거리 사상 함수 및 RBF 네트워크의 2 단계 알고리즘을 적용한 서류 레이아웃 분할 방법

(A Two-Stage Document Page Segmentation Method using Morphological Distance Map and RBF Network)

신 현 경 †

(Shin Hyun Kyung)

요 약 본 논문에서는 2 단계 서류 레이아웃 분할 방법을 제안한다. 서류 분할의 1 차 단계는 top-down 계열의 영역 추출로서 모폴로지 기반의 거리 함수를 사용하여 주어진 영상 데이타를 사각형 영역들로 분할한다. 거리 사상 함수를 통한 예비 결과는 성능 개선을 위한 2 차 단계의 입력 변수로 작용한다. 서류 분할의 2 차 단계로서 기계 학습 이론을 적용한다. 통계 모델을 따르는 RBF 신경망을 선택하였고, 은닉 층의 설계를 위해 코호낸 네트워크의 자기 조직화 성격을 활용한 데이타 군집화 기법을 기반으로 하였다. 본 논문에서는 300 개의 영상에서 추출된 영역 데이타를 통해 학습된 신경망이 1 차 단계에서 도출된 예비 결과를 개선함을 연구 결과로 제시하였다.

키워드 : RBF 망, 자기 조직화 사상, 서류 페이지 레이아웃 분할

Abstract We propose a two-stage document layout segmentation method. At the first stage, as top-down segmentation, morphological distance map algorithm extracts a collection of rectangular regions from a given input image. This preliminary result from the first stage is employed as input parameters for the process of next stage. At the second stage, a machine-learning algorithm is adopted. RBF network, one of neural networks based on statistical model, is selected. In order for constructing the hidden layer of RBF network, a data clustering technique based on the self-organizing property of Kohonen network is utilized. We present a result showing that the supervised neural network, trained by 300 number of sample data, improves the preliminary results of the first stage.

Key words : radial basis function network, self-organization map, and document page segmentation

1. Introduction

Document page segmentation, as a definition, is an image analysis technique partitioning a page of document image into the separate regions in accordance with geometrical and logical (semantic) relations among their textual contents [1]. Document segmentation is a low-level process as an important part of TIE (textual information extraction) project. Consequently, performance of TIE from images (or videos) strongly depends on the outputs of document segmentation. Once a page of document is partitioned into the sub-regions, each sub-region can be further categorized as either textual or non-textual region by a method based on the analysis of its content. The texts included in a textual region can be converted to ASCII through OCR [2] and then can be understood semantically through text processing [3].

Several research groups related to TIE project have addressed a variety of techniques on document page layout analysis. Most of the methods

† 정 회 원 : 경원대학교 수학정보학과 교수
hyunkyung@kyungwon.ac.kr
논문접수 : 2008년 5월 20일
심사완료 : 2008년 9월 1일

정보과학회논문지 : 소프트웨어 및 응용 제35권 제9호(2008.9)

can be classified as either the region based top-down approach or the pixel based bottom-up approach. The bottom-up method normally uses statistical information of local neighborhood of a pixel under consideration. Strength of this method lies in correct result of layout analysis while weakness lies in complexity of its algorithms and heavy intensity of computation. Haralick et al. [4] uses the small bounding boxes around the textual entities identified by connected components. Gupta et al. [5] utilizes the bounding boxes for pre-processing of document layout analysis. Kim et al. [6] used cluster-based templates and K-mean clustering algorithm. On the other hand, the top-down method utilizes rather global information of the white strip regions within a page. Haralick et al. [7] uses X-Y min-cut to decompose document image recursively. Shi and Malik [8], Borenstein and Ullman [9] also use X-Y min-cut. Some of researchers focus on white regions (or background) in document image. Baird [10] and Antonacopoulos [11] use the white tiling method. Lee and Ryu [12] suggest a method of multi-resolution structural analysis using conventional pyramid of quad-tree.

Recently, the practice of machine-learning algorithm [13] for document understanding is one of the extremely appealing and challenging tasks. Rosenfeld et al. [14] introduced mixed learning algorithm method. Chen et al. [15] adopted a method of unsupervised clustering to improve drawbacks from supervised learning.

