Gesture Classification of Single-Pixel-Imaging Reconstruction by Using Deep Learning

Hiroki Takatsuka[1], Masaki Yasugi[1], Naoya Mukoujima[1], Shiro Suyama[1], Hirotsugu Yamamoto[1]

hirotsugu@yamamotolab.science
[1]Utsunomiya Univ., 7-1-2 Yoto, Utsunomiya, Tochigi 321-0904, Japan

Keywords: single-pixel imaging, deep learning, gesture classification

ABSTRACT

We report classification of images reconstructed with different number of masks using single pixel imaging. The target images to be classified are hand gestures (rock, scissors, paper). Deep learning based on LeNet is used for classification. Images that were reconstructed with single-pixel imaging by use of with 400 or more masks were accurately classified by 90% or more.

1 Introduction

Modern cameras typically use an array of millions of detector pixels to capture images. In contrast, single-pixel cameras use a sequence of mask patterns to filter the scene along with the corresponding measurements of the transmitted intensity which is recorded using a single-pixel detector [1]. Single pixel imaging which employs active illumination to acquire spatial information is an innovative imaging scheme and has received increasing attention. It is applicable to imaging at non-visible wavelengths and imaging under low light conditions. However, single-pixel imaging has once encountered problems of low reconstruction quality and long data-acquisition time [2]. In this imaging method, the illumination is repeatedly modulated with respect to the subject and the imaging 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 reduce the number of measurements, deep learning is used to improve image quality of the reconstructed images [3]. Another feature of single-pixel imaging is privacy prevention. Single-pixel imaging enables us to detect a shadow picture of a user by use of a single detector and specially encoded display as the illumination. Thus, our goal is to realize privacy-prevented gesture interface by utilizing single-pixel-imaging and deep learning.

Human gesture recognition involves deriving meaningful reasoning from human movements. Applications of gesture recognition include human-computer interactions, patient monitoring, surveillance, robotics, and sign language recognition [4]. Recently, gesture recognition methods by use of deep learning have gained wide acceptance due to their generalization capabilities and high accuracy in detecting and classifying gestures. These methods have been shown to work well for clean datasets. However, gesture recognition under degraded conditions, for example, in cases where gestures are partially occluded, or in low illumination conditions, remains a challenge. In these degraded environments, the gestures may not be fully recorded during the camera pickup process, which can make gesture recognition under such conditions more challenging [5]. Thus, gesture classification of images that are reconstructed by use of a limited number of measurements in single pixel imaging is still a challenge.

In this paper, we use deep learning to classify images reconstructed with single pixel imaging. Single pixel imaging changes the quality of the reconstructed image depending on the number of illumination masks. We investigate the relationship between the number of masks and the accuracy of image classification.

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 conducted by use of 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. The restored image with a small number of measurements contains a lot of noise. The intensity correlation function can be expressed as:

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

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 n-th randomly patterned mask in the coordinates (x,y). $\Delta A(n)$ is the deviation of the average value of the light intensity pixel detector. $A(n)$ can also be given by

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

where $T(x,y)$ denotes the transmission function [6–8].

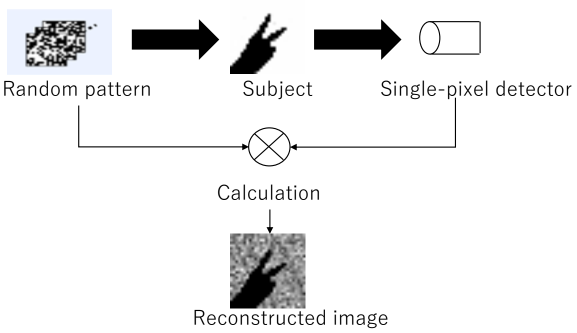


Fig. 1 Schematic diagram of single-pixel imaging to detect shadow picture.

3 Experiments

3.1 Training Data Set for Hand Gesture

We have built the training data set used in this study for deep learning. There are three types of hand gestures taken: rock, scissors, and paper. We shot a video of these three different hand gestures in front of a lighted display. Each frame of this video was taken as an image and resized to 28×28 pixel images [6]. From these images, 18000 images were used in this study (PNG format). These images are composed of 6000 images each of rock, scissors and paper. Port of those data are shown in Fig. 2. These images were modulated in single-pixel imaging to obtain 12000 training data, 4500 validation data and 1500 test data, respectively. The data generated by this method was performed with 10000, 5000 and 2500 masks. In addition, from 1000 masks to 100 masks in 100 masks. As the number of modulations at 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 [9].

