Effects of Training Images on CNN-Based Demodulation for Digital Signage and Image Sensor-Based VLC

Yuki Iyoda, Kentaro Kobayashi*, and Wataru Chujo

Department of Electrical and Electronic Engineering, Meijo University, Tempaku-ku, Japan;
Email: 213427004@ccmailg.meijo-u.ac.jp (Y.I.), wchujo@meijo-u.ac.jp (W.C.)
*Correspondence: kkobaysh@meijo-u.ac.jp (K.K.)

***Abstract*—This paper studies a visible light communication (VLC) system using a digital signage and an image sensor. The authors have focused on the demodulation part of the communication system, which modulates data signals without disturbing the visual information on the digital signage, and have proposed a novel concept that uses machine learning to demodulate the data signals from images received by the image sensor. However, it has not been fully clarified which parameters of the training images contribute to the performance of the machine learning-based demodulation. This paper extends the convolutional neural network (CNN)-based demodulation method and clarifies how much the number of parallelized data signals and the number of patterns of data signals in the training images contribute to the demodulation performance. The results show that the performance improves with the number of parallelized data signals in the training images, and that half of the signal patterns are sufficient for learning.**

***Keywords*—visible light communication, digital signage, image sensor, demodulation, machine learning**

I. Introduction

Digital signage is a medium that displays advertising images and various useful information using electronic display devices and has already been installed in many places [1]. There is image sensor-based VLC technology that realizes communication by modulating data signals on displayed images on the digital signage so that the data signals cannot be perceived by the human eye, and then capturing the displayed images with an image sensor such as a mobile phone camera [2, 3]. It modulates the data signal by changing the luminance or color components. In demodulation, the transmitted data signal is obtained from the captured images by signal processing. This method can be used to established only with existing digital signage and terminals with a built-in image sensor, without installing new communication equipment. With the image sensor-based VLC system, viewers can receive not only the displayed contents from the digital signage itself, but also the value-added information such as timely and location-specific benefits and augmented reality.

One of the difficulties is that when data signals are superimposed with high signal intensity to improve the communication quality, the visual quality of the signage image deteriorates, i.e., the superimposition of data signals is easily perceived by the human eye. Conversely, if the signal intensity is reduced to improve the visual quality, the data signals cannot be received correctly. One of the methods to solve this problem is to transmit data signals by placing multiple markers that change luminance or color at low speed, taking advantage of the human visual characteristic of being insensitive to gradual changes in images [4]. This method can be implemented without modifying the digital signage and image sensor, but only allows communication at about 10 bps. Another method that cannot be perceived by the human eye has been proposed by modulating the backlight of the digital signage at high speed, but this method requires modification of the digital signage [5]. In the researches focusing on data superimposition at the transmitter side, modulation using wavelet transform [6] and discrete cosine transform [7], which are modulation methods in frequency components, have been investigated. In addition, [8-11] have investigated modulation methods with less visual quality degradation by superimposing data signals with color components that are difficult for the human eye to perceive.

The above-mentioned studies mainly focus on the processing at the transmitter side. Focusing on the processing at the receiver side, all of them detect the digital signage from the captured image and demodulate the transmitted data signal. It can be regarded as the process of determining which data signal is displayed on the digital signage in the captured image, and in such image processing, machine learning is rapidly developing. The previous study [12] proposed a demodulation method using CNN-based machine learning, which is trained with images simulating noise, blur, and misalignment that may occur in the received images, and showed that the

Manuscript received July 13, 2022; revised September 23, 2022; accepted November 23, 2022.

doi:10.12720/jcm.18.6.385-390

proposed method can achieve better demodulation performance than the conventional threshold-based method. However, it has not been fully clarified which parameters of the training images contribute to the demodulation performance.

This paper extends the CNN-based demodulation model and clarifies which parameters of the training images contribute to the performance of the machine learning-based demodulation. Specifically, this paper clarifies how much the number of parallelized data signals and the number of patterns of data signals in the training images contribute to the demodulation performance, and provides a measure of the dataset to achieve the desired performance.

This paper is organized as follows: Section II describes the system model of the considered VLC. Section III describes the details of the proposed CNN-based demodulation method. Section IV shows the numerical results of the performance evaluation. Section V summarizes the conclusion.