We utilize a two-stage approach: morphological top-down method and machine-learning method. At the first stage, a distance map algorithm on the re-sampled image is used. At the second stage, with the preliminary results from the top-down approach, self-learning technology of machine-learning algorithm divides or merges the segmented regions. For the learning algorithm, we adopt a RBFN with unsupervised self-organization. Our two-stage method is designed to use flexible interface with a generic form of the statistical decision models including neural network. The first stage of segmentation can be regarded as a process of assigning a-priori probability, and the second stage of segmentation can be regarded as a process of determining posteriori probability.

Contents of this paper are organized as follow: In section 2, the distance map method for top-down document layout analysis is described. In order for improving the results of the first stage of document segmentation as described in section 2, the second stage of document segmentation using machine learning technique is explained in the section 3. The radial basis function network is selected as a machine learning algorithm for document page segmentation. In section 4, experimental results of the study performed in this paper will be presented. In section 5, discussion on the technology applied in this paper will be described.

2. Distance Map Method for Top-Down Segmentation

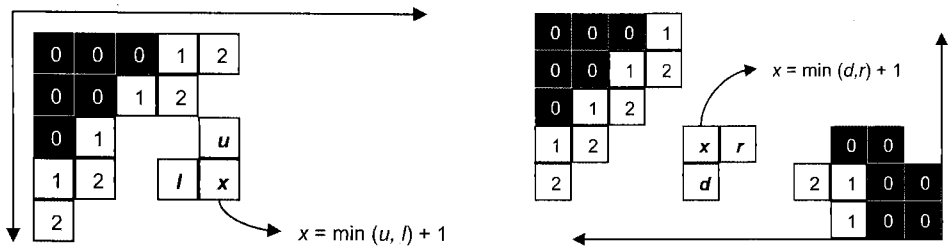
Our definition of the distance map is the following: given a black and white bi-tonal image, a distance map is a labeled image on the white pixels, where the value of label indicates distance to the nearest black pixel. The top-down segmentation method adopted in this paper is non-iterative (two-stage) distance map construction. Before applying distance map algorithm, as initialization, the black pixels and the boundary pixels are assigned to 0 and 0xFF, respectively.

At the first stage of distance map algorithm, as seen in Figure 1(a), input image is swept from top-left corner of image to bottom-right corner. While sweeping the image, for the pixel having non-zero value, the algorithm labels a new value x to the pixel under consideration using the following formula

$$x = min\ (x_u,\ x_l) + 1 \qquad (1)$$

where x_u, x_l indicate the pixel values at north and west side of given pixel, respectively. It should be mentioned that the algorithm ignores the pixels of zero value (the black pixels). Directional bias for sweeping used at the first stage is obvious. In order for resolving the bias, we apply another sweep with opposite direction, which is the second stage of our distance map algorithm. Refer to Figure 1(b).

The resulting labeled image after the stage 1 and 2 is called the distance map. Once a distance map



(a) stage 1 - from top-left to bottom-right (b) stage 2 - from bottom-right to top-left

Figure 1 two-stage distance map construction

is constructed, an integer-valued image, it is then converted into a binary image by applying conventional threshold method. Refer to Figure 2(a)(b)(c). We apply a connected components algorithm on the binary image obtained from the distance map. The bounding boxes of the connected components, as seen in Figure 2(d), are considered as the segmented regions.

Choice of the value for threshold is important in the sense that the number of segmented regions is sensitively related with it. In our system, the maximum and the minimum numbers of the resulting segmented regions are the user inputs. Once the number of resulting segmented region is out of the range, the control process is designed to adjust the value of threshold iteratively.

For the purpose of rapid computation, we adopted the image pyramid method: the distance map algorithm is applied on the re-sampled image (2×2 or 4×4, depending on the size of original input image). The pyramidal method is legitimate, since the top-down method is global processing insensitive to minor details of edge configuration of black pixels. Our distance map algorithm is presented in Figure 3:

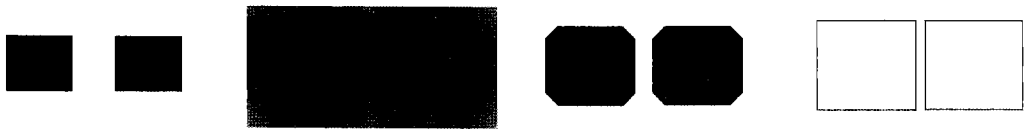
We present the results of top-down segmentation using the distance map algorithm in section 5.

