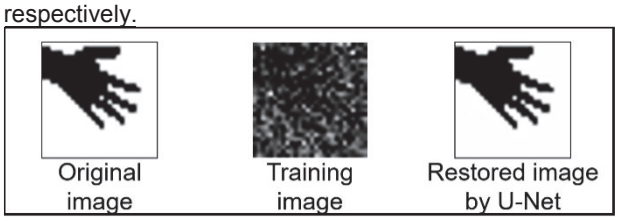
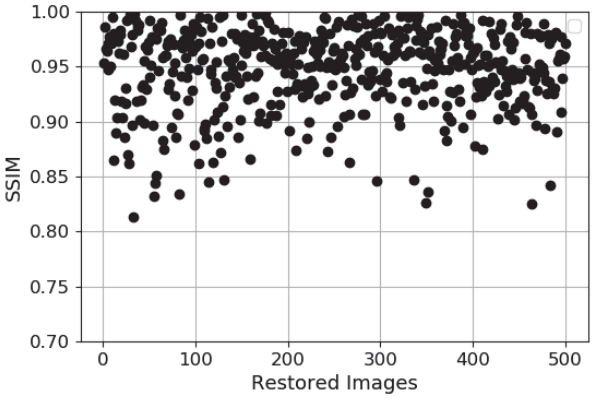


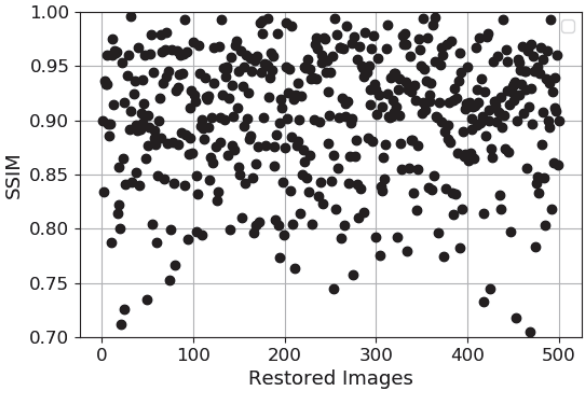
Fig.6 Images for each calculation of processing.

Table 1
The values of SSIM for each modulation frequency of the illuminating light.

Illumination pattern	SSIM	
	Modulation: 500	250
Front	0.947	0.900
Back	0.956	0.955

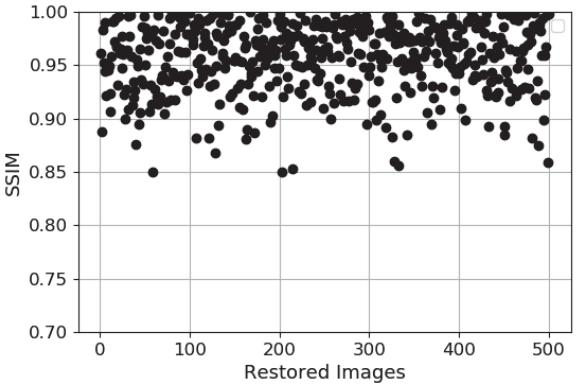


(a) A scatter diagram of results for 500 modulations and front illumination.

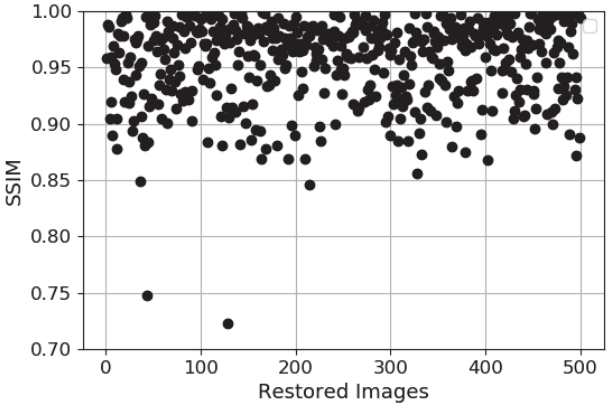


(b) A scatter diagram of results for 250 modulations and front illumination.

Fig.7 A scatter diagram of results (a) in results for 500 modulations and front illumination (b) in results for 250 modulations and front illumination.



(a) A scatter diagram of results for 500 modulations and back illumination.



(b) A scatter diagram of results for 250 modulations and back illumination.

Fig.8 A scatter diagram of results (a) in results for 500 modulations and back illumination (b) in results for 250 modulations and back illumination.