II. System Model

This paper assumes a VLC system like [8-11] shown in Fig. 1. The transmitter generates a mapping data signal based on the input data signal, adds it to the luminance or color component of the background image, and displays the resulting image on the digital signage as the transmitted image. The receiver obtains the data signal by capturing the digital signage with an image sensor. After detecting the signal, the receiver removes the background image by making the difference of two consecutive received images in the time direction and demodulates the data signal from the difference image.

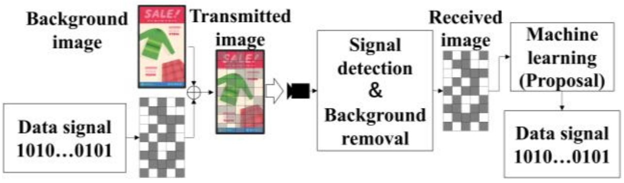


Figure 1. Digital signage and image sensor-based VLC system.

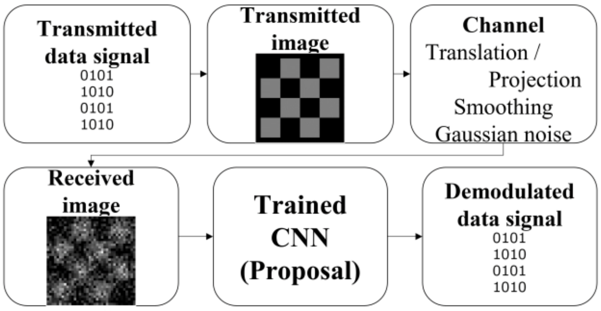


Figure 2. System model.

The purpose of this paper is to clarify the factors that contribute to performance by using a simplified model that simulates the system, rather than using actual transmitted and received images. Fig. 2 shows the simplified system model that focuses on the signal demodulation in the above system. Specifically, the data signal is modulated into the transmitted image, and the image received through the communication channel is demodulated by the proposed method to obtain the data signal. In the communication channel, it is assumed that there is misalignment in signal detection, blur due to light diffusion and lens, and noise due to illumination and background image, respectively, and these are modeled as translation or projection transformation, smoothing, and Gaussian noise on the transmitted image. The details are described in the following sections.

A. *Modulation*

The data signal $\boldsymbol{d}$ to be transmitted is parallelized and defined as $M \times N$ binary symbols as follows.

$$\boldsymbol{d} = \begin{bmatrix} d_{1,1} & \cdots & d_{1,N} \\ \vdots & \ddots & \vdots \\ d_{M,1} & \cdots & d_{M,N} \end{bmatrix}$$

$$d_{m,n} = \{0,1\}(m = 1,2 \cdots M\ , n = 1,2 \cdots N)$$

Fig. 3 shows the conceptual diagram of the generation of the transmitted image D corresponding to the data signal $\boldsymbol{d}$. D is I pixels in width and J pixels in height and has $M \times N$ rectangular areas. Each rectangular area is called a cell. The cell corresponding to the m-th row and n-th column is denoted as $C_{m,n}$, and each cell contains pixels of width I/M and height J/N. Based on the corresponding data signal, the cell $C_{m,n}$ is given as follows, where the value α is the signal intensity.

$$C_{m,n} = \begin{cases} \begin{bmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{bmatrix} & (d_{m,n} = 0) \\ \begin{bmatrix} \alpha & \cdots & \alpha \\ \vdots & \ddots & \vdots \\ \alpha & \cdots & \alpha \end{bmatrix} & (d_{m,n} = 1) \end{cases}$$

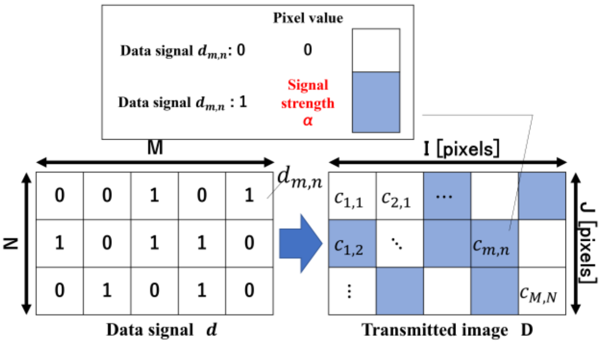


Figure 3. Correspondence between data signal and transmitted image.