3. RBFN(Radial Basis Function Networks)

The distance map method described in section 2 is a simple classifier concerning only a value of distance to the nearest black pixel. Obviously a more complicated classifier having higher degree of freedom is necessary. We adopt neural network framework to improve the results obtained via the distance map method. The main purpose of this section is to propose a new feature selection method which enables problem of document segmentation to be feasible by supervised learning.

In order for implementing supervised learning paradigm, RBFN is adopted among others. It has two main advantages [16,17]. First of all, it has a simple structure of synapses in terms of updating weight values: two layers of synapse (one from an input neuron layer to a hidden neuron layer and the other from a hidden neuron layer to an output neuron layer) are updated sequentially. Secondly, the process of updating weights on the synapses is simple: in case of regularization method, the learning process can be achieved by one-time direct computation; in case of gradient descent method, simplicity of synaptic layer structure allows faster computation. We select the regularization method.

An overview of RBF networks is shown in Figure 4. The networks are consisted of three neuron layers and two synaptic layers. Three neuron layers (the input, the hidden, and the output) are



(a) input image (b) distance map (c) thresholded image (d) segmentation

Figure 2 a simple example of a top-down segmentation using the distance map

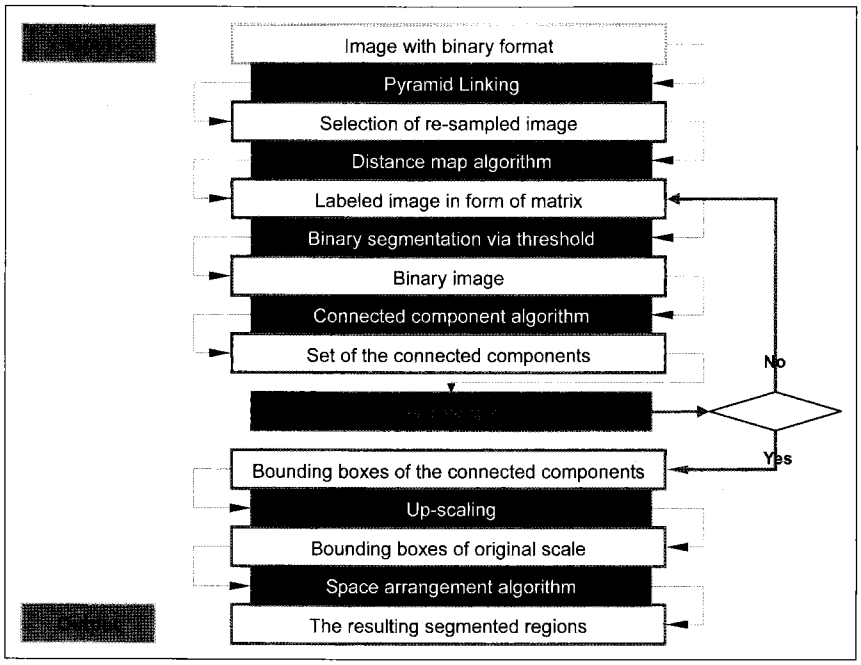


Figure 3 Flow chart of top-down document page segmentation using distance map algorithm. M and m denote maximum and minimum number of bounding boxes allowed, respectively

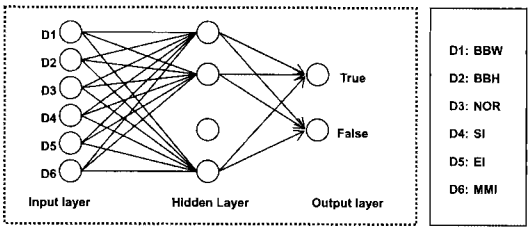


Figure 4 a structural overview of RBFN: the number of neurons allocated in the hidden layer is assigned dynamically via self-organization strategy. The acronyms in the legend are explained in section 3. 2

linked sequentially by the synapse layers.

