Optical Character Recognition System Using BP Algorithm

***Sang Sung Park*†, *Won Gyo Jung*†, *Young Geun Shin*†, *Dong-Sik Jang*†**

† Department of Industrial Systems and Information Engineering, Korea University, Sungbuk-gu Anam-dong 5 Ga 1, Seoul 136-701, South Korea

Summary
Most government agencies and companies have kept proof data and documentations which are passed certain period of time and exchanged electronic forms by the regulation of an office management. The method that saving relevant documents by scanning or entering manually on computer was used for document's digitalizing. So that the government agencies and companies are trying to reduce these inconvenience nowadays. They use OCR (OCR : Optical Character Recognition) technique which is that saving relevant documents to DB after extracting text by using OCR(Optical Character Recognition). However, there is inconvenience in general OCR. That is, text should be entered to DB after classifying segments one by one in realized whole document after doing character recognition through OCR. In this paper, in order to solve this problem, we constructed OCR system that saves abstracted characters to DB automatically after extracting only equivalent and necessary characters from a large amount of documents by using BP algorithm that is one of Artificial neural network.

Key words:
ANN, BP, OCR, Digitalizing Documents, C#.

1. Introduction

Recently, most government agencies and companies have kept proof data and documentations which are passed certain period of time and exchanged electronic forms by the regulation of an office management. The method that saving relevant documents by scanning or entering manually on computer was used for document's digitalization. So that the government agencies and companies are trying to reduce these inconvenience nowadays. They use OCR (OCR : Optical Character Recognition) technique which is that saving relevant documents to DB after extracting text by using OCR. However, there is inconvenience in general OCR. That is, text should be entered to DB after classifying segments one by one in realized whole document after doing character recognition through OCR. In this paper, in order to solve this problem, we constructed OCR system that saves abstracted characters to DB automatically after extracting only equivalent and necessary characters from a large amount of documents by using BP algorithm that is one of Artificial neural network.
This paper is consisted as following. Describing problems of an existing study and this study's necessity in Paragraph 1. Introducing a related study about character recognition in Paragraph 2. Explaining a composition of proposed system in Paragraph 3. Describing the method to embody about proposed system in Paragraph 4. Explaining a sequence of constructed system in paragraph 5. And lastly, the paper is concluded with making reference to an effect of manufactured system and hereafter study direction in Paragraph 6.

2. Related study

2.1 The method for character recognition

The character recognition method is divided into roughly two branches including a deterministic method and a syntax method. A deterministic method is comparing an input pattern and a standard pattern by analyzing a literal pattern which is in document image. Then, recognizing their patterns by estimating the similarity of each other.

On the other hands, a syntax method is following a given syntax rule which is introducing similarity with syntax of language and pattern structure, and then identify the structure of patterns according to a given syntax rule [2] [3].

In this paper, We used deterministic method that compares input pattern and standard pattern. The character recognition is able to classify into Template matching method, Statistical method, Structural analysis method and etc according to a classification process. The Template matching method is that finding the most similar form by comparing with Template pattern , and it classifies a literal pattern according to a arrangement form. This method which is mainly using the character of one fixed form used a lot in the beginning. however frequency of use is less in the present due to the problems. Statistical character recognition method is that recognizing character by extracting a characteristic vector in indicator target .In this method , find the characters of statistical probability distribution of characteristic vector through the learning step, and separate the space of characteristic vector with an each class by using it. This classification model was defined well mathematically and in this method pattern of express, it is a very important issue that how define well

Manuscript received December 5, 2008
Manuscript revised December 20, 2008

of input pattern's character and how extract the character. The structure analytic character recognition method is extracting the base element of composed character such as stokes in a Chinese character and its correlation based on a literal composition principle .This method has gotten a fine theoretical array and simple method, but there is a shortcoming that it takes long realization time because the regulation of characteristic character is very various according to the fonts.

In order to recognize the character pattern, the study that using Neural Networks Model which is one of artificial intelligence system is getting into the spotlight. Neural networks model is modeling human's structure of brain, and it presupposed that displays good performance through connection of simple calculation element with neuron that is standard unit for composing a brain. Therefore, neural networks model is a suitable model in the problems that require lots of computational complexities and parallelism such as analysis voice, character, image and etc[7].