B. *Communication Channel*

First, to reproduce the misalignment, a translation image is generated by translating the transmitted image D to the surrounding 8 neighborhoods by β-translation in addition to the centered (unmoved) image, or a projection image is generated by projection transformation in which the four vertices of D are shifted by β-translation to unmoved, inside, or outside, respectively. The resulting translation/projection image is treated as a misaligned image D_t of width $(I + 2\beta)$ × height $(J + 2\beta)$. Next, to reproduce the blur, a filtered image D_{tf} is generated by smoothing the misaligned image D_t with a Gaussian filter of standard deviation σ_f. Finally, to reproduce the noise, a Gaussian noise with mean 0 and standard deviation σ_n is added to each pixel of the filtered image D_{tf}, and then absolute value processing is performed to generate the received image D_{tfn}.

C. *Demodulation*

The received image D_{tfn} is input to a trained demodulator to obtain the demodulated data signal $\widehat{\boldsymbol{d}}$. The demodulator is constructed by learning various received images that reproduce the degradation caused by the communication channel. The proposed demodulation method is described in the next section.

III. Demodulation Method Using Machine Learning

A. *Machine Learning Model*

This paper proposes two machine learning models to clarify how much the number of parallelized data signals in the training images contributes to the demodulation performance.

The machine learning models are shown in Table I. In the mode (I), the data signal to be demodulated is $4 \times 4\,(= 16)$ binary symbols. This model takes a received image of 36×36 pixels containing 4×4 cells as input and 16 demodulated data symbols as output. In the other model (II), the data signal to be demodulated is $3 \times 3\,(= 9)$ binary symbols. This model takes a received image of 28×28 pixels containing 3×3 cells as input and 9 demodulated data symbols as output. Both machine learning models have the same layer structure, differing only in the size of the input image and the number of demodulated data signals to be output.

TABLE I. Machine Learning Model

	Layer	Output size	
		(I) 4×4 cells	(II) 3×3 cells
Input	Received image	36×36×1	28×28×1
Layer1	Conv. (5×5×3)	32×32×3	24×24×3
Layer2	Average Pooling (2×2)	16×16×3	12×12×3
Layer3	Conv. (5×5×128)	12×12×128	8×8×128
Layer4	Conv. (5×5×256)	8×8×256	4×4×256
Layer5	Average Pooling (2×2)	4×4×256	2×2×256
Layer6	Dropout (0.5)	4×4×256	2×2×256
Layer7	Flatten	4096	1024
Layer8	Dense (ReLU)	128	128
Layer9	Dropout (0.25)	128	128
Output	Dense (Sigmoid)	16	9

The proposed method uses the CNN-based deep learning of Keras, which is a high-level API of TensorFlow. The input layer is the layer that inputs the received images. The convolutional layer is a layer that performs convolutional operations on its input to extract features, and the activation function is ReLU. The pooling layer is a layer that summarizes the output of the convolutional layer and emphasizes the features. The dropout layer improves the robustness of the model by randomly dropping its input at an arbitrary rate. The flatten layer converts its input into a one-dimensional vector for input to the full coupling layer. The dense layer is the output layer that converts its input into probabilities using the activation function based on the features obtained so far, and performs maximum likelihood estimation. The proposed models are based on the well-known VGG16 [13]. In VGG16, it was shown that the reduction of parameters and the improvement of accuracy can be achieved by constructing the convolutional layer with three layers of 3×3 instead of one layer of 7×7, which shows the advantage of a deep neural network with a small kernel in the convolutional layer. However, in the proposed models, since the ratio of noise to the input signal is large, the kernel of the convolutional layer is set to 5×5 and the convolution is limited to three layers due to the small size of the input image. In addition, since Max Pooling tends to lose the information of cells with a pixel value of 0, Average Pooling is used. Unlike the general classification, the output layer outputs the likelihood of each binary symbol of the data signal in the range of 0 to 1 using the sigmoid function. The likelihood of each cell is thresholded with a threshold value of 0.5, and the resulting value is used as the demodulated data signal.

