Research on Methods to Increase Recognition Rate of Korean Sign Language using Deep Learning

[1]So-Young Kwon, [2]*Yong-Hwan Lee

Abstract

Deaf people who use sign language as their first language sometimes have difficulty communicating because they do not know spoken Korean. Deaf people are also members of society, so we must support to create a society where everyone can live together. In this paper, we present a method to increase the recognition rate of Korean sign language using a CNN model. When the original image was used as input to the CNN model, the accuracy was 0.96, and when the image corresponding to the skin area in the YCbCr color space was used as input, the accuracy was 0.72. It was confirmed that inserting the original image itself would lead to better results. In other studies, the accuracy of the combined Conv1d and LSTM model was 0.92, and the accuracy of the AlexNet model was 0.92. The CNN model proposed in this paper is 0.96 and is proven to be helpful in recognizing Korean sign language.

Keywords: *Deep learning, CNN, Sign language, Deaf, Hand detection, Image processing*

I. Introduction

Communication scholars believe that the fundamental reason why humans became the lord of all creation is because humans have a better communication system than other animals [1][2]. Humans can establish mutual relationships through communication, and furthermore, they can perform friendship functions, information acquisition functions, persuasion functions, decision-making functions, and confirmation functions [3][4]. Modern society is changing rapidly with the advancement of science and technology, and effective communication skills are essential to respond to these changes [5].

Korean sign language is a visual-motor language system that is naturally occurring and non-verbal rule-dominant for deaf people, and the Korean sign language act was enacted in 2016 after long efforts by the agricultural society and related organizations [6]. Korean sign language is a unique language that was created based on vision and movement in Korea's deaf culture, and is the unique language of the deaf with equal qualifications to the Korean language [6]. Deaf refers to a person who is hearing impaired and uses sign language as their everyday language. Not all hearing impaired people use sign language, and there are people who communicate primarily through spoken language and people who communicate primarily through sign language. Among Korean deaf people, there are people who use Korean signed language as their mother tongue, and spoken Korean is their second language. Among deaf people who use Korean sign language as their first language, there are those who cannot speak spoken Korean properly, and because they do not know this, there are cases where it is difficult to communicate with each other, and the farming community is closed. There are cases where people with malicious intentions who are aware of this situation in the farming community are committing fraud targeting deaf people [7]. In 2017, the topic 'The secret of a man who gave grace to the deaf' was aired on KBS's 'In Depth 60 Minutes', a public program in Korea. It was an incident in which a fraudster committed fraud targeting deaf people, resulting in approximately 500 victims and 28 billion won in damages [7]. Since deaf people are members of society, everyone should make efforts to create a society

[1]*Ph.d Candidate, Kumoh National Institute of Technology, Dept. of Electronic Engineering (papaya4040@kumoh.ac.kr)*
[2]* Corresponding Author *Professor, Kumoh National Institute of Technology, School of Electronic Engineering (yhlee@kumoh.ac.kr)*

Received: Jan. 23, 2024, Revised: Feb. 13, 2024, Accepted: Feb. 16, 2024

where deaf people can naturally communicate and live together [7].

Deaf people use Text to Speech(TTS) devices and Augmentative Alternative Communication(AAC) devices as assistive devices to solve the inconvenience of communication [8]. AAC device is a device that outputs a pre-recorded voice when you select a picture that fits the situation. Devices using TTS technology are devices that output voice when text is input, and are inefficient because they have to find and input an expression for each situation [9]. To solve these problems, software-based sign language interpretation programs using hardware such as gloves with sensors, image processing, or Kinect are being developed and are providing much help to the deaf [10][11][12].

This paper presents a method to increase the recognition rate of Korean sign language using image processing and Convolutional Neural Networks(CNN). In previous studies, certain parts of the body with skin color were detected through color conversion such as YCbCr and HSI of the input image. Afterwards, images of hands with skin color were extracted and the motion characteristics were identified using edge detection and labeling [13][14][15][16]. We compared the accuracy of Korean sign language recognition between the converted input image into a CNN model using the existing method and the original input image into the CNN model. It was confirmed that the accuracy of the method using the original input image into the CNN model was 0.96, which was higher than when only the hand region was extracted and used.