2. Artificial neural networks algorithm

Artificial Neural Network(ANN) used Back Propagation(BP) algorithm which is efficient learning of Multi-Layer Perceptron(MLP).[8].BP is consist of three layer forward structure that has hidden layer between input layer and output layer. Fig. 1 shows BP's structure. BP's learning method passes through a learning process that exchange early connection weight value to suitable value for data. The forward step presents input pattern of neural net and produce output by using a input function and activation function to an each node. All value can contain only binary value (0 or 1).

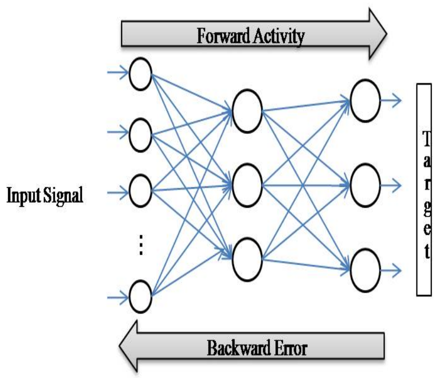

Fig. 1 The structure of BP algorithm

Activation function which is used in calculation of connection weight value used sigmoid function same as equation (1).

$$logsig(x) = \frac{1}{1+exp^{-x}} \quad (1)$$

Output value of hidden layer and output value of output layer which are according to activation function formula can express equation (2)and (3).

$$h = logsig(\sum_{i=1}^{n} w_{ij} x_i) \quad (2)$$
$$y = logsig(\sum_{j=1}^{m} w_{jk} x_j) \quad (3)$$

x_i = Input variation, w_{jk} = Connection weight between hidden layer and output layer, w_{ij} = Connection weight between input layer and hidden layer

Backward is a step for renewal a connection weight value which is important element in learning process. In this step, measuring an error through the formula by calculating difference of desire value and output value . And then resetting a connection weight between layer and layer for making a minimize error value in order of input layer and output layer. Resetting the final connection weight value which has an minimize error value via forward and backward steps.

$$e = \frac{\sum_{k=1}^{l}(t-y)^2}{2} \quad (4)$$

3. System composition

Component of optical character recognition system that proposed in this paper is divided at character learning part, document style setting part, image register part, character realization part and data saving part. System schematic diagram is same as Fig. 2.

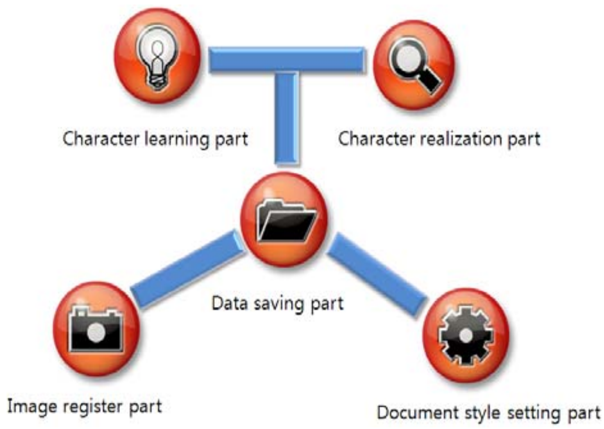

Fig. 2 System schematic diagram

3.1 Character learning part

Character learning part creates resetting connection weight that inputs character image and extracts information of pixel value about relevant image by using BP algorithm. Created connection weight becomes standard value for comparing and stored in data saving part during the recognition step. That is, it learns various fonts of character image and stores in data saving part in order to create literal standard value.

3.2 Document style setting part

Document style setting part sets recognition area that wish to achieve the character recognition and distinguish the document form. That is, it presets the document which has fixed form (ex) patent specification, financial products contract, etc) and necessary recognition area before doing character recognition.

3.3 Image register part

Image register part inputs document image to perform character recognition. Image register part involves the devices such as a scanner that converts document to an imaging.

3.4 Character realization part

Character realization department recognizes the character which is compared with the standard value that comes out the learning character step within a setting area for recognition, and the character should be corresponded to the document image that is matching a setting document style in the document style setting part. That is, when any document image was inputted through image register part, it realize the style of document image automatically. And if there is a conformable document style which is set in document style setting part, it realize the character.

3.5 Data saving part

Data saving part saves data that generated in document style setting part, image register part and character realization part. That is, it saves that document style data which occurs from document style setting part , any document image data which includes the character for recognition process in image register part and completely recognized character realization data and etc.

In these composition, inputted any document is applicable to the document style that is set in the realization part , the system is able to recognize only necessary segment so that mass document can be analyzed fast and enables character recognition.

4. System embodiment method

4.1 Character learning step

Fig. 3 is a flowchart that display character learning step which is preprocess phase of character recognition step.