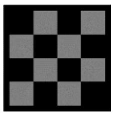
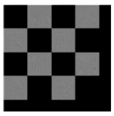
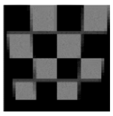
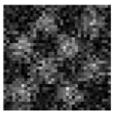
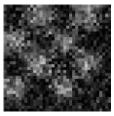
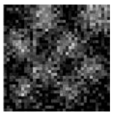
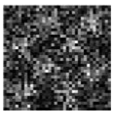
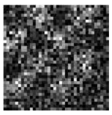
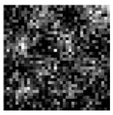
B. *Dataset*

The dataset is shown in Table II. If the transmitted image D is represented by 4×4 cells, $2^{16} = 65536$ different data signals can be represented, and if it is represented by 3×3 cells, $2^9 = 512$ different data signals can be represented. Therefore, in this dataset, different received images D_{tfn} are generated for all 65536 or 512 patterns of data signals by the procedure described in Chapter 2. The size of D is $I \times J = 32 \times 32$ pixels for 4×4 cells and $I \times J = 24 \times 24$ pixels for 3×3 cells. The signal intensity is set to $\alpha = 10$ in both cases. D_{tf} is obtained by applying a Gaussian filter with filter size 7×7 and standard deviation σ_f to a total of 90 patterns of D_t: 9 translation patterns in which D is translated to the center and its 8 neighborhoods, and 81 projection patterns in which the four vertices are translated to unmoved, inside, and outside, respectively $(\beta = 2)$. D_{tfn} is obtained by adding a Gaussian noise with mean 0 and standard deviation σ_n to D_{tf}. The standard deviations (σ_f, σ_n) of the Gaussian filter and Gaussian noise are chosen in six ways: (0,0), (1,1), (3,3), (3,5), (5,3), and (5,5). Thus, 65536 or 512 received images are generated for each of the 90 misalignment patterns and 6 combinations of smoothing and noise intensity, i.e., 540 patterns, and 10 sets of the above images are generated for each (σ_f, σ_n) except for $(\sigma_f, \sigma_n) = (0, 0)$ due to the randomness of Gaussian noise. 9 of the sets are used as training images and 1 set as unknown validation images. Table III shows examples of the dataset for one of the 65536 data signal patterns.

TABLE II. DATASET

σ_f, σ_n	Misalignment	Set	Training images of (I) Training images of (II)	Validation images of (I) Validation images of (II)
0,0	None	Training:1 Validation:1	65536 512	65536 512
1,1		Training:9 Validation:1	589824 4608	65536 512
3,3				
5,3				
3,5				
5,5				
0,0	8 translations	Training:1 Validation:1	524288 4096	524288 4096
1,1		Training:9 Validation:1	4718592 36864	524288 4096
3,3				
5,3				
3,5				
5,5				
0,0	81 projections	Training:1 Validation:1	5308416 41472	5308416 41472
1,1		Training:9 Validation:1	47775744 373248	5308416 41472
3,3				
5,3				
3,5				
5,5				

TABLE III. EXAMPLES OF DATASET

σ_f, σ_n	Misalignment		
	None	Translation	Projection
0,0			
3,3			
5,5			

C. Training Method

The machine learning models shown in Table I are trained on the dataset shown in Table II and the data signals that are the correct answers. Due to the large size of the dataset, the contribution of the training order to the demodulation performance is also investigated in the following two ways.

1) Learning of images containing all smoothing, noise, and misalignment patterns for every 64 data signal patterns.
2) Learning of images containing all data signal patterns with all smoothing and noise patterns for every misalignment pattern.

The loss function is mean-square error, and mini-batch learning is performed. The mini-batch size is set to 4096 images, which can be stably expanded in memory, and the number of training sessions is set to 15.

IV. PERFORMANCE EVALUATION

A. Evaluation Method

The first point to be clarified is how much the number of parallelized data signals contributes to the demodulation performance. This is verified by comparing the threshold decision method with the proposed methods: the model (I) with 4×4 cells and the model (II) with 3×3 cells. The demodulation performance is evaluated by using the bit error rate as an evaluation metric, which is calculated using the validation images described in Section III.B. The validation images are processed in the same way as the training images; however, they are unknown and different from the training images because the Gaussian noise is random. In the threshold decision method, the demodulated data signals are obtained by thresholding each averaged cell value of 4×4 cells of the unknown validation images with half of the signal intensity α without considering the misalignment. In the proposed method, the demodulated data signals are obtained by inputting the unknown validation images to the trained neural network. Without information on the data signal intensity, smoothing, noise, and misalignment, the trained neural network determines the output only by the weights obtained during training. To show the effect of training contents on the demodulation performance, the following three evaluations were performed for two models and two training orders (A) and (B) shown in Section III.C.

1) Evaluation with the validation images including smoothing and noise only.
2) Evaluation with the validation images including translation in addition to smoothing and noise, excluding the images of (1).
3) Evaluation with the validation images including projection transformation in addition to smoothing and noise, excluding the images of (1) and (2).