II. Related research

2.1. Hand detection

The types of color models mainly used in image processing include RGB, HSI, and YCbCr. RGB data is raw data obtained from a camera sensor and refers to a method of expressing one color by appropriately mixing Red, Green, and Blue. HSI consists of Hue, Saturation, and Intensity, where Hue is the type of color and Saturation is the vividness of the color [17]. Intensity represents the brightness of light. The YCbCr color model consists of luminance and chrominance components and is one of the ways to encode RGB information rather than an absolute color space. Y represents brightness, Cr represents red intensity, and Cb represents blue intensity.

Existing studies are conducting preprocessing work to extract skin color areas when detecting hands in images. R, G, and B used in Eq. 1, Eq. 2, and Eq. 3 mean Red, Green, and Blue, respectively. In the RGB color area, when Red is over 100, Green is over 65, and Blue is over 40, it corresponds to skin color, and the corresponding conditions are shown in Eq. 1.

$$(100 \leq \mathrm{R}) and\ (65 \leq G) and\ (40 \leq B) \tag{1}$$

Figure 1 shows the image extracted from the input image and the region corresponding to the conditions in Eq. 1 as a binary image. Image processing was done using C++. In general, the RGB color space is affected by various environmental factors such as weather, time, and lighting because the distribution of colors varies depending on the brightness of the lighting. It can be seen that skin area of RGB color space was not properly detected, as can be seen in Figure 1.

Figure 1. (a) Input Image (b) Skin area in RGB color space

Eq. 2 is an equation for converting from the RGB color space to the HIS color space, and Eq. 3 is the skin area condition in HSI color space.

$$H = \begin{cases} cos^{-1}\left(\frac{0.5[(R-G)+(R-B)]}{\sqrt{(R-G)^2+(R-B)(G-B)}}\right), \text{if) G} \geq \text{B} \\ 360 - cos^{-1}\left(\frac{0.5[(R-G)+(R-B)]}{\sqrt{(R-G)^2+(R-B)(G-B)}}\right), otherwise \end{cases}$$
$$I = \frac{1}{3}(R+G+B)$$
$$S = 1 - \frac{1}{R+G+B}[\min(R,G,B)] \quad (2)$$

$$\text{Skin area} = (0 \leq \text{H}) and (H \leq 50) and (20 \leq S) and (S \leq 255) \quad (3)$$

Eq. 4 shows the equation for converting from RGB color space to YCbCr color space, and Eq. 5 shows the conditions for the skin color area in the YCbCr color area. In [10], AND operation was performed on image RGB, HSI, and YCbCr to detect hand areas in various environments. Figure 2 shows the skin area of the HSI color space, and the skin area of the YCbCr color space, and the results of AND processing all skin color areas of RGB, HSI, and YCbCr.

$$Y = (0.299 * R) + (0.587 * G) + (0.144 * B)$$
$$Cr = (R - Y) * 0.713 + 128$$
$$Cb = (B - Y) * 0.564 + 128 \quad (4)$$

$$\text{Skin area} = (77 \leq \text{Cb}) and (Cb \leq 127) and (133 \leq Cr) and (Cr \leq 173) \quad (5)$$

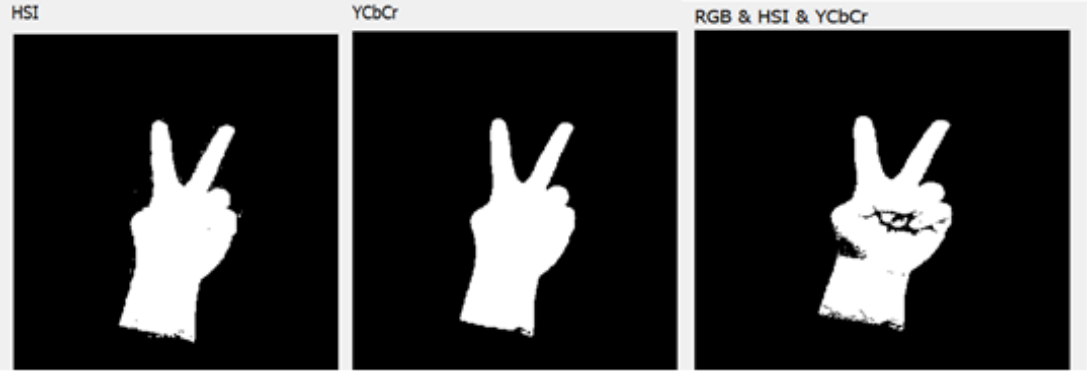

Figure 2. (a) Skin area in HSI color space (b) Skin area in YCbCr color space (c) AND-processed image of YCbCr, HSI, and RGB

People have a variety of skin colors, but since all humans have red color regardless of skin brightness, it is effective to detect skin color using only the color difference component. In this paper, the accuracy was obtained by using the skin area binary image in YCbCr color space, the skin area image in YCbCr color space, and the original image as input to the CNN model.