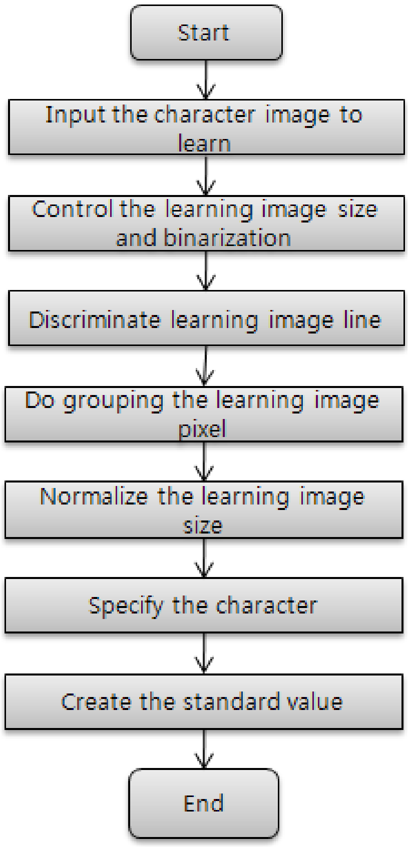

Fig. 3 Character learning step

In the character learning step, Input a character image to study for creating a literal standard value which is used at priority character recognition. Next, control whole size of inputted image and binarize whole pixel of image. The whole image size can be controlled, and it will adjust according to the specific size that is already preset. If the image is bigger than preset size, reduce an image size for reducing the data, and if not it doesn't need any control. After controlling the size of image, divide and binarize the whole pixels as 0 for white and 1 for black.

0 means background and 1 means character and thereby divide character and background definitely .
Next, in order recognize the character efficiently discriminate the line in the binarized image as Fig. 4.

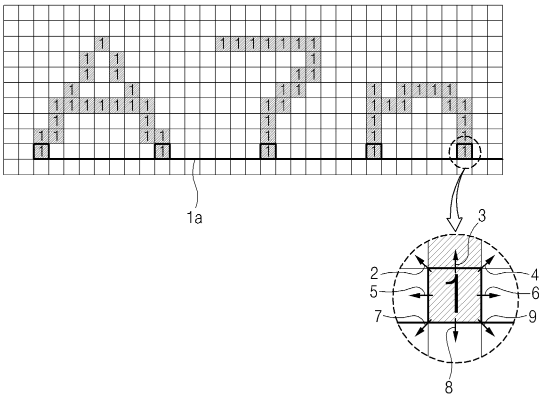

Fig. 4 Discriminate line

Discriminate line means that connect pixels which are located in the lowest position among characters that realized as 1 and recognize by line (1 a) , so it doesn't mean creating the line. At this time, measure the similarity of pixels which are located on adjacent 8 directions to lowest position pixel such as top-left, top-mid, top-right, left, right, bottom-left, mid-left, bottom-left., and then it can find the last point at the any line's end.

After discriminating the line, grouping the adjacent pixels which mean character and have 1 of pixel value by the line [9][10].

In this occasion, measure the similarity of pixels as well as a line discrimination step, and the grouping method through measurement of similarity is same Fig. 5.

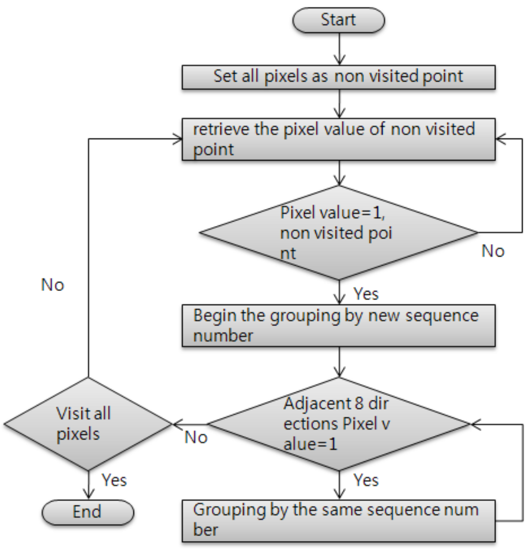

Fig. 5 Character grouping algorithm

Set these all pixels of learning image which is already binarized in a line discrimination step as non visited point, then search the pixel value of non visited. It begins from first pixel of whole image because there is no pixel that visit at first. Next, confirm a pixel value is 1 or not. If the pixel value is 0 , it search again the pixel value of non visited point except visited pixels before in this case. And if the pixel value is 1 which means character , begin grouping with new number .As illustrated to Fig. 4, confirms whether non visited pixel bring pixel value 1 or not among the searching pixels which have 1 pixel value and its adjacen 8 directions (2 or 9) 's pixels.