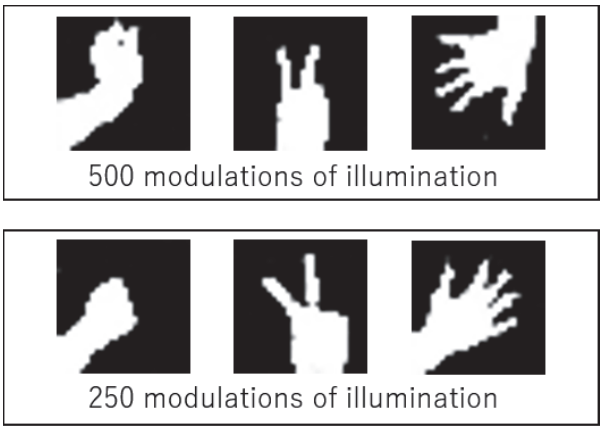


Fig.9 Restored images by deep learning when front illumination.

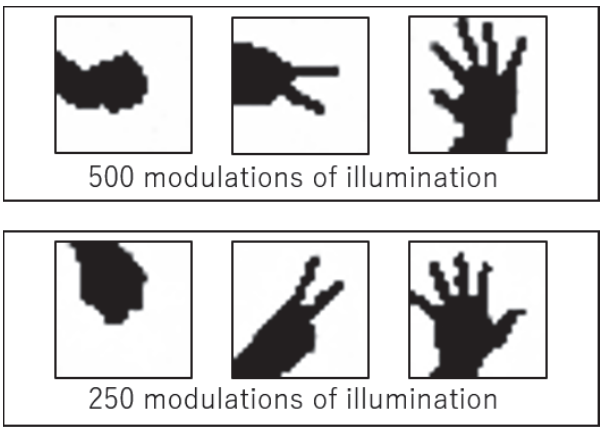


Fig.10 Restored images by deep learning when back illumination.

5 DISCUSSION

In Fig. 7 and Fig. 8, there is a slight variation. This is probably due to the fact that the number of training data is only 6800. Therefore, we can expect better results by increasing the amount of training data. The only difference between the two original images used is the processing of negative and positive inversions. The fact that the difference in the restoration is due to the inversion of the luminance value alone shows the significance of the case where the image is irradiated from behind the gesture. The randomly generated masks used in the modulation of illumination in this paper were performed as 500 and 250. The goal is to reduce this number even further by changing the parameters of the deep learning model and by considering other models that are better at feature extraction. In addition, for future applications, in order to perform gesture recognition to a person, which is the objective, the enlargement of the single-pixel detector lens is necessary, considering the large size of the subject to be measured. Therefore, using pAIRR, the gesture image of a hand touching an aerial display image can be converted to a single-pixel image. This enables us to measure the gesture with a single-pixel detector, and we are considering to detect the position and recognition of the gesture [8,9].

6 CONCLUSION

In this paper, we used single-pixel imaging and deep learning to protect privacy, and compared the modulation of illumination from behind and in front of the gesture. A comparison of structural similarity shows the effectiveness of the gesture's shadow that means the method of modulating illumination from behind.

A part of this work was supported by JST/ACCEL (grant no. JPMJAC1601) and JSPS KAKENHI (19H00871, 20H05702).

REFERENCES

[1] M. F. Duarte, M. A. Davenport, D. Takhar, J. N. Laska, T. Sun, K. F. Kelly, R. G. Baraniuk, "Single-pixel imaging via compressive sampling", IEEE Signal Processing Magazine, Vol. 25, No. 2, pp. 83-91 (2008).

[2] T. Tokimoto, S. Suyama, H. Yamamoto, "4320-Hz LED Display with Pulse-Width Mosulation by Use of a Nonlinear Clock", Journal of Display Technology, Vol. 12, No. 12, pp. 1581-1587 (2016).

[3] S. Onose, M. Takahashi, Y. Mizutani, T. Yasui, H. Yamamoto, "Single Pixel Imaging with a High-Frame-Rate LED Digital Signage", Proc. IDW/AD'16, pp. 1495-1498 (2016).

[4] K. Shibuya, K. Nakae, Y. Mizutani, T. Iwata, "Comparison of reconstructed images between ghost imaging and Hadamard transform imaging", Opt. Rev. **22**, pp. 897-902 (2015).

[5] D. V. Strekalov, A. V. Sergienko, D. N. Klyshko, T. H. Shih, "Observation of Two-Photon "Ghost" Interference and Diffraction", Phys. Rev. Lett. **74**, pp. 3600-3603 (1995).

[6] O. Ronneberger, P. Fischer, T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation", arXiv, 1505.04597v1 (2015).

[7] M. Yasugi, Y. Mizutani, T. Yasui, H. Yamamoto, "Deep Learning for Single-Pixel Imaging Without Normalization and Image Output", Proc. Of JSAP-OSA Joint Symposia 2020, 9p-Z10-6 (2020).

[8] S. Morita, H. Yamamoto, "Single Pixel Imaging with pAIRR", OPJ-OSA Joint Symposia on Nanophotonics and Digital Photonics, 31aOD5 (2017).

[9] S. Morita, S. Onose, M. Sasaki, H. Yamamoto, "Single Pixel Imaging on Aerial Display with AIRR", Proc. IDW'17, pp. 958-961 (2017).