3.1 Design of the Hidden Layer: Clustering

Determination of a number of neurons in hidden layer is one of the two most critical phases in design of RBFN (the other is synaptic weight update scheme). We pick a self-organization strategy among the others [17]. The main reason for opting self-organization is two-folded. The one is to avoid a bias from initialization, and the other is to deal efficiently with the high dimensional input data. For the self-organization, we applied the Kohonen network [18,19] as described as follows. With the training data set, Kohonen self-organizing map automatically (without supervision) divides collection of the data into a reasonable number of clusters. The two-dimensional Kohonen layer used in Kohonen self-organization map is designed to have 10-by-10 lattice of nodes. Once the centers of clusters are identified, a post-process determines the centers of clusters through analysis on the size of clusters. The final result of post-processing gives the exact number of clusters which is in turn used to build the hidden layers.

3.2 Design of Input Layer: Feature Selection

Formation of an input neuron layer depends on feature selection process. We create 6-dimensional string vector in which each component represents as follows:

- D1 represents a width of bounding box.
- D2 represents a height of bounding box.
- D3 represents a number of rows contained in the bounding box.
- D4 represents an amount of first-line indent.

The Infancy of the ROK Army Korea

Maj. Robert K. Sawyer, USA

As the U.S. troops left Korea during late 1948 and the first half of 1949, they turned over part of their equipment to the ROK forces in accordance with the military agreement signed by Rhee and Hodge. By November 1948 between 60 and 80 percent of the Constabulary's small arms and automatic weapons were American, but there was a dearth of mortars and heavy machine guns. Although only 52 of an allotted 90 105mm infantry cannon (M3) had been received, a full issue of 37mm anti-tank guns was on hand. Unfortunately, spare parts and all types of sighting and aiming equipment were in extremely short supply.

During the spring of 1948, the South Koreans had added six infantry regiments to the original nine formed for the Constabulary and had created a fourth and a fifth brigade. They wanted to form a sixth brigade but were warned that the United States could not furnish equipment for further expansion at that time.

Disturbance

In October 1948 the fledgling government and Army had undergone its sternest test. Domestic disturbances were common throughout South Korea, but the mutiny of the ROK 14th Regiment at Yosu on 19 October represented a serious threat. The uprising spread quickly to the towns of Sunch'on, Posong, Polgyo-ri, and Kwangyang. The American mission in Korea felt that Korea's future depended upon Rhee and his associates bringing the dissident elements under control quickly and firmly.

The Yosu revolt appears to have been prematurely touched off by an order issued in October assigning the 14th Regiment to the island of Cheju-do to help quell disorders there. U.S. M1 rifles were issued to the regiment, though it retained its Japanese 99s as well. Among the high-ranking noncommissioned officers of the regiment were a number of Communists, and they apparently planned to use the extra rifles to arm fellow travelers in the surrounding villages. Then came a second order instructing the regiment to move at once, and the Communist leaders had to act hastily or lose the opportunity for concerted action and for arming their coworkers. They fanned the anti-police sentiment among the troops of the 14th and incited an attack to take over the town.

Capt. Hausman

As the rebellion spread to surrounding towns, loyal Constabulary units and police detachments were rushed from districts throughout southwest Korea. Brig. Gen. Song Ho Seung, the chief of the Constabulary, was placed in command of the government forces, and General Roberts sent Captain Hausman and two other advisors along to assist Song. The Korean general and the advisors set up headquarters at Kwangju. Soon five additional advisors joined the group, including two lieutenants who had been imprisoned by the rebels in Yosu and later escaped.

Capt. Hausman

As the operation grew in scope and seriousness and because of the extremely poor communications between dispersed Constabulary units, a need for additional advisors arose. Coulter and Roberts sent a senior officer, Col. Hurley E. Fuller, to take over the responsibilities.