If there are pixels that pixel value is 1, grouping as the same number. And then search repeatedly the pixel value with the pixels that grouping most recently and their adjacen 8 directions repeatedly to check the value is 1 .

After this process, confirming whether visited these all pixels of learning image which evolve in case that there is no entirely non visited pixel among pixel s of 8 directions which adjoin pixel value 1.If visited all pixels, end the grouping and if there is a non visited pixel, go back to the step of searching non visited pixel and restart the whole process have done so far[11].

Next, if the size of learning image pixel which is grouping is 10X20 or 30X40 , normalize as 20 X30 as the preset size .For example , specifying the preset learning data to the normalized character, specified character will be the learning data such as 'A, B, C,D, …, Z, a, b, c, …, z, 0, 1, 2, …,9'.

Learning data is decided , so it is 'A' to be the grouping by first ,and it is 'B 'to be grouping by second, and it is '9' to be grouping the last. After do grouping of many learning image which is specified to learning , reducing the errors between desire value and output value by using BP algorithm and pass through the process that update connection weight value.

For modeling BP algorithm, input layer, hidden layer ,output layer Learning Rage and Sigmoid must be set.

Number of neuron to come in input layer is pixel value that is equivalent to each character which is grouping before. That is, number of input layer is 20X30 becoming 600. And number of hidden layer sets 900 that is 1.5 times of 600[12]. Though outputs layer should be compared with character which is desire value but it is displayed as binary value so used Uicode (16). Thus, designed Artificial Neural Netwok is same as Fig. 6.

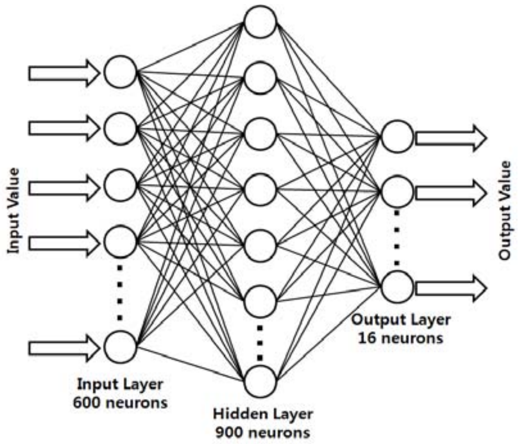

Fig. 6 Modeling of an Artificial Neural Network

For setting the Learning Rage and Sigmoid Slope, executed Variation by giving different values. Table1, Table 2 display the result of Variation.

Table 1: Learning Rate Variation

Learning Rate	**…**	**130**	**140**	**150**	**160**	**…**
Error Value	…	15.76	8.14	1.43	3.61	…

Table 2: Sigmoid Slope Variation Error value

Sigmoid Slope	**…**	**0.010**	**0.012**	**0.014**	**0.016**	**…**
Error Value	…	4.00	2.78	0.05	0.13	…

As the result of table 1, when inputs the value of 150 Learning Rate displayed best results so Learning Rate set as 150. And as the result of table 2, when inputs the value of 0.014 Sigmoid Slope displayed best results so Sigmoid Slope set as 0.014. Learning Rate set as 150 and Sigmoid Slope set as 0.014 to modeled artificial neural network and executed learning. Execute the learning process which was set as critical value of error value by below 0.002 and the maximum number of replication by 10000, then store each equivalent weight value which is updated artificial neural network's connection weight value in data saving part same as Fig. 7.

Weight[1 , 0 , 0] = -14.39005
Weight[1 , 0 , 1] = -302.5181
Weight[1 , 0 , 2] = 46.76268
Weight[1 , 0 , 3] = 240.5509
Weight[1 , 0 , 4] = 22.23077
Weight[1 , 0 , 5] = 59.12272
Weight[1 , 0 , 6] = -878.856

Fig. 7 Connection weight of BP algorithm

4.2 Document style setting step

After saving the style of document that wish to recognize in the document style setting step, save the area that wish to recognize character at the style of document. If take 'Patent specification' for the example, formed areas can be included so that can record contents of patent specification title, inventor, applicant, application number, technical area, the execution example and etc in style of document that wish to recognize. The Style and realization area of this document can set as vector, and the form of vector can express as '[Document title, Data sequence, Bottom- left point, Bottom- right point, Top -left point, Top- right point]'. For example, If the coordinate of position which is the area that wish to recognize in document which is called 'Patent specification 1' has two location such as '(100, 150), (200, 150), (100, 180), (200, 180) ' and '(300, 150), (400, 150), (300, 180), (400, 180)', the form of vector can express by ' [Patent specification 1, 0, 100,150, 200,150, 100,180, 200,180]' and '[Patent specification 1, 1, 300,150, 400,150, 300,180, 400,180]'.