2.2. CNN

CNN is a neural network in which a convolutional layer and a pooling layer are added to an artificial neural network composed of a fully-connected neural network [18]. CNN is a method that complements the problems that arise when processing data such as images or videos in existing deep neural networks. In the convolution layer, a convolution mask (or window, or filter) is overlaid on each pixel of the input image, then the weight is multiplied by the input data, and the summed value is applied to the activation function and passed to the next layer. In particular, this convolution mask is applied to all pixels of the input image using a sliding window method. The convolution operation using the mask M for input data

I can be expressed as Eq. 6. In the case of the convolutional layer, unlike the fully connected layer, the weight that needs to be adjusted during the model learning process is the number of coefficients included in the mask, so the number of computer operations is dramatically reduced.

$$O(i,j) = \sum_{s=-1}^{1}\sum_{t=-1}^{1} I(i+s, j+t)\cdot M(s,t) \tag{6}$$

When performing a convolution operation, if the mask is placed on a boundary line, a space without pixel values may exist. Additionally, as the convolution operation is performed repeatedly, the size of the input data becomes smaller. Padding is a method of preventing the size of the input data from being reduced by performing a convolution operation by forcibly filling in the corners of the input data with specific values when there is no information. Zero-padding is a method that assumes that 0 is filled, and mirroring is a method that pads with the closest pixel value. Figure 3 shows pictures of zero padding and mirror padding.

0	0	0	0	0
0	5	6	2	0
0	3	2	4	0
0	3	2	3	0
0	0	0	0	0

5	5	6	2	2
5	5	6	2	2
3	3	2	4	4
3	3	2	3	3
3	3	2	3	3

Figure 3. (a)Zero padding (b)mirror padding

Stride refers to the moving interval when applying a mask to input data and moving it using a sliding window method. When the stride is 1, the size of the input data is not reduced when padding, etc. is performed, but when the stride is 2, the size of the input data is reduced by about 1/2. Eq. 7 represents the stride equation. O means the size of the output image, I means the size of the input image, and M means the filter size. P is padding and S is stride.

$$O_{width} = \frac{I_{width} + 2P + M_{width}}{S} + 1$$
$$O_{height} = \frac{I_{height} + 2P + M_{height}}{S} + 1 \tag{7}$$

In the pooling layer, pooling is an operation that reduces space in the horizontal and vertical directions. Maximum pooling is a method of reducing the size by selecting the maximum value from the target area, and average value pooling is a method of reducing the size by calculating the average of the values of the target area. In the case of the target area, set the window according to the size of the set stride and move at that interval. Figure 4 shows the max pooling and average pooling methods.

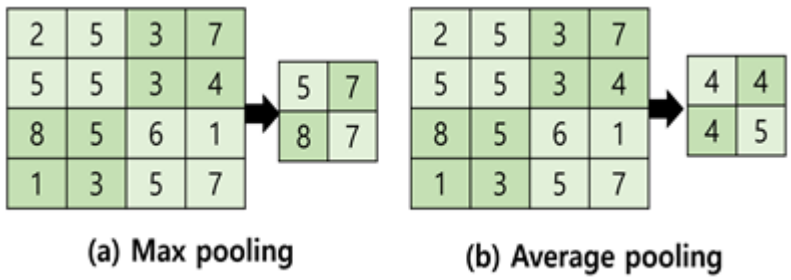

Figure 4. (a)Max pooling (b)Average pooling

When using CNN, Max Pooling is used more often than Average Pooling. When performing pooling, a method is often used to calculate the pooling size and stride value using (2, 2) so that there is no overlap. Sign language recognition research using deep learning includes sign language recognition research using CNN. The CNN model is widely used as the best performing model among deep learning architecture models in image recognition [19].

