# Parameter-Free Geometric Document Layout Analysis

Seong-Whan Lee, *Senior Member*, *IEEE*, and Dae-Seok Ryu

**Abstract**—Automatic transformation of paper documents into electronic documents requires geometric document layout analysis at the first stage. However, variations in character font sizes, text line spacing, and document layout structures have made it difficult to design a general-purpose document layout analysis algorithm for many years. The use of some parameters has therefore been unavoidable in previous methods. In this paper, we propose a parameter-free method for segmenting the document images into maximal homogeneous regions and identifying them as texts, images, tables, and ruling lines. A pyramidal quadtree structure is constructed for multiscale analysis and a periodicity measure is suggested to find a periodical attribute of text regions for page segmentation. To obtain robust page segmentation results, a confirmation procedure using texture analysis is applied to only ambiguous regions. Based on the proposed periodicity measure, multiscale analysis, and confirmation procedure, we could develop a robust method for geometric document layout analysis independent of character font sizes, text line spacing, and document layout structures. The proposed method was experimented with the document database from the University of Washington and the MediaTeam Document Database. The results of these tests have shown that the proposed method provides more accurate results than the previous ones.

**Index Terms**—Geometric document layout analysis, parameter-free method, periodicity estimation, multiscale analysis, page segmentation.

## 1 Introduction

With the advance in information technology and the increased need for information, the volume of documents containing information has increased more and more. In spite of the use of electronic documents, the amount of paper documents has never decreased because the publication of newspapers, periodicals, reports, books, etc., has increased continuously and most human beings prefer paper documents for reading and archiving. However, it is very difficult and cumbersome to store and retrieve the ever increasing number of paper documents. On the contrary, electronic documents have several advantages in storage, retrieval, and updating. If paper documents could be converted to electronic documents automatically, it would be possible to search for the contents of the documents in seconds and to retrieve and update them efficiently. However, it is not trivial to automatically transform paper documents into electronic format. First of all, to achieve this transformation, geometric document layout analysis should be performed. Geometric document layout analysis involves specifying the geometry of the maximal homogeneous regions and classifying them into text, image, table, drawing, etc. As shown in Fig. 1, each region, which is segmented and classified by geometric document layout analysis, could be processed efficiently as follows: Text regions are converted to ASCII codes by Optical Character Recognition (OCR), images are compresse, tables are reconstructed, and drawings are vectorized.

In this paper, we propose a novel method for geometric document layout analysis, with which it is possible to segment document images into maximal homogeneous regions and to identify them as texts, images, tables, and ruling lines. The main advantage of the proposed method is the ability to segment the document images having various character font sizes, text line spacing, and document layout structures by the suggested periodicity measure, multiscale analysis, and confirmation procedure, independently of any parameters.

This paper is organized as follows: In Section 2, previous works related to geometric document layout analysis and page segmentation are reviewed and the main ideas of the proposed method are described. In Section 3, a parameter-free method for page segmentation and region identification is proposed. In Section 4, to verify the performance of the proposed method, experimental results with the document database from the University of Washington [17] and the MediaTeam Document Database [18] are analyzed. Finally, conclusions and further research directions are given in Section 5.

## 2 Background

### 2.1 Previous Works

A number of approaches have been proposed in various ways for page segmentation and geometric document layout analysis.

Traditional approaches for page segmentation and geometric document layout analysis are typically referred

- *S.-W. Lee is with the Center for Artificial Vision Research, Korea University, Anam-dong, Seongbuk-ku, Seoul 136-701, Korea. E-mail: swlee@image.korea.ac.kr.*
- *D.-S. Ryu is with Hyundai Information Technology, Co. Ltd., San 1-8, Mabuk-ri, Gusung-Myun, Yongin-si, Kyunggi-do, 449-910, Korea. E-mail: ryu75@shinbiro.com.*

*Manuscript received 25 Jan. 2000; revised 28 Dec. 2000; accepted 11 Apr. 2001.*
*Recommended for acceptance by D. Dori.*
*For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number 111305.*

0162-8828/01/$10.00 © 2001 IEEE
Authorized licensed use limited to: Korea University. Downloaded on October 25, 2008 at 01:45 from IEEE Xplore. Restrictions apply.

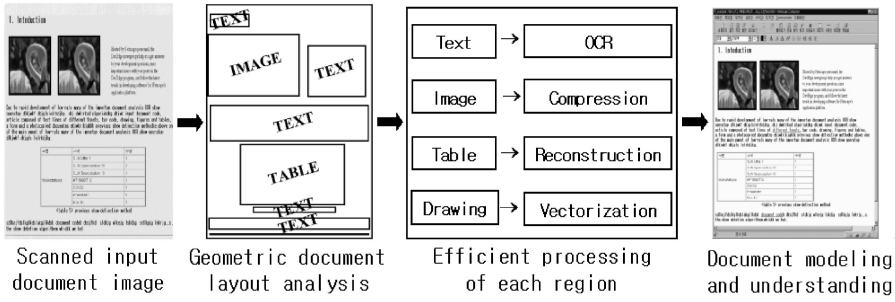



Fig. 1. The role of geometric document layout analysis.

to as bottom-up methods [1], [2], [3]. These methods start with local information (concerning the black pixels or the connected components), determine the words, merge the words into text lines, and merge the text lines into paragraphs. A well-known method in bottom-up approaches is the connected components grouping method [2], where connected components are extracted from the image. Then, components of the same type are iteratively grouped together to form progressively higher-level descriptions of the printed regions of the document (words, text lines, paragraphs, etc.). A disadvantage of this approach is that the identification, analysis, and grouping of connected components are, in general, time-consuming processes, especially when there are many components in the image (e.g., a newspaper page). To reduce this computational complexity, a new bottom-up layout analysis method [3] has been presented. It detects the layout hierarchy as a minimal-cost spanning tree and additional heuristics have been applied. But, this algorithm may suffer from the traditional problems of bottom-up strategies, such as incorrect segmentation due to wrong early groupings.

On the other hand, top-down approaches look for global information on the page, e.g., black and white stripes, and on the basis of this, split the page into columns, the columns into blocks, the blocks into text lines, and the text lines into words. The advantages of these approaches are that the time complexity is lower than that of bottom-up approaches and it is natural for human beings to see an object from a coarse to fine resolution. However, with the previous methods [4], [5], the complex document layout which is composed of nonrectangular images and various character font sizes makes it difficult to segment correctly in a top-down manner.

A lot of page segmentation and geometric document layout analysis methods regard a homogeneous region such as text, image or graphic, in a document image as a textured region. Then, page segmentation is implemented by finding textured regions in gray scale images. One major problem associated with such texture-based approaches [6], [7] is that the time complexity is too high since different filters are tuned to capture a desired local spatial frequency and the orientation characteristics of a textured region, so that many masks are used for extracting local features. If a small mask is chosen, it is difficult to detect large-scale textures such as large fonts. On the contrary, if a large mask is chosen, the computational cost will increase dramatically. To avoid this problem, the multiscale analysis has been applied in recent researches [8], [9]. In these methods, local characteristics can be extracted at each resolution with only two filters. However, the process of getting multiscale texture information is quite time-consuming and, in some cases, regions of different types but with similar texture can be confused or merged.

In addition to the above approaches which share some common features, Antonacopoulos [10] used white tiles to extract contours of regions and features for classification. Jain and Yu [11] used the traditional bottom-up approach for page segmentation and region identification and employed a document model to preserve top-down generation of information for logical layout analysis. However, one problem associated with these methods is the dependence on thresholds and parameters, so adaptively setting the various thresholds and parameters has been unavoidable.

### 2.2 Motivation of this Work

Variations in character font sizes, text line spacings, and document layout structures have made it difficult to design a general-purpose document layout analysis algorithm [11]. Therefore, the use of some parameters has been unavoidable in earlier methods, as discussed in a previous section. To overcome this problem, we present a parameter-free method in which we are able to segment a document image, even though various character font sizes, text line spacing, and document layout structures may exist.

Human beings, at first glance, see an object in a low resolution and then, gradually look at it closer in a higher resolution. In order to simulate this, we construct a pyramidal quadtree structure and analyze it in a top-down manner. The other point of view is that text regions can be easily distinguished from the other regions by the property of being aligned horizontally or vertically. To find this property of text regions, we concentrate on extracting the periodical feature of text regions and segment document images by using the suggested periodicity measure. However, there are some cases that make it difficult to segment correctly using only the suggested periodicity measure, therefore, in those cases, we confirm their segmentation results by their texture analysis.

Consequently, in this paper, we make a pyramidal quadtree structure for multiscale analysis and a top-down approach and suggest a periodicity measure, which can take out the periodical attribute of a region, and the suggested periodicity measure is applied to document segmentation. For obtaining more convincing segmentation results, a confirmation procedure using texture analysis is applied only to ambiguous regions.

Authorized licensed use limited to: Korea University. Downloaded on October 25, 2008 at 01:45 from IEEE Xplore. Restrictions apply.

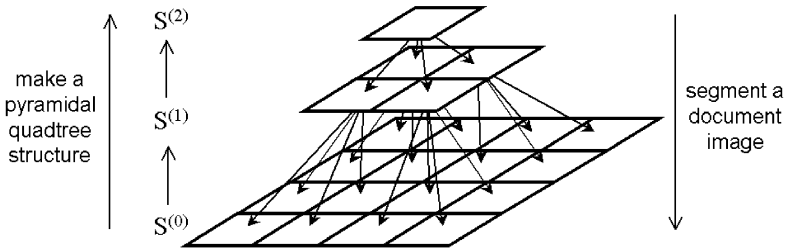



Fig. 2. The pyramidal quadtree structure in the case of three levels.

## 3 Proposed Parameter-Free Geometric Document Layout Analysis Method

The proposed parameter-free method for page segmentation and region identification is described as follows: Given a binary document image,

1. Construct a pyramidal quadtree structure and get multiscale images of the input document image (Section 3.1).
2. Extract bounding boxes of connected components [5] for the top-level image.
3. Estimate the periodicity of each region along the horizontal and vertical directions at each level except the top level and get the periodical attribute of a region (Section 3.2).
4. If a region is not a single periodical region, find a position for splitting and split it into two regions (Section 3.3).
5. Repeat Steps 3-4 until the region becomes a single periodical region.
6. Confirm the segmentation results (Section 3.4).
7. Identify each maximal homogeneous region as text, image, table, and ruling line (Section 3.5).

### 3.1 Construction of Pyramidal Quadtree Structure

As previously mentioned, human beings at first glance see an object in a low resolution and then, gradually look at it closer in a higher resolution. In order to simulate this, the pyramidal quadtree structure is employed. The pyramidal quadtree structure is made by reducing the resolution of an input image by a quarter repeatedly while its size is larger than $50 \times 50$, and then execute a page segmentation algorithm from the lowest to the highest resolution. Each pixel $s^{(n)}$ at scale $n$ corresponds to a group of four pixels $s_i^{(n-1)}(i = 1, 2, 3, 4)$ at the finer scale $n-1$, where $s_i^{(n-1)}$ are the children of $s^{(n)}$. If any one of four pixels $s_i^{(n-1)}(i = 1, 2, 3, 4)$ is black, its parent pixel $s^{(n)}$ is assigned to black. Its equation is described below, where black pixel is 1 and white pixel is 0.

$$s^{(n)} = \bigcup_{i=1}^{4} s_i^{(n-1)}. \tag{1}$$

Therefore, the number of pixels in $S^{(n)}$ is one quarter of the number of pixels in $S^{(n-1)}$. In this way, the pyramidal quadtree structure is constructed, as shown in Fig. 2. As a result of this procedure, the multiscale images shown in Fig. 3 are obtained.

### 3.2 Periodicity Estimation

Text regions can be easily distinguished from the other regions by the property that text lines of a paragraph are aligned horizontally or vertically and have almost the same line spacing. If we apply this property of text regions to page segmentation, we could easily segment not only text regions but also the other regions into each homogeneous area. For this purpose, we concentrate on extracting the periodical attribute of a text region and we call it a periodicity.

In order to extract the periodicity of a region, we first find the page skew angle ($\theta$) by a skew detection algorithm [14] and then get the horizontal (or vertical) projection profile according to the page skew angle ($\theta$), as shown in (2). In this equation, $I(x, y)$ is the intensity value in a $width \times height$ image, and $P_H^{(n)}$ is the horizontal projection profile at the $n$th level. An example is given in Fig. 4.

$$P_H^{(n)} = \left\{ p_y \mid p_y = \sum_{x=0}^{width-1} I(x, y + xtan\theta),\ 0 \le y < height \right\}. \tag{2}$$

At the second step, smoothing is applied to control the amount of details in the projection profile, as shown in (3), where $s$ is the kernel size and $m_y$ is the integer value. Fig. 5 describes this example, which is the result of smoothing for the projection profile in Fig. 4.

$$M_H^{(n)} = \left\{ m_y \mid m_y = \lfloor \frac{1}{s} \sum_{i=y-s/2}^{y+s/2} p_i \rfloor,\ 0 \le y < height,\ p_i \in P_H^{(n)} \right\}. \tag{3}$$

At the third step, we find the first derivative of the smoothed projection profile or get the gradients of neighbored elements in the smoothed projection profile, as described in (4). This example is shown in Fig. 6 which is the result of the first derivative of the smoothed projection profile shown in Fig. 5.

$$F_H^{(n)} = \left\{ f_y | f_y = \frac{dm_y}{dx}, 0 \le y < height, m_y \in M_H^{(n)} \right\}. \tag{4}$$

At the fourth step, as shown in Fig. 7, we find the zero crossing points in the first derivative of the smoothed projection profile, as described in (5). The zero crossing points, found from the equation, correspond to local maxima or minima points of the smoothed projection profile.

$$\begin{aligned} Z^{(n)} = \Big\{ z \mid & (f_z < 0 \ and \ f_{z+1} \ge 0) \\ & or\ (f_z > 0 \ and \ f_{z+1} \le 0), f_z \in F_H^{(n)} \Big\}. \end{aligned} \tag{5}$$

Authorized licensed use limited to: Korea University. Downloaded on October 25, 2008 at 01:45 from IEEE Xplore. Restrictions apply.

Fig. 3. Multiscale images: (a) top-level = level 4 (50x66), (b) level 3 (100x131), (c) level 2 (199x261), (d) level 1 (398x522), and (e) bottom-level = level 0 (796x1043).

Therefore, the variation in the distances between neighboring zero crossing points, calculated by (6), indicates the frequency distribution of its region. That the variation of a region is low means that the frequency distribution of text lines of the region is almost the same, that is, the region is composed of a homogeneous region. On the other hand, that the variation of a region is high means that the frequency distribution of text lines of the region is jagged, so the region needs to be segmented further.

$$V = \frac{\sum_{i=1}^{n}(d_i - m)^2}{n}, \qquad where \begin{cases} d_i = z_{i+1} - z_i,\ z_i \in Z^{(n)}, \\ m = \dfrac{\sum_{i=1}^{n} d_i}{n}. \end{cases} \tag{6}$$

For normalizing the range of variance $V$, we form a normalization function, as shown in (7). By this equation, if the variance $V$ of a region comes to 0, $P$ becomes 1, and if the variance $V$ of a region comes to infinity, $P$ becomes 0. Therefore, the periodicity of a region is expressed by the value $P$ which is in the range from 0 to 1.

Authorized licensed use limited to: Korea University. Downloaded on October 25, 2008 at 01:45 from IEEE Xplore. Restrictions apply.

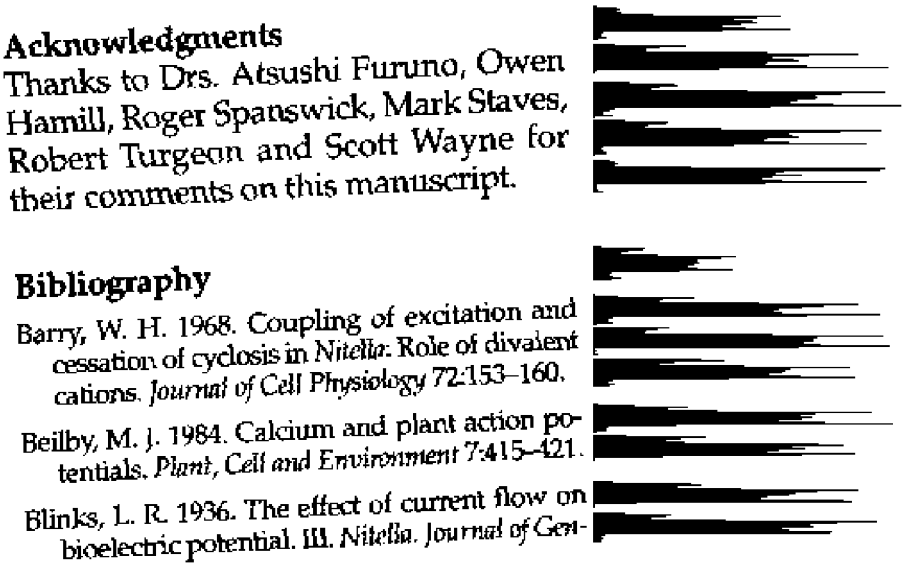



Fig. 4. The result of the projection profile according to the page skew angle ($\theta$).

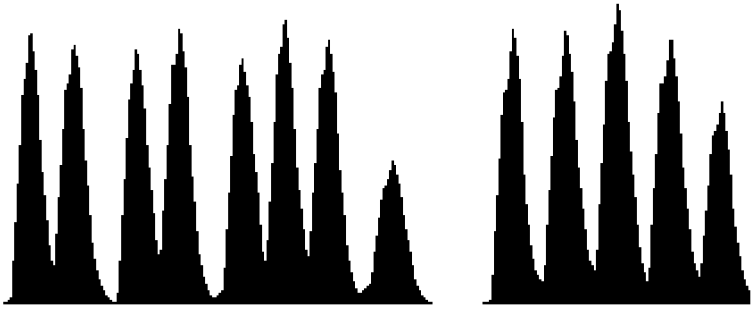


Fig. 5. The result of smoothing for the projection profile.

Fig. 6. The result of the first derivative of the smoothed projection profile.

$$P = 2 - \frac{2}{1 + e^{-V}}. \tag{7}$$

Therefore, we can estimate whether a region is composed of a single period and frequency or not by the decision value $D(P)$ of (8).

$$D(P) = \begin{cases} 1, & if\ P > TH \\ 0, & otherwise. \end{cases} \tag{8}$$

In this equation, we set the threshold value: $TH = 0.5$. It is approximately equal to 1.099 of the variance $V$ in (6). The value, which is not dependent on the changes in character font size, text line spacing, and document layout structure, is applied equally to each region to decide whether it is a homogeneous region or not. If the decision value $D(P)$ of (8) is 0, the variance between the differences of zero crossing points is high and the region is not a homogeneous region; therefore, it requires further splitting. On the other hand, if the decision value $D(P)$ of (8) is 1, the variance between the differences of zero crossing points is low and the region is a homogeneous one; therefore, it requires no further splitting.

In the same way, we can compute the vertical projection profile, its variance, and the periodicity of a region in the vertical direction.

### 3.3 Determination of Splitting Position

When a region is determined to require further splitting because it is not a homogeneous region, there are two cases in the horizontal direction, as shown in Fig. 8.

The case in Fig. 8a needs to be split more because one white area is larger than the other white areas and the case in Fig. 8b needs to be split more because one black area is larger than the other black areas. In these two cases, we find

Authorized licensed use limited to: Korea University. Downloaded on October 25, 2008 at 01:45 from IEEE Xplore. Restrictions apply.

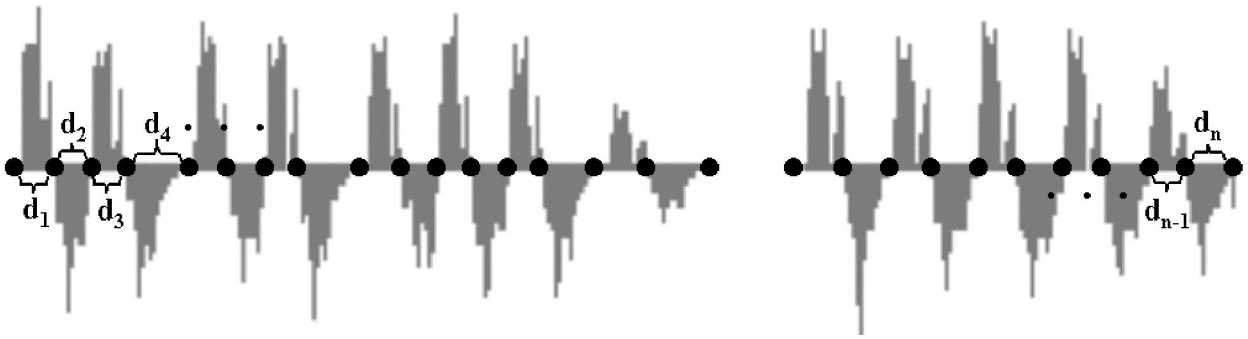



Fig. 7. The zero crossing points in the first derivative of the smoothed projection profile.

a suitable position for splitting, using the method given below, and split it into two regions.

The processes described in Sections 3.2 and 3.3 are repeated for each region until no further splitting is required.

Let $W$ denote the set of the white areas of a region, $w_i$
Let $B$ denote the set of the black areas of a region, $b_i$
Sort the set $W$(or $B$) in the increasing order of the magnitude of $w_i$(or $b_i$)
$w_{med} \leftarrow$ the median element of $W$
$b_{med} \leftarrow$ the median element of $B$
$w_{max} \leftarrow$ the last element of $W$
$b_{max} \leftarrow$ the last element of $B$
if $w_i > w_{med}$ and $w_i = w_{max}$, split $w_i$
if $b_i > b_{med}$ and $b_i = b_{max}$, split $w_{i-1}$

### 3.4 Confirmation of Segmentation Results

By repeating the processes described in Sections 3.2 and 3.3, we segmented a document image into maximal homogeneous regions. However, there are three types of segmentation or identification errors if only the periodicity estimation is used, as shown in Fig. 9. In this figure, the samples in Fig. 9a cause a segmentation error because some regions are touched or overlapped, the samples in Fig. 9b cause a

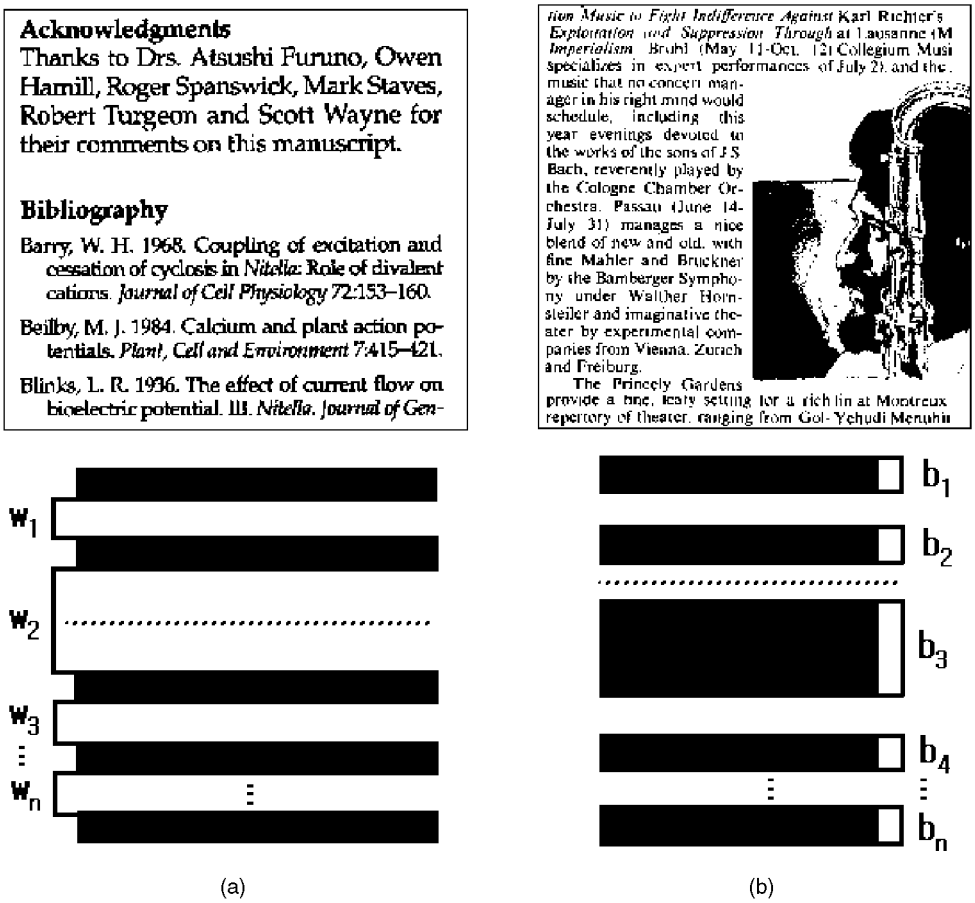
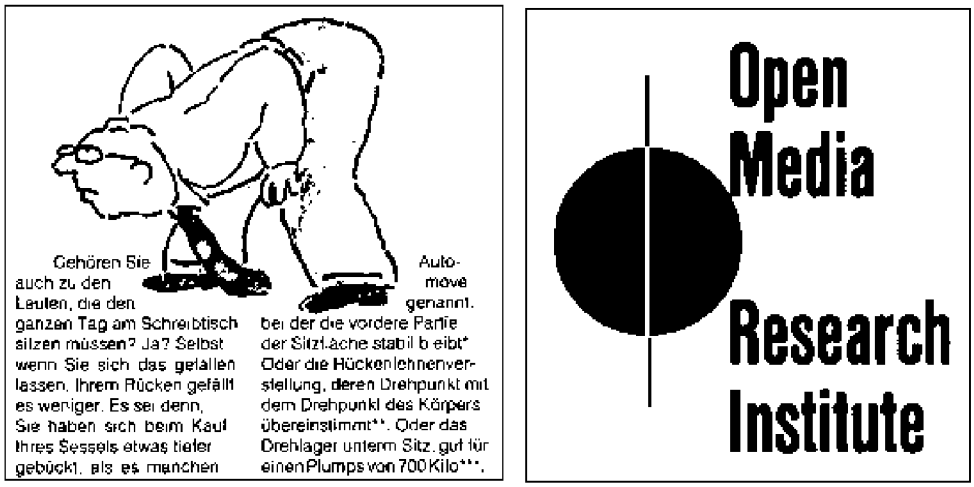
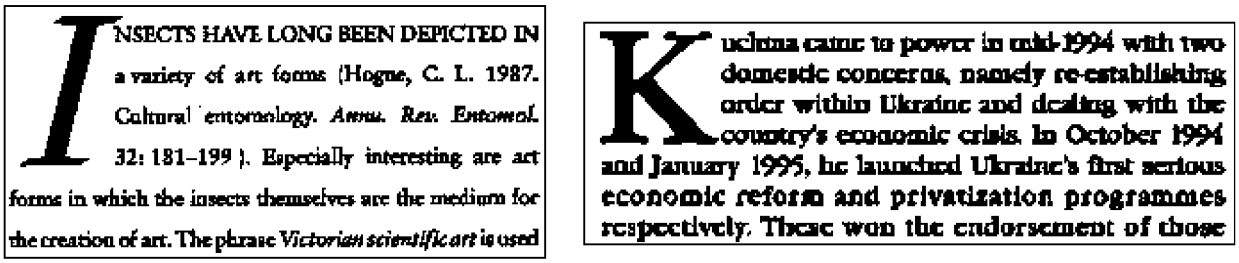



Fig. 8. Two cases requiring further horizontal splitting: (a) A distinct white area exists. (b) A distinct black area exists.

Authorized licensed use limited to: Korea University. Downloaded on October 25, 2008 at 01:45 from IEEE Xplore. Restrictions apply.




(a)




(b)

(c)

Fig. 9. Three cases that cause segmentation or identification errors with the periodicity estimation.

segmentation error because of a region beginning with a very large character, and the samples in Fig. 9c cause an identification error since a region which is composed of a few words and a single text line gives insufficient information for estimating its periodicity.

If a region is composed of a single period and has just one peak in the smoothed horizontal projection profile, that is, the decision value $D(P)$ of (8) in the horizontal direction is equal to 1 and the number of zero crossing points of the smoothed horizontal projection profile is smaller than or equal to 3, we cannot be confident with only the result of the periodicity estimation. Therefore, in this case, we confirm whether it is a homogeneous region or not by its texture analysis.

In this paper, the wavelet multiresolution (MRA) is adopted for texture analysis. In the next part, the basic notation of MRA will be given.

Let $L^2(R)$ be the Hilbert space of all the square integrable functions on $R$. A sequence $\{V_j\}$ of closed subsequence of $L^2(R)$ is called a multiresolution approximation if the following properties are satisfied:

1. $\forall\, j \in Z$, $V_{j-1} \supset V_j$;
2. $\forall\, (j,k) \in Z^2$, $f(x) \in V_j \Longleftrightarrow f(x - 2^j k) \in V_j$;
3. $\forall\, j \in Z$, $f(x) \in V_j \Longleftrightarrow f(\frac{x}{2}) \in V_{j+1}$;
4. $\lim_{j \to -\infty} V_j = \overline{(\bigcup_{j \in Z} V_j)} = L^2(R)$;
5. $\lim_{j \to +\infty} V_j = \bigcap_{j \in Z} V_j = \{0\}$;
6. There exists $\varphi_0(x) \in L^2(R)$ such that $\{\varphi_0(x-n)\}_{n \in Z}$ is a Riesz basis of $V_0$.

Such a function $\varphi(x)$ is called a scaling function and is said to generate the one-dimensional $MRA(V_j)$. Furthermore, if the compactly supported scaling function exists, whose integer translation form an orthonormal basis of $V_0$, then the $MRA(V_j)$ is called an orthogonal MRA.

For an orthogonal $MRA(V_j)$, let $W_j = V_{j-1} \ominus V_j$. Especially, $W_0 = V_{-1} - V_0$. We can define a function $\psi(x)$ whose dilation $\{\psi(x-n)\}_{n \in Z}$ being the orthonormal basis of $V_0$, such a function $\psi(x)$ is called an orthonormal wavelet function. In general, for any $j \in Z$, the dilation and dyadic translations

$$\left\{2^{-\frac{j}{2}}\varphi(2^{-j}x - n)\right\}_{n \in Z} \text{ and } \left\{2^{-\frac{j}{2}}\psi(2^{-j}x - n)\right\}_{n \in Z}$$

are the orthonormal bases of $V_j$ and $W_j$, respectively.

Authorized licensed use limited to: Korea University. Downloaded on October 25, 2008 at 01:45 from IEEE Xplore. Restrictions apply.

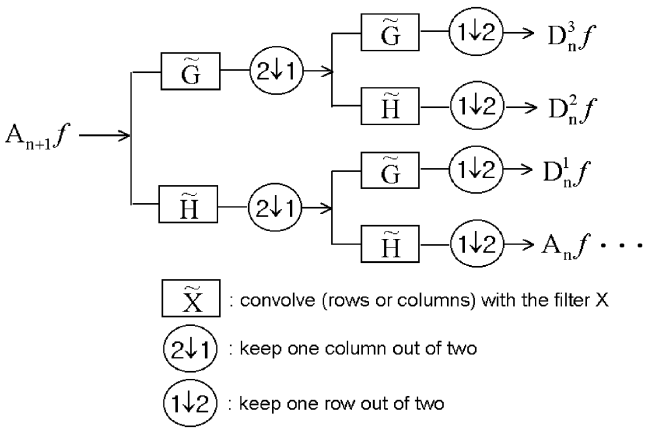



Fig. 10. Decomposition of an image $A_{n+1}f$ into $A_nf$, $D_n^1f$, $D_n^2f$, and $D_n^3f$ [12].

Since $V_0 = V_1 \oplus W_1$, $\forall f(x) \in L^2(R)$ can be projected onto $V_1$ and $W_1$, associated with the projection, $f(x)$ is decomposed as

$$f(x) = P_{V_1} f(x) \oplus P_{W_1} f(x), \tag{9}$$

we called the projection functions, which indicate the low and high frequency components of the original signal, L and H subsignals. The decomposition can be continued due to having $V_1 = V_2 \oplus W_2 = V_3 \oplus W_3 \oplus W_2 = \cdots$.

Similar to 1D MRA, we can produce a very significant formula,

$$\begin{aligned} V_{j-1} \otimes V_{j-1} &= (V_j \oplus W_j) \otimes (V_j \oplus W_j) \\ &= (V_j \otimes V_j) \oplus (V_j \otimes W_j) \oplus (W_j \otimes V_j) \oplus (W_j \otimes W_j). \end{aligned} \tag{10}$$

For $\forall f(x,y) \in V_{j-1} \otimes V_{j-1}$, it can be decomposed as

$$\begin{aligned} f(x,y) = {} & f_{V_j \otimes V_j}(x,y) \oplus f_{V_j \otimes W_j}(x,y) \\ & \oplus f_{W_j \otimes V_j}(x,y) \oplus f_{W_j \otimes W_j}(x,y). \end{aligned} \tag{11}$$

These four functions are the projections of f(x,y) onto the four spaces on the right side of (10).

The dyadic wavelet is chosen in the decomposition because the width and height of the subimage are just half as that of the original image. Thus, the computing speed should be much faster without affecting the result.

The 2D Harr wavelet transform [12] has been adopted for texture analysis since it is the simplest and fastest transform among the various MRA transforms and can be applied to binary images.

In two dimensions, the wavelet representation can be computed with a pyramidal algorithm similar to the one-dimensional algorithm. The two-dimensional wavelet transform can be seen as a one-dimensional wavelet transform along the $x$ and $y$ axes. By repeating a one-dimensional algorithm, we can show that a two-dimensional wavelet transform can be computed with a separable extension of the one-dimensional decomposition algorithm. At each step, we decompose $A_nf$, $D_n^1f$, $D_n^2f$, and $D_n^3f$, as illustrated in Fig. 10. We first convolve the rows of $A_{n+1}f$ with a one-dimensional filter, retain every other row, convolve the columns of the resulting signals with another one-dimensional filter, and retain every other column. By wavelet transform shown in Fig. 10, a document image is transformed into four subimages: $LL$, $LH$, $HL$, and $HH$. The $LL$ subimage has low frequencies in both horizontal and vertical directions and it corresponds to $A_nf$. The $LH$ subimage has low frequencies in the horizontal direction and high frequencies in the vertical direction and it corresponds to $D_n^1f$. The $HL$ subimage has high frequencies in the horizontal direction and low frequencies in vertical direction and it corresponds to $D_n^2f$. Finally, the $HH$ subimage has high frequencies in both horizontal and vertical directions and it corresponds to $D_n^3f$ [13].

In the case of 2D Harr wavelet for binary images, the low pass filter $\tilde{H}$ can be considered as the logical OR and the high pass filter $\tilde{G}$ can be considered as the exclusive OR. Therefore, $LL^{(n)}$, $LH^{(n)}$, $HL^{(n)}$, and $HH^{(n)}$ subimages corresponding to $A_nf$, $D_n^1f$, $D_n^2f$, and $D_n^3f$, respectively, can be computed as (12). By applying (12) repeatedly, the multiscale $LL$, $LH$, $HL$, and $HH$ subimages is obtained as shown in Fig. 11.

$$\begin{aligned} LL_{(i,j)}^{(n+1)} &= \left(LL_{(2i,2j)}^{(n)} \vee LL_{(2i+2,2j)}^{(n)}\right) \wedge \left(LL_{(2i,2j)}^{(n)} \vee LL_{(2i,2j+2)}^{(n)}\right) \\ LH_{(i,j)}^{(n+1)} &= \left(LL_{(2i,2j)}^{(n)} \vee LL_{(2i+2,2j)}^{(n)}\right) \wedge \left(LL_{(2i,2j+1)}^{(n)} \otimes LL_{(2i,2j+3)}^{(n)}\right) \\ HL_{(i,j)}^{(n+1)} &= \left(LL_{(2i+1,2j)}^{(n)} \otimes LL_{(2i+3,2j)}^{(n)}\right) \wedge \left(LL_{(2i,2j)}^{(n)} \vee LL_{(2i,2j+2)}^{(n)}\right) \\ HH_{(i,j)}^{(n+1)} &= \left(LL_{(2i+1,2j)}^{(n)} \otimes LL_{(2i+3,2j)}^{(n)}\right) \\ &\quad \wedge \left(LL_{(2i,2j+1)}^{(n)} \otimes LL_{(2i,2j+3)}^{(n)}\right). \end{aligned} \tag{12}$$

In order to classify the transform coefficients into images, texts, horizontal ruling lines, and vertical ruling lines, we first assign the coefficients $c, c \in \{I, T, H, V\}$, to the corresponding classes (image, text, horizontal ruling line, or vertical ruling line) according to Table 1, at each level. If the coefficient is high in the $LH$ subimage and low in the $HL$ subimage, it means that there are only horizontal components, therefore, we assign it to horizontal ruling line(H). On the other hand, if the coefficient is low in the $LH$ subimage and high in the $HL$ subimage, it means that there are only vertical components and we therefore assign it to vertical line(V). If the coefficient is high in both $LH$ and $HL$ subimages, it means that there are both horizontal and vertical components, and it is therefore assigned to text(T). If the coefficient is low in both $LH$ and $HL$ subimages, it would be an image or background which has neither horizontal nor vertical components. In that case, we reference the corresponding coefficients of the $LL$ subimage. Thus, if its corresponding coefficient is low in the $LL$ subimage, it is assigned to image(I) and, if its corresponding coefficient is high in the $LL$ subimage, it is assigned to background(X).

At the second step, the assigned coefficients are weighted by their neighbors, which are also assigned by Table 1. Weight matrices corresponding to each class are given in Fig. 12. The assigned coefficients are weighted with neighboring coefficients of equal class type, as shown in (13). In this equation, $m_i^c$ is the $i$th element of the weight

Authorized licensed use limited to: Korea University. Downloaded on October 25, 2008 at 01:45 from IEEE Xplore. Restrictions apply.

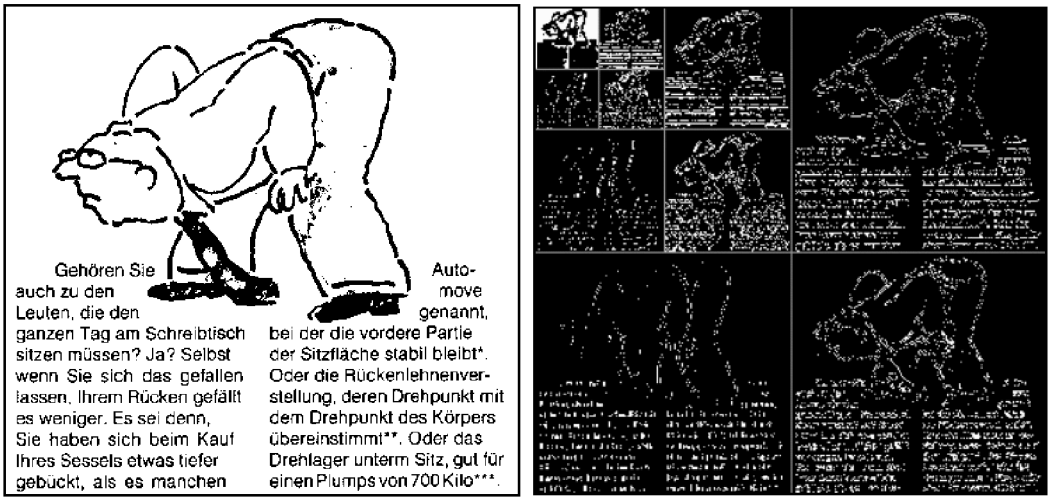



Fig. 11. Original image (left) and multiscale subimages applied 2D Harr wavelet transform (right).

matrix of the corresponding class and $b_i$ is 1 if the assigned class of $b_i$ is equal to class $c, c \in \{I, T, H, V\}$, otherwise, $b_i$ is 0. $V^n_{x,y}(c)$ is the weighted value of the corresponding coefficient at point $(x, y)$ of the $n$th level.

$$V^n_{x,y}(c) = \sum_{i=1}^{9} m^c_i \cdot b_i. \tag{13}$$

It the following step, we accumulate the weighted value $V^n_{x,y}(c)$ of each level, as shown in (14). By applying this equation from the top level (level N) to the bottom level (level 1), we can obtain the final accumulated value of corresponding coefficient $c$, which is equal to $A^1_{x,y}(c)$.

$$A^n_{x,y}(c) = V^n_{x,y}(c) + A^{n+1}_{\lfloor x/2 \rfloor, \lfloor y/2 \rfloor}(c), \quad n = N-1, \cdots, 1. \tag{14}$$

In the final step, we decide the classes of the corresponding coefficients with images, texts, horizontal ruling lines, or vertical ruling lines at level 1. As shown in (15), we decide the class of the corresponding coefficient $c$ by its maximal value among the accumulated values of I, T, H, and V.

$$D(x, y) = \arg \max_c A^1_{x,y}(c), \quad c \in \{I, T, H, V\}. \tag{15}$$

As a result of this procedure, we can get the classification results shown in Fig. 13. After removing the nontext regions (image, horizontal ruling line, and vertical line drawing) and small connected components of the text region, we apply the processes described in Sections 3.2 and 3.3 once again. Through this procedure, better segmentation results are obtained.

TABLE 1
Assigning the Coefficients to the Corresponding Classes at Each Level

| | | LH | |
|---|---|---|---|
| | | low (0) | high (1) |
| HL | low (0) | I / X | H |
| | high (1) | V | T |

### 3.5 Homogeneous Region Identification

By repeating the processes described in Sections 3.2 and 3.3 and applying Section 3.4 for ambiguous regions, given a document image as input, we get the maximal homogeneous regions of that image clearly. In this section, we present the method for identifying the segmented maximal homogeneous regions as texts, images, tables, and ruling lines at the base image, level 0. The heuristics for region identification are as follows:

- Text identification. If a region has a horizontal periodicity and is composed of one or more peaks in the smoothed horizontal projection profile, that is, the decision value $D(P)$ of (8) in the horizontal direction is 1 and the number of zero crossing points of the smoothed horizontal projection profile is larger than 3, it is classified as a text block. If a region has a vertical periodicity and is composed of just one peak in the smoothed horizontal projection profile, that is, the decision value $D(P)$ of (8) in the vertical direction is 1 and the number of zero crossing points of the smoothed horizontal projection profile is smaller than or equal to 3, it is classified as a single text line, such as title, subtitle, legend, and caption.
- Ruling line identification. After applying the Run Length Smoothing Algorithm (RLSA) [1] for linking disconnected ruling lines because of the binarization, if a region is composed of one connected component and its ratio of width to height or height to width is larger than 5, it is classified as a ruling line.
- Table identification. For the remaining regions, we extract any existing horizontal lines. The top and

$$M^I, M^T = \begin{bmatrix} 0 & \frac{1}{2} & 0 \\ \frac{1}{2} & 1 & \frac{1}{2} \\ 0 & \frac{1}{2} & 0 \end{bmatrix} \qquad M^H = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 1 & 1 \\ 0 & 0 & 0 \end{bmatrix} \qquad M^V = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

Fig. 12. Weight matrices of each class showing the relationship between neighbors.

Authorized licensed use limited to: Korea University. Downloaded on October 25, 2008 at 01:45 from IEEE Xplore. Restrictions apply.

Fig. 13. Original images (left) and their classification results (right) (blue: image, green: text, red: horizontal ruling line, yellow: vertical ruling line, and black: background).

bottom ruling lines, which should be similar in length with the width of the region, are detected. Furthermore, there should be more than one horizontal ruling line in the region other than the top and bottom ruling lines.

- Image identification. The rest of the regions are regarded as image regions.

## 4 Experimental Results and Analysis

Our proposed method has been tested on 150 images from the document database volume II available from the University of Washington (UWDB) [17] and all 233 images in the article category out of 512 images from the MediaTeam Document Database [18].

The example of the processes using our method is shown in Fig. 15. The performance result with UWDB [17] and MediaTeam DB [18] is quantitatively evaluated, as shown in Table 2. In addition, the corresponding graph is illustrated in Fig. 14 to show the performance comparison of the proposed method with recursive X-Y cut method [4] and Block Adjacency Graph (BAG) method [11]. Experimental results indicate that our method provides more accurate results than previous ones.

We experimented on a Pentium 200 MHz PC with 64 MB memory. The average computation time was 0.22 seconds for constructing a pyramidal quadtree structure, 1.46 seconds for segmenting a document image into homogeneous regions, and 0.58 seconds for conforming segmentation results by using texture information. So, the total average computation time to analyze a document image was 2.26 seconds.

Authorized licensed use limited to: Korea University. Downloaded on October 25, 2008 at 01:45 from IEEE Xplore. Restrictions apply.