4.3 Character realization step

Fig. 8 is a flowchart that display Character realization step.

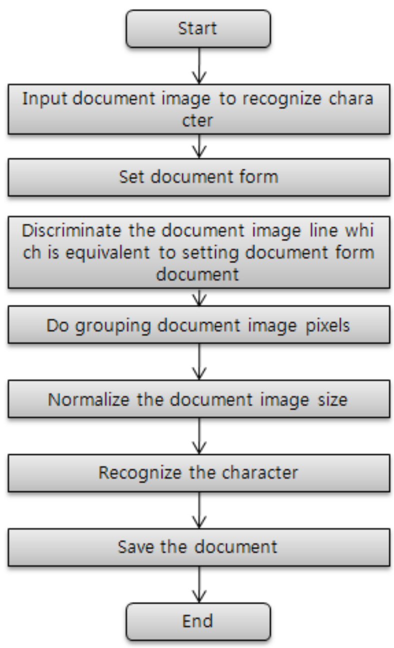

Fig. 8 Flowchart of character realization step

If any document image which is wished to recognize character is inputted through the image register part, reduce and binarize the whole size of inputted document image as the same way that referred during the explain of character learning step.

If the binarized document image comes under a document style which is set at document style setting step, discriminate the document image line and do grouping the document image pixels as the same way of learning image line realization step and learning pixel grouping step.

By using form of vector which was referred in representation of document style setting step can know that the document image is applying to the setting style of the document or not

That is, in case of document style was set by vector form '[Document title, Data sequence, Bottom- left point, Bottom- right point, Top -left point, Top- right point]',it can know recognizing the character applying coordinate of area that is set 'Document title' in inputted document image.

'Document title' area passes through Character realization step and realized the character first, then execute the character recognition which is equivalent to setting character document style under the remainder setting area .

Next, normalize the grouping pixel size of the document same as the way in character learning step. After this process , Among the saved document style recognize the character of document's grouping image pixel which is equivalent coordinate area except 'Document name' At this time, by using the weight value which was generated at the learning character step can compose the same artificial neural network , and input the literal pixel value that wish to recognize. Through this process it can realize the character.

5. Constructed system

The system is constructed by the proposed method, and experimented by scanned documents. System is developed by using Visual C#. Fig. 9 is a screen of system that is doing character learniing. Standard value is set by calling a predeterminate image. And Fig. 10 is a screen that recognize character in document by using saved standard value beforehand.