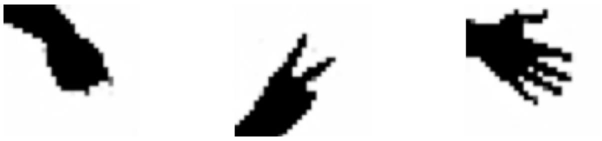
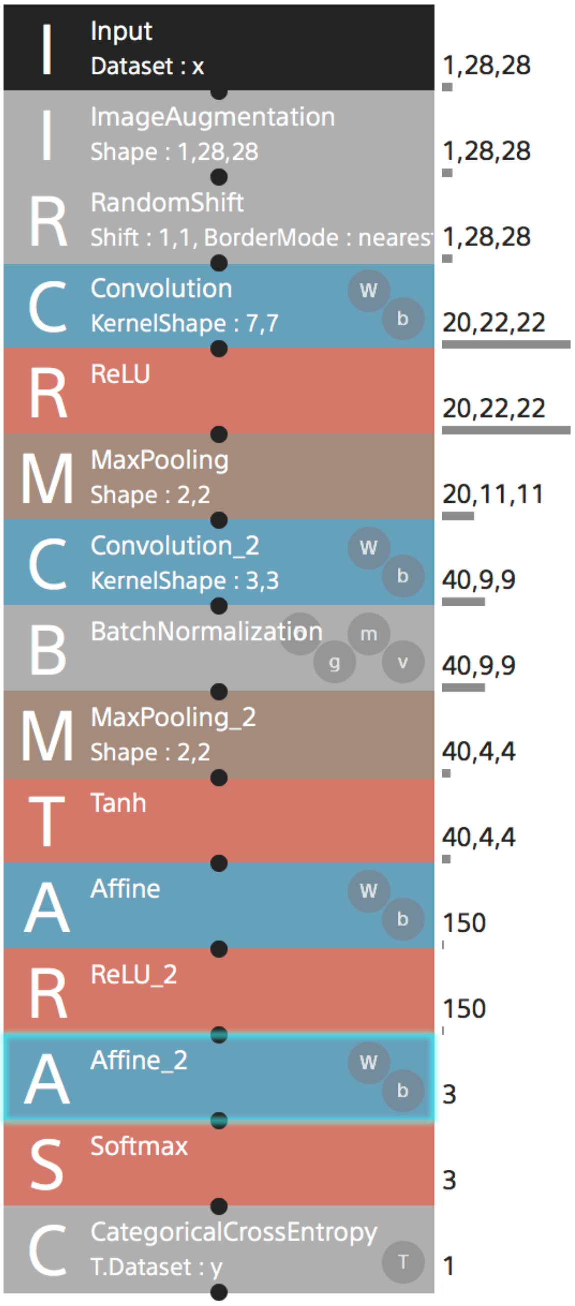


Fig.2 Port of training data.

3.2 Network Model for Deep Learning

In this paper, deep learning was performed by using Sony's NNC (Neural Network Console). The network model for deep learning used in this research in shown in Fig. 3. This model is based on LeNet [10]. This model inputs 28×28 pixel images. First, training data patterns are operated by ImageAugumentation and RandomShift, where ImageAugmentation flips left and right and rotates 1 radian left and right; and RnadomShift shifts in the range of 1 pixel up, down, left, and right. Next, repeated convolution, ReLU, and MaxPooling follow. Then, we use activation treatment and affine to reduce the output to three. Finally, the probabilities of each of the three are output. Batch size is 1024. Optimization method is AMSGrad. The learning rate is 0.001 and the number of epoch is 100.



Input
Dataset : x
1,28,28

ImageAugmentation
Shape : 1,28,28
1,28,28

RandomShift
Shift : 1,1, BorderMode : neares
1,28,28

Convolution
KernelShape : 7,7
20,22,22

ReLU
20,22,22

MaxPooling
Shape : 2,2
20,11,11

Convolution_2
KernelShape : 3,3
40,9,9

BatchNormalization
40,9,9

MaxPooling_2
Shape : 2,2
40,4,4

Tanh
40,4,4

Affine
150

ReLU_2
150

Affine_2
3

Softmax
3

CategoricalCrossEntropy
T.Dataset : y
1

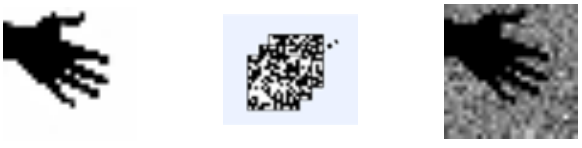
Fig. 3 Network model.