III. Proposed system

In this paper, the accuracy was compared by using the skin area binary image in YCbCr color space, the skin area image in YCbCr color space, and the original image as input to the CNN model. The confusion matrix is used to evaluate the performance of Korean sign language recognition implemented in this paper. Table 1. shows the confusion matrix [20]. TP is a case where something that is true is correctly classified as true, and FP is a case where something that is true is classified as false. FN is a case where a lie is classified as true, and TN is a case where a lie is correctly classified as false.

Table 1. Confusion Matrix

	Actually Positive	Actually Negative
Predicted Positive	TP (True Positive)	FP (False Positive)
Predicted Negative	FN (False Negative)	TN (True Negative)

The confusion matrix is an indicator for evaluating the performance of a classification model, and through it, the accuracy, precision, recall, F1-Score, etc. of the model can be determined. In this paper, accuracy is used as an indicator. Accuracy refers to the proportion of correct predictions made by the model and is shown in Eq. 8. It is the value obtained by dividing the sum of TP, FN, FP, and TN from TP + TN, which correctly predicted the result.

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \tag{8}$$

The experimental environment of this paper is shown in Table 2. There are a total of 42,000 pieces of data, of which 33,600 are training data and 8,400 test data.

Table 2. Experimental Environment

CPU	Intel i9-13900
RAM	64GB
OS	Window 10, 64-bit

Figure 5 shows the Korean sign language. The binary image of the skin area in the YCbCr color space is shown in Figure 6. Figure 7 is an image of the skin area in YCbCr color space.

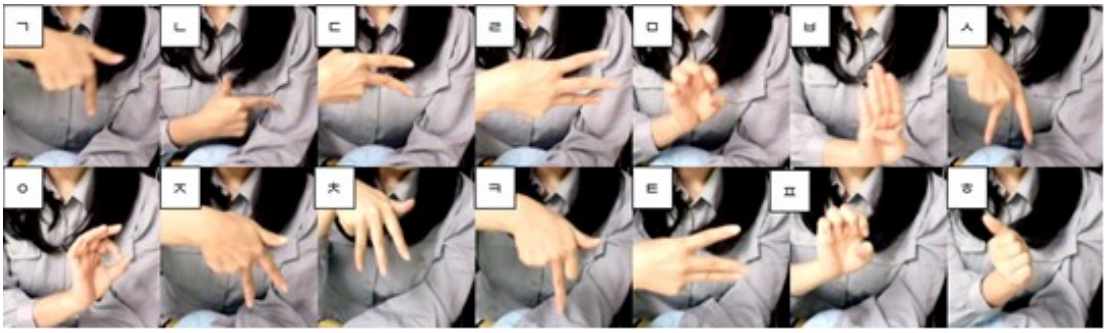

Figure 5. Korean sign language

Figure 6. Binary image of the skin area in the YCbCr color space

Figure 7. Image of the skin area in the YCbCr color space

Figure 8 shows the CNN model used in this paper. Rectified Linear Unit(ReLu) was used as an activation function. This function does not cause gradient loss problems and the learning speed is sufficiently fast. To extract important features and reduce dimensionality, max pooling of size (2, 2) was used. After flattening, 20% of the data is dropped out to prevent overfitting. A fully connected layer, which is a layer that predicts to classify images, exists after dropout. The fully connected layer consists of three layers, and the activation function of two layers used ReLu, and the activation function of one layer used softmax. When trained with the same model, when a binary image from which only the hand region is extracted is input to the CNN model, the accuracy is 0.68, and when an image from which only the hand region is extracted is input, the accuracy is 0.72. When the original video is input, the accuracy is 0.96. The verification accuracy graph of the system is shown in Figure 9. It was proved that the accuracy of using the original image itself in the CNN model for Korean sign language recognition was the highest at 0.96.