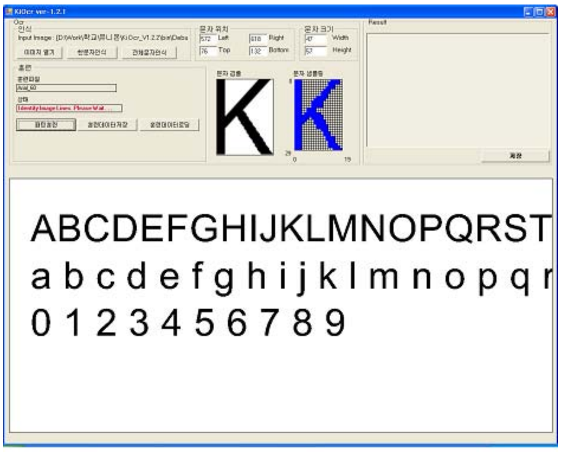

Fig. 9 The screen of character learning by the constructed system

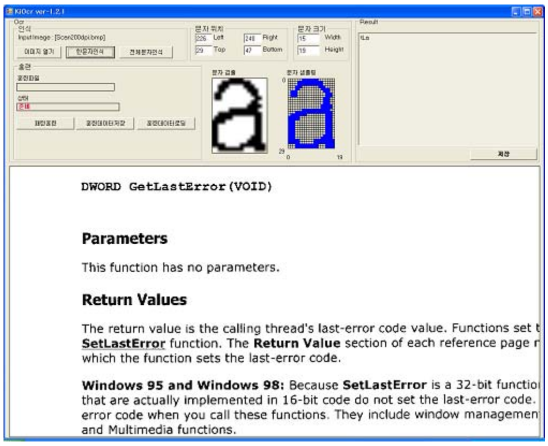

Fig. 10 The screen of character recognition by the constructed system

6. Conclusion

In this paper, in order to recognize the character efficiently which is in mass document constructed the Optical Character Recognition System, and the section that needs for recognizing standardized document style and characters are predesignated. Optical Character Recognition System that is constructed by using artificial neural net work algorithm expects to be efficient in character recognition of mass standardized document. An experiment that measure accuracy and efficiency of developed system in paper has to be achieved by using various document more than hereafter.

Acknowledgments

- This work was supported by the Brain Korea 21 Project in 2008.

- This research was supported by the MKE(Ministry of Knowledge Economy), Korea, under the ITRC(Information Technology Research Center) support program supervised by the IITA(Institute of Information Technology Advancement). (IITA-2008-(C1090-0801-0025))
- This work was supported by the IT R&D program of MKE/IITA. [2007-S019-02, Development of Digital Forensic System for Information Transparency]

Corresponding Author

- E-mail address : jang@korea.ac.kr (Dong-Sik Jang)

References

[1] K. Fukushima, "A Neural Network for Visual Pattern Recognition, " IEEE Computer, Vol. 21, No. 3, March 1988.
[2] W. Y. Huang and R. P. Lippmann, "Comparisons between Neural Net and Conventional Classifiers," in Proceedings of the IEEE ICNN, Vol. IV, pp.485-493, 1987.
[3] Mardia, Kanti V. "Statistics and Images", Vol 1, Vol 2Cartax Publishing Company, Oxford, UK, 1995.
[4] T. Matsuoka, H. Hamada and R. Nakatsu, "Syllable Recognition Using Integrated Neural Networks," in Proceedings of the IEEE IJCNN, Vol. I, pp. 251-258, 1989.
[5] A. Waibel, "Connectionist Glue: Modular Design of Neural Speech Systems," in Proceedings of the 1988 Connectionist Models Summer School, pp. 417-425, 1988.
[6] J. L. Mcclelland and D.E. Rumelhart, "Learning Internal Prepresentation by Error Propagation,", Parrallel Distributed Processing, Vol. 1, 1986.
[7] Chaudhuri, B.B., Pal, U., "A complete Bangla OCR system", Pattern Recognition, Vol. 31, pp. 531–549, 1998.
[8] Cash, G. and Hatamian, M. "Optical character recognition by the method of moments". Computer Vision, Graphics, and Image Processing, Vol. 39, pp. 291-310, 1987.
[9] De Luca, P. and Gisotti, A. "Printed character preclassification based on word structure". Pattern Recognition, Vol. 24, pp. 609-615, 1991.
[10] R.M.K. Sinha, et al., "Hybrid contextual text recognition with string matching", IEEE PAMI, Vol. 15, pp. 915–923, 1993.
[11] J. L. Mcclelland and D.E. Rumelhart, "Learning Internal Prepresentation by Error Propagation,", Parrallel Distributed Processing, Vol. 1, 1986.
[12] J. Fournier, M. Cord, S. Philipp-Foliguet, "Back-Propagation Algorithm for Relevance Feedback in Image Retrieval", IEEE, 2001.

Sang Sung Park received the Ph.D. degree in industrial system and information engineering from Korea University. He is a Research Professor of Ubiquitous Information Security Research Division at Korea University. His research interests include computer vision, pattern classification and knowledge management.

Won Gyo Jung received his B.S. degree in industrial engineering from Kyunghee University. He is currently a M.S. candidate in Division of Information Management Engineering at Korea University. His research interests include pattern classification, e-business, information system and artificial intelligence.

Young Geun Shin received his B.S. degree in industrial system and information engineering from Korea University. He is currently a Integrated M.S. and Ph.D. andidate in Division of Information Management Engineering at Korea University. His research interests include pattern classification, scheduling and artificial intelligence.

Dong Sik Jang received the Ph.D. degree in industrial and systems engineering from the Dwight Look College of Engineering, Texas A&M University, in 1988, the M.S. degree in Operations Research and Industrial Engineering from The University of Texas at Austin, in 1985, and the B.S. degree in Industrial Engineering from Korea University, in 1979. He is a Professor of Division of Information Management Engineering at Korea University. His research interests include computer vision, optimization theory and computer algorithm.