The other point to be clarified is how much the number of data signal patterns contributes to the demodulation performance. This is verified by comparing the model (I) with different numbers of data signal patterns contained in the dataset. Here, in order to focus on the effect of the data signal patterns, the proposed method was trained and verified using the dataset without misalignment.

B. Results

Fig. 4 shows the bit error rate of the threshold decision method, the model (I) with 4×4 cells, and the model (II) with 3×3 cells, for the validation images (1), (2), and (3). The performance of the two machine learning models is shown for two training orders (A) and (B). For each case, the left bar (blue, monochromatic) shows the bit error rate for the validation images with smoothing and noise only, the middle bar (orange, diagonal) shows the bit error rate for the validation images with translation in addition to smoothing and noise, and the right bar (black, checkerboard) shows the bit error rate for the validation images with projection transformation in addition to smoothing and noise. As shown in Fig. 4, both models (I) and (II) achieve better performance than the threshold decision method. The performance of both models (I) and (II) depends on the training order. For the model (I) with 4×4 cells, the training order (A) is better than (B). For the model (II) with 3×3 cells, the training order (B) is better than (A). In the case of model (I), for each of the 65536 patterns of data signals, there are 90 patterns of misalignments and 9 sets of 6 combinations of smoothing and noise, so there are many more patterns of data signals than patterns of these channel disturbances. In contrast, in the case of model (II), there are 512 patterns of data signals, so there are more patterns of channel disturbances than patterns of data signals. From these results, it is expected that the performance will be improved by training the dataset in the order of larger patterns of data signals or channel disturbances, since the learned images in each training session will contain more different information, but a better ordering is a topic for future work.

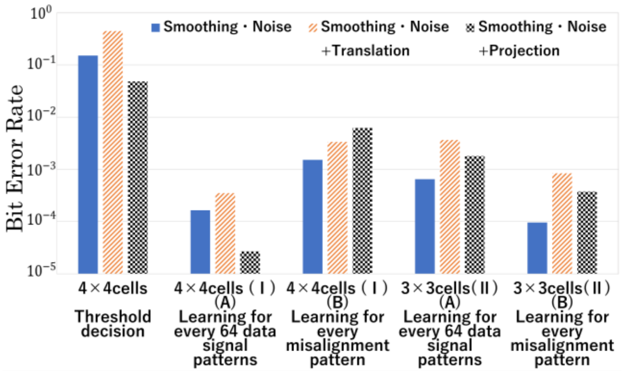


Figure 4. Bit error rate for each type of validation images.

Comparing the better result of the model (I) with that of the model (II), the bit error rate of the models (II) is about 10 times worse. In general, it is known that machine learning performance improves as the amount of training data increases, and in this case, the reason is that the number of training images is 128 times smaller due to the number of data signal patterns. In addition, the number of learned channel disturbances is reduced in proportion to the smaller number of data signal patterns, so the performance will degrade. Therefore, for the same number of learned channel disturbance patterns, the model with the larger number of parallelized data signals will have better performance. However, since the training time is proportional to the number of training images, the time required to construct the model (II) is about 120 times less than that of the model (I), which is a significant reduction.

Fig. 5 shows the relationship between the number of learned data signal patterns and the bit error rate for the model (I). Here, the verification was performed after each training of 1024 data signal patterns randomly sampled with replacement. In the model (I), bit errors occurred only for $(\sigma_f, \sigma_n) = (3,5)$ and $(5,5)$, where the standard deviation of the noise is large. It can be seen that after learning about half of the total number of data signal patterns, the performance converges even as the number of patterns to learn increases. Therefore, it is not necessary to learn all the data signal patterns, but only about half of them. As a future study, it will be possible to achieve higher performance with even fewer training patterns by selecting patterns based on image correlation and other factors, and a further reduction in learning time is expected.

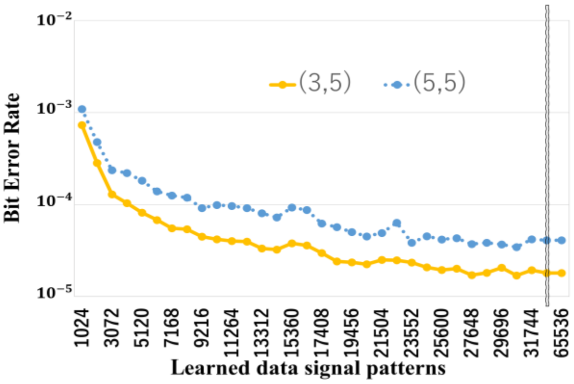


Figure 5. Bit error rate for the number of learned data signal patterns.

V. Conclusion

This paper proposed a method for demodulating data signals by CNN-based machine learning for digital signage and image sensor-based VLC. The effects of the training images on the demodulation performance were evaluated. The simulation results showed that the case with the larger number of parallelized data signals in the training images has better performance. It was also shown that it is not necessary to train all the data signal patterns, but only about half of them are sufficient. Further studies to improve the performance and reduce the learning time based on the results of this study are left for future work.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Yuki Iyoda and Kentaro Kobayashi conceived the idea of this study. Yuki Iyoda made substantial contributions to the system implementation and data analysis. Kentaro Kobayashi and Wataru Chujo contributed significantly to the interpretation of the results and supervised the conduct of this study. Yuki Iyoda drafted the original manuscript. Kentaro Kobayashi and Wataru Chujo critically revised the manuscript for intellectual content. All authors approved the final version of the manuscript.

REFERENCES

[1] Digital Signage Consortium. [Online]. Available: https://digitalsignage.jp/

[2] N. Saeed *et al*., "Optical camera communications: Survey, use cases, challenges, and future trends," *Physical Communication*, vol. 37, Oct. 2019.

[3] W. Liu and Z. Xu, "Some practical constraints and solutions for optical camera communication," *Phil. Trans. R. Soc. A*, vol, 378, no. 2169, Mar. 2020.

[4] K. Kuraki, S. Nakagata, R. Tanaka, and T. Anan, "Data transfer technology to enable communication between displays and smart devices," *FUJITSU Sci. Tech. J.*, vol. 50, no. 1, pp. 40-45, Jan. 2014.

[5] H. Aoyama and M. Oshima, "Visible light communication using a conventional image sensor," in *Proc. IEEE Consumer Commun. Netw. Conf. (CCNC)*, pp. 103- 108, Jan. 2015.

[6] Y. Lin, T. Wada, K. Mukumoto, and H. Okada, "Performance evaluation of information embedding schemes based on wavelet transform for parallel transmission visible light communication systems," in *Proc. IEEE Global Conf. Consumer Electron. (GCCE)*, pp. 1-2, Oct. 2017.

[7] R. Mushu, T. Wada, K. Mukumoto, and H. Okada, "A proposal of information embedding scheme based on discrete cosine transform in parallel transmission visible light communications," in *Proc. IEEE Global Conf. Consumer Electron. (GCCE)*, pp. 175-176, Oct. 2018.

[8] H. Okada, S. Sato, T. Wada, K. Kobayashi, and M. Katayama, "Preventing degradation of the quality of visual information in digital signage and image-sensor-based visible light communication systems," *IEEE Photon. J.*, vol.10, no. 3, Art. no. 7903509, Apr. 2018.

[9] S. Abe, T. Hiraki, S. Fukushima, and T. Naemura, "Imperceptible color vibration for screen-camera communication via 2D binary pattern," *ITE Trans. Media Technology and Applications*, vol. 8, no. 3, pp. 170-185, July 2020.

[10] K. Zhang *et al*., "ChromaCode: A fully imperceptible screen-camera communication system," *IEEE Trans. Mobile Computing*, vol. 20, no. 3, pp. 861-876, 1 Mar. 2021.

[11] K. Shimei, K. Kobayashi, and W. Chujo, "Data signal modulation based on uniform color space for digital signage and image sensor based visible light communication," *IEICE ComEX*, vol. 11, no. 1, pp. 26-32, Oct. 2021.

[12] Y. Iyoda, K. Kobayashi, and W. Chujo, "Data signal demodulation based on machine learning for digital signage and image sensor based visible light communication," *IEICE ComEX*, vol. 10, no. 12, pp. 912-917, June. 2021.

[13] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. International Conf. on Learning Representations (ICLR)*, Apr. 2015.

Copyright © 2023 by the authors. This is an open access article distributed under the Creative Commons Attribution License (CC BY-NC-ND 4.0), which permits use, distribution and reproduction in any medium, provided that the article is properly cited, the use is non-commercial and no modifications or adaptations are made.