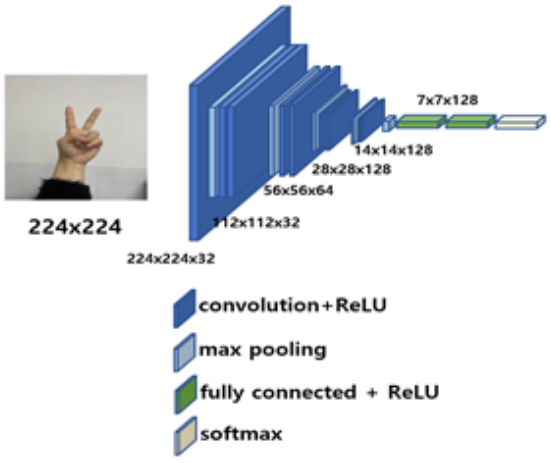

Figure 8. Proposed CNN model

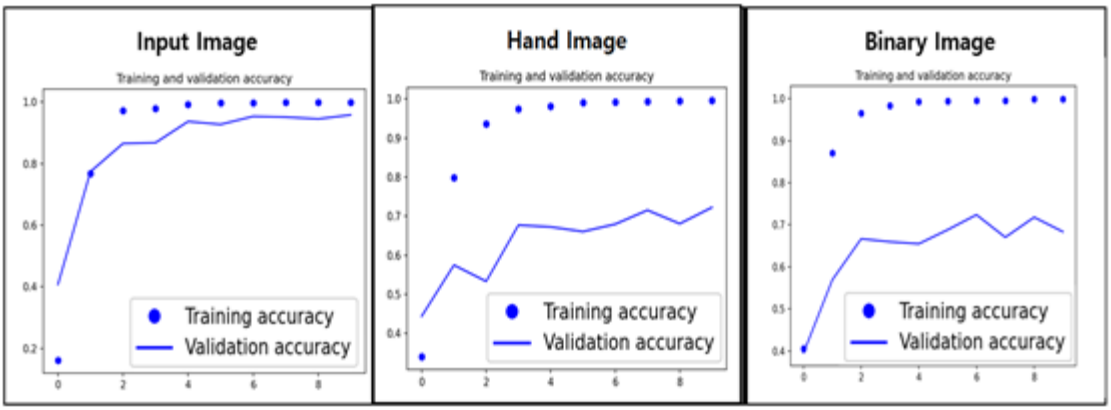

Figure 9. Traning and validation accuracy

Table 3. shows the accuracy when using the method used in other studies and when using the method proposed in this paper. The models were written in Python. When Conv1d and LSTM were combined, the accuracy was 0.92 [21]. The accuracy of the VGG16 model was 0.96, and the accuracy of the AlexNet model was 0.92 [22][23]. The CNN model proposed in this paper has the accuracy of 0.96, which is similar to or higher than the model proposed in other papers. Therefore, it is expected to be helpful in recognizing Korean sign language.

Table 3. Accuracy

Model	Accuracy
Conv1d + LSTM	0.92
VGG16	0.96
AlexNet	0.92
Proposed model	0.96

IV. Conclusion

Communication is an essential element for human interaction. Among deaf members of society, there are people whose first language is sign language. In this paper, we present a method to improve the accuracy of Korean sign language in sign language programs to facilitate easy communication between hearing and hearing people. The accuracy was compared when using the original image and when using an image from which only the hand area was extracted, and the experimental results showed that when the original image itself was used in the CNN model, the accuracy was 0.96, which was the highest accuracy. Additionally, it was experimentally measured that the method combining Conv1d and LSTM had an accuracy of 0.92, the method using VGG16 had an accuracy of 0.96, and the method using AlexNet had an accuracy of 0.92. The accuracy of the CNN model proposed in this paper is 0.96, which is similar or higher than other models, and is expected to help increase the recognition rate of Korean sign language programs.

X. Acknowledgments

This research was supported by Kumoh National Institute of Technology(2022～2023).