4 Results

Single pixel imaging is performed on hand gesture shadow pictures. An example of reconstructed images by use of 10000 masks is shown in Fig. 4. Note that Fig. 4 shows the input image (hand gesture shadow picture), examples of random pattern masks, and the reconstructed image.

Then, image classification was performed using deep learning using data obtained by single pixel imaging with 10000 masks. The neural network used for image classification is shown in Fig. 3. the learning curves and results of image classification using this neural network are shown in Fig. 5 and Table 1, where labels are 0 for rock, 1 for scissors, and 2 for paper; and y’_0, y’_1 and y’_2 are results of classification, respectively. All the images were classified correctly.

Then, single pixel imaging was performed by use of various mask numbers, and image classification was performed on the reconstructed shadow-picture images. The reconstructed images by use of a variety number of masks are shown in Fig. 6. As the number of masks decreases, the type of gesture becomes harder to recognize. We have performed gesture classification on these reconstructed images by using the LeNet network model. The relationship between the number of masks and the accuracy rate is shown in Fig. 7. The result of accuracy rate for different gestures is shown in Fig. 8. The paper maintained a relatively high accuracy rate.



Original image

random pattern mask

Reconstructed image

Fig. 4 Reconstructed image with single pixel imaging by use of 10000 random pattern masks.

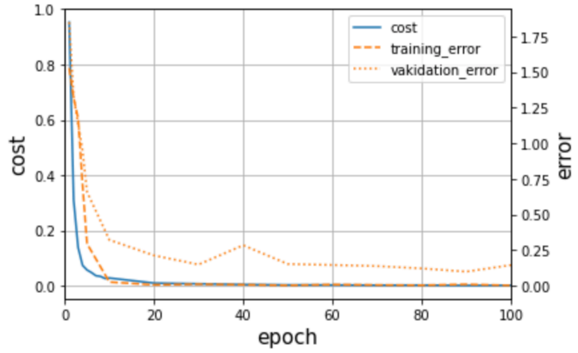
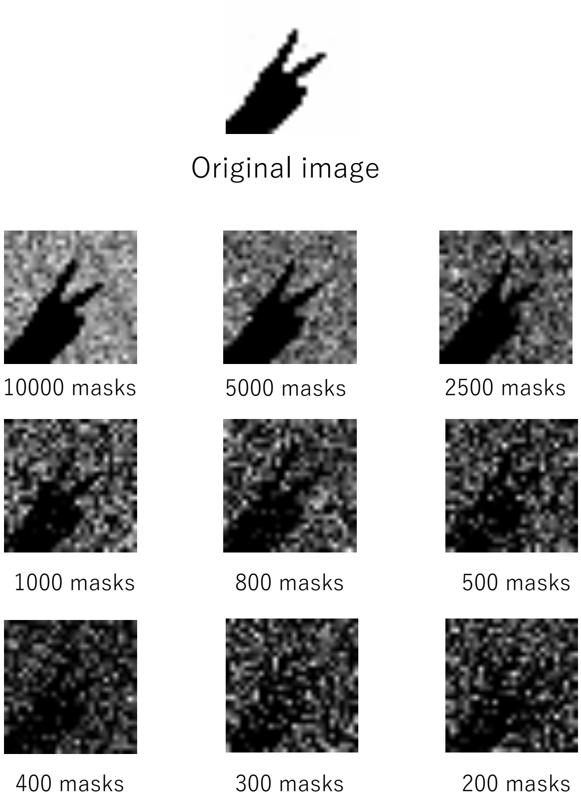


Fig. 5 Learning curves of gesture recognition of images reconstructed by use of 10000 masks.

Table 1 Results of image classification of images reconstructed by use of 10000 masks.

	y’_0	y’_1	y’_2
Label=0	500	0	0
Label=1	0	500	0
Label=2	0	0	500
Precision	1	1	1



Original image

10000 masks 5000 masks 2500 masks

1000 masks 800 masks 500 masks

400 masks 300 masks 200 masks

Fig. 6 Reconstructed images by use of different number of masks.

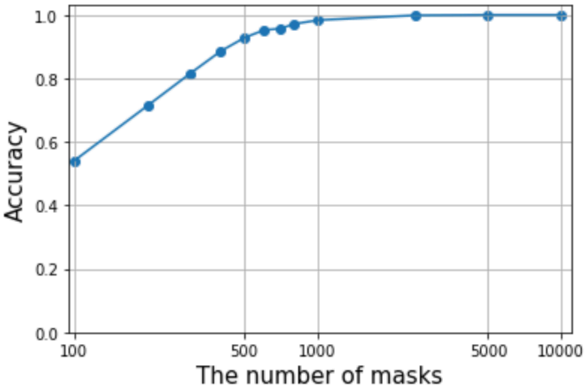


Fig. 7 Relationship between the number of masks and the accuracy rate on gesture classification.