In the meantime, Capt. Hausman and Capt. John P. Reed managed to round up part of a Constabulary regiment and on 21 October halted the rebel drive until local reinforcements could arrive. A counterattack rewon Sunch'on on 22 October, and two days later the local forces entered Kwangyang and Posong. Polgyo-ri fell on the 25th, and the rebels then made their final stand at Yosu. After two days of fierce fighting, the opposition ended, many of the mutineers slipped off to the rugged mountains to the north. Here, in the Chiri-san area, the rebel remnants became guerrilla fighters and a constant thorn in the side of the government.

The effort made by Capt. Hausman and his fellow advisors in rallying the Constabulary and aiding in the conduct of the campaign was first rate. Without their assistance the revolt might well have become far more serious. The Yosu rebellion was also important in providing a training ground for the commanders, staff, and units of the Constabulary. Although anti-guerrilla activity earlier in 1948 at Cheju-do had given the Constabulary valuable experience, the Yosu affair allowed units to be employed for the first time on something approaching a major scale.

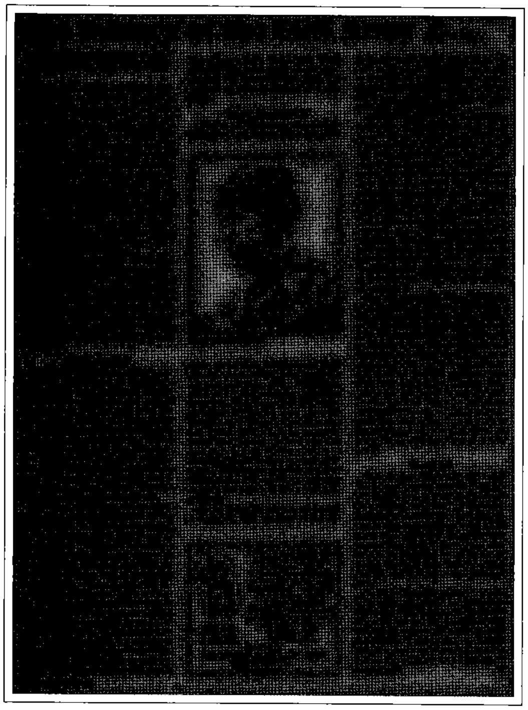
The 14th

Almost immediately after the fall of Yosu, repercussions began throughout the ROK forces. A purge to eliminate Communist influence in the Constabulary got under way and, according to General Roberts, over 1,500 were uncovered and removed from the service. Except for a brief outbreak of Communist activity at Taegu that was soon brought under control, the Constabulary remained loyal.

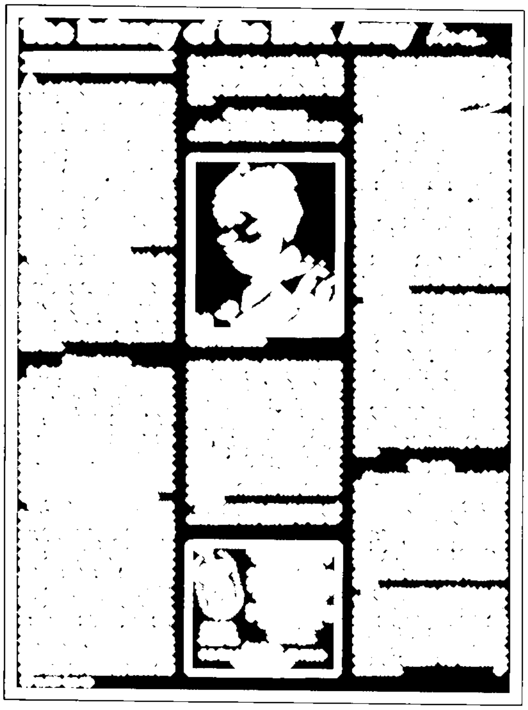
In an interesting sidelight to the Yosu affair, the ROK leaders abolished the 14th Regiment on 28 October for its shameful conduct and burned its colors. All units bearing the number 4, either alone or in combination, were redesignated and the use of the number was henceforth abandoned.