XI. References

[1] J. Z. Young, “Biological Point of View”, Approaches to Human Communication, Bud, Richard W and Brent D.Ruben (els), New Jersey : Hayden Book Co., Inc., 1972.
[2] Y. K. An, "Communication, Reason and Artificial Intelligence," Journal of AI Humanities, vol. 3, pp. 99-120, 2019.
[3] J. H. Ahn, “Communication and Human Relations”, Cogito, 34, pp. 163-188.
[4] I. J. Kim, “Effective Communication Skills”, Engineering education and technology transfer, vol. 6, no.3/4 , 1999, pp.55-59.
[5] K. C. Hong, H. S. Kim and Y. H. Han, “CNN-based Sign Language Translation Program for the Deaf”, KicsP, vol. 22, no. 4, 2021, pp.206-212.
[6] ″Korean Sign Language Method”, National Law Information Center. https://www.law.go.kr/%EB%B2%95%EB%A0%B9/%ED%95%9C%EA%B5%AD%EC%88%98%ED%99%94%EC%96%B8%EC%96%B4%EB%B2%95
[7] S. M. Youn, “Korean Sign Language(KSL) is Another Korean Language”, Korean Language and Literature, vol. 78, no. 78, 2021, pp. 121-144.
[8] S. Glennen, D. C. Decoste, “The Handbook of Augmentative and Alternative Communication”, Singular Publishing Group, INC. San Diego. London, 1997.
[9] K. P. Gwon and J. H. Yoo, “Numeric Sign Language Interpreting Algorithm Based on Hand Image Processing”, IEMEK, vol. 14, no. 3, pp. 133-142, 2019.
[10] Joshua R. New, “A Method for Hand Gesture Recognition”, Proceedings of IEEE Communication Systems and Network Technologies, pp. 919-923, 2002.
[11] C. Dong, C. Leu and Z. Yin, “American Sign Language Alphabet Recognition Using Microsoft Kinect”, Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 44-25, 2015.
[12] EnableTalk website, Avaliavle on : http://enabletalk.com.
[13] S. Kolkur, D. Kalbandee, P. Shimpi, C. Bapat and J. Jatakia, “Human Skin Detection Using RGB, HSB, and YCbCr Color Models”, Proceedings of IEEE Conference on Acoustic, Speech and Signal processing, pp. 324-332, 2017.
[14] H. S. Park, “Vehicle Tracking System using HSV Color Space at nighttime”, jkiiect, vol. 8, no. 4, pp. 270-274, 2015.
[15] Y. Xu and G. Pok, “Identification of Hand Region Based on YCgCr color Representation”, Journal of Applied Engineering Research, vol. 12, no. 6, pp. 1031-1034, 2017.
[16] Z. Zhengzhen and S. Yuexiang, “Skin Color Detecting Unite YCgCb Color Space with YCgCr Color Space”, Proceedings of Conference on Image Analysis and Signal Processing, pp. 221-225, 2009.
[17] Y. W. Choi, W. M. Yook and G. S. Cho, “Development of Building 3D Spatial Information Extracting System using HSI Color Model”, journal of Korean Society for Geospatial Information Science, vol. 21, no. 4, pp. 151-159, 2013.
[18] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard and L. D. Jackel , “Backpropagation Applied to Handwritten Zip Code Recognition”, Neural Computation, vol. 1, pp. 541-551, 1989.
[19] P. Molchanov, S Gupta, K. Kim and J. Kautz, “Hand Gesture Recognition with 3D Convolutional Neural Networks”, Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1-7, 2015.
[20] S. H. Park, J. M. Goo and C. H. Jo, “Receiver operating characteristic (ROC) curve: practical review for radiologists”, Korean journal of radiology. vol. 5, no. 1, pp. 11– 18, 2012.
[21] G. C. Kim and R. Ha, “Real-time Hand Expression Recognition Translation using Deep Learning”, Korean Institute of Information Scientists and Engineers, pp. 1774-1776, 2022.
[22] S. Gnanapriya, K. Rahimunnisa, M. Sowmiya, P. Deepika and S. Praveena Rachel Kamala, “Hand Detection and Gesture Recognition in Complex Backgrounds”, ICCMC-2023, pp. 829-833, 2023.
[23] K. C. Hong, H. S. Kim and Y. H. Han, “CNN-based Sign Language Translation Program for the Deaf”, JISPS, vol. 22, no. 4, pp. 206-212, 2021.

Authors

So-Young Kwon

2019 : M.S Degree in Dept. Of Electronic Engineering, Kumoh National Institute of Technology
2022~ Present : Ph.D candidate in Dept. Of Electronic Engineering, Kumoh National Institute of Technology

Research Interests : Digital SoC, Image Processing, Verilog HDL

Yong-Hwan Lee

1999~2002 : Research Engineer, Hynix Semiconductor
2003~2004 : Senior Research Engineer, Samsung Electronics
2004~Present : Professor, School of Electronic Engineering, Kumoh National Institute of Technology

Research Interests : Digital SoC, MIPI, Verilog HDL