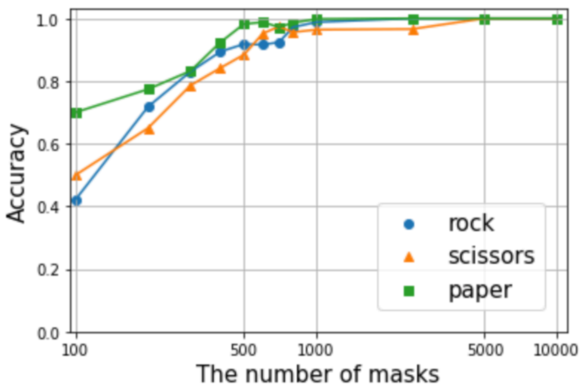


Fig. 8 Accuracy rate for different gestures.

5 Discussion

In Fig. 7, image classification accuracy is affected by the number of masks for single pixel imaging. The accuracy rate of image classification began to decrease when the number of masks fell below 1000, and the accuracy rate of image classification began to drop sharply when the number of masks fell below 500. Consequently, the characteristics of each gesture were lost, and it seems that the accuracy rate of image classification began to decrease even in deep learning.

In Fig. 8, which compares the image classification accuracy rate for each gesture, the accuracy rate of image classification with a small number of masks was relatively high for paper and low for rock and scissors. The reasons is that the finger part of scissors becomes unclear and it cannot be distinguished from rock.

In order to perform more accurate classification under degraded conditions, it is necessary to accurately capture the characteristics of each classification image. A prospective solution on this problem is to insert a noise-canceling function such as an autoencoder to the network and restoring using U-net.

6 Conclusion

We have realized gesture classification network for the shadow pictures of hand gestures were reconstructed by single pixel imaging by use of different numbers of masks. Although single pixel imaging requires a large number of masks to reconstruct a high resolution image, deep learning shows a possibility to classify gestures with a smaller number of masks. When the number of masks is 400 or more, the accuracy of image classification exceeds 90%.

References

[1] G. M. Gibson, S. D. Johnson, M. J. Padgett, “Single-pixel imaging 12 years on: a review,” Optics Express, Vol. 28, Issue 19, pp. 28190–28208 (2020).

[2] Z. Zhang, X, Wang, G. Zheng, J. Zhong, “Hadamard single-pixel imaging versus Fourier single-pixel imaging”, Optics Express, Vol. 25, Issue 16, pp. 19619–19639 (2017).

[3] N. Mukojima, M. Yasugi, Y. Mizutani, T. Yasui, H. Yamamoto, “Deep-Learning-Assisted Single-Pixel Imaging for Gesture Recognition in Consideration of Privacy,” IEICE Transactions on Electronics., accepted.

[4] S. Mitra, T. Acharya, “Gesture Recognition: A Survey,” IEEE Transactions on Systems, Man, and Cybernetics, Part C, Vol. 37, Issue 3, pp. 311–324 (2007).

[5] G. Krishnan, R. Joshi, T. O’Connor, F. Pla, B. Javidi, “Human gesture recognition under degraded environments using 3D-integral imaging and deep learning,” Optics Express, Vol. 28, Issue 13, pp.19711–19725 (2020).

[6] K. Shibuya, K. Nakae, Y. Mizutani, T. Iwata, “Comparison of reconstructed images between ghost imaging and Hadamard transform imaging”, Optical Review, Vol. 22, pp. 897–902 (2015).

[7] D. V. Strekalov, A. V. Sergienko, D. N. Klyshko, T. H. Shih, “Observation of Two-Python “Ghost” interference and Diffraction”, Physical Review Letters, Vol. 74, pp. 3600―3603 (1995).

[8] N. Mukoujima, M. Yasugi, Y. Mizutani, T. Yasui, H. Yamamoto, “Deep-Learning-Assisted Single-Pixel Imaging for Gesture Recognition Considering Privacy,” Proc. IDW, Vol. 27, pp.985–988 (2020).

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

[10] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, L. D. Jackel, “Backpropagation Applied to Handwritten Zip Code Recognition”, Neural Computation, Vol. 1, Issue 4, pp. 541–551 (1989).