20 MILITARY

(a) Original document image



(b) Labeled image by distance map



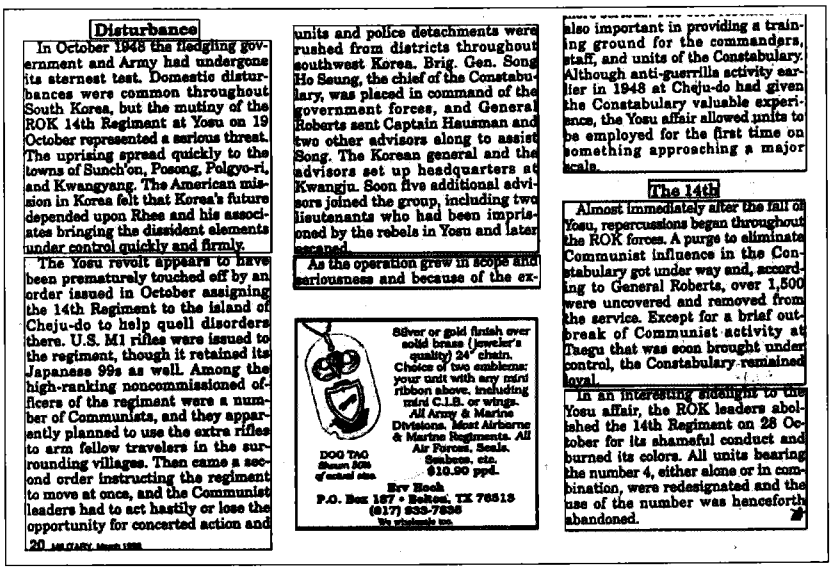
(c) Threshold operation on the labeled image

The Infancy of the ROK Army Korea

Maj. Robert K. Sawyer, USA

(d) Segmentation result

Figure 5 a result of document segmentation by a top-down method based on the distance map.



A result of document segmentation based on the RBFN

Figure 6 a result of document segmentation after applying self-learning algorithm. The results shown in Figure 5-D are improved

- D5 represents a distance from the end of last row to right end of the bounding box.
- D6 represents a ratio of the longest to the shortest intervals among the connected components inside a bounding box.

4. Experimental Results

For the purpose of visualization, we present a series of outputs obtained from the segmentation processes using the distance map method. Refer to Figure 5: in panel A, the original input image having color format is shown; in panel B, an output of distance map; in panel C, a result of threshold operation; in panel D, the final result of segmentation is presented.

In Figure 6, the output of the second stage is presented. For clarity of presentation, we only show the lower half part of the resulting image. Evidence of improvement on segmentation is shown: in Figure 5-D, the third block from the top and the first one from the left containing a word 'Disturbance' has successfully divided into two sub-regions in Figure 6. The word 'Disturbance' now is isolated from the rest of phrase and recognized as a title of section. Moreover, the body of phrase is also divided into the two semantically independent phrases.

However, as seen at the bottom of blocks in the second column (containing the region of drawing of 'dog tag'), we could not separate between the region of drawing and the region of word phrase.

5. Conclusion

Two-stage document segmentation has two main purposes: fast processing speed and robustness. Rapid computation is provided by the top-down method adopted at the first stage, and robustness is given by the self-learning method assumed at the second stage. Given an image of size n × n in pixels, the exact number of operations for the distance map algorithm is $5n^2 = 2 \times 2 \times n^2$ (two- time sweeps of morphological dilations × minimum operation and addition × size of image in pixel) + n^2 (threshold operation). In order to reduce the computational cost, we applied multi-resolution pyramid linking method. For this purpose, we constructed three layers of re-sampled image-the operation numbers are 21/16 n^2. Consequently, the number of operations for the distance map with re-sampled image is $5r^2$ where r is n/4 or n/8 depending on the resolution of original image. The total number of operations is in between

26/16 n^2 and 89/64 n^2, which shows great reduction from $5n^2$. For the identification of connected component, we implemented the Floyd-Warshall labeling algorithm- complexity is 16 m where m is the number of black pixels. Typically the number of black pixels in document image is less than 0.1 n^2. As a result, the computational cost for the first stage is approximately 2 n^2.

Robustness of segmentation results relies on the adaptive property of machine learning algorithms. The robustness here is meant by continuity of segmentation outputs on the input images. The dimension of input vector (feature selection data) is the key parameter on this issue. The greater the dimension guarantees the better robustness. We investigated on statistical independence between the components of the input vector (refer to section 3.2 for the details), and decided to use 6-dimensional vector space. We trained the network using 300 training data.

We should mention that the top-down method using a distance map algorithm often shows quite good document segmentation results, refer to Figure 5-D, especially when the input image does not contain complex layout.

References

[1] Haralick, R. M., "Document Image Understanding: Geometric and Logical Layout," CVPR94: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 385-390, 1994.

[2] Ferrer, M., and Valveny, E., "Combination of OCR Engines for Page Segmentation Based on Performance Evaluation," Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), pp. 784-788, Vol. 2, 2007.

[3] Wu, V., Manmatha, R. and Riseman, E. M., "Finding Text in Images," ACM, DL, pp. 3-12, 1997.

[4] Liang, J., Ha, J., Haralick, R. and Phillips, I., "Document layout structure extraction using bounding boxes of different entities," Proceedings Third IEEE Workshop on Applications of Computer Vision (Sarasota, FL), pp. 278-283, 1996.

[5] Gupta, G., Niranjan, S., Shrivastava, A., and Sinha, R.M.K., "Document Layout Analysis and Classification and Its Application in OCR," 10th IEEE International Enterprise Distributed Object Computing Conference Workshops, 2006.

[6] Kim, E., Jung, K., Jeong, K., and Kim, H., "Automatic Text Region Extraction Using Cluster-based Templates," Proc. of International Conference on Advances in Pattern Recognition and Digital Techniques, pp. 418-421, 2000.

[7] Ha, J., Haralick, R. and Phillips, I., "Recursive X-Y Cut Using Bounding Boxes of Connected Components," Proc. Third Int'l Conf. Document Analysis and Recognition, pp. 952-955, 1995.

[8] Shi, J., and Malik, J., "Normalized Cuts and Image Segmentation," IEEE Conference on Computer Vision and Pattern Recognition, pp. 731-737, 1997.

[9] Borenstein, E. and Ullman, S., "Class-specific top-down segmentation," In Proc of the 7th European Conference on Computer Vision, Copenhagen, Denmark, pp. 109-124, 2002.

[10] Baird, H. S., "Background Structure in Document Images," Document Image Analysis, pp. 17-34, 1994.

[11] Antonacopoulos, A., "Page segmentation using the description of the background," Computer Vision and Image Understanding, 70(3):350-369, June, 1998.

[12] Lee, S. and Ryu, B., "Parameter-Free Geometric Document Layout Analysis," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 23, no. 11, pp. 1240-1256, Nov. 2001.

[13] Hassoun, M. H., "Fundamentals of Artificial Neural Networks," The MIT Press, Cambridge, MA, 1995.

[14] Rosenfeld, B., Feldman, R., and Aumann, Y., "Structural extraction from visual layout of documents," CIKM '02: Proceedings of the eleventh international conference on Information and knowledge management, pp. 203-210, 2002.

[15] Chen, S., Mao, S., and Thoma, G., "Simultaneous Layout Style and Logical Entity Recognition in a Heterogeneous Collection of Documents," Ninth International Conference on Document Analysis and Recognition, Vol. 1, pp. 118-122, 2007.

[16] Bishop, C. M., "Neural Networks for Pattern Recognition," Oxford University Press, Oxford, UK, 2003.

[17] Haykin, S., "Neural Networks - A Comprehensive Foundation," Prentice Hall, Upper Saddle River, NJ, 1999.

[18] Kohonen, T, "Exploration of very large databases by self-organizing maps," 1997 International Conference on Neural Networks, vol. I, pp. PL1-PL6, Houston, 1997.

[19] Kohonen, T., "Self-Organizing Maps," 2nd edition, Springer-Verlag, Berlin, 1997.

신 현 경

2002년 State University of New York (Stony Brook) 대학원 응용수학과(공학박사). 2008년 현재. 경원대학교 수학정보학과 조교수. 관심분야는 Neural network. Machine learning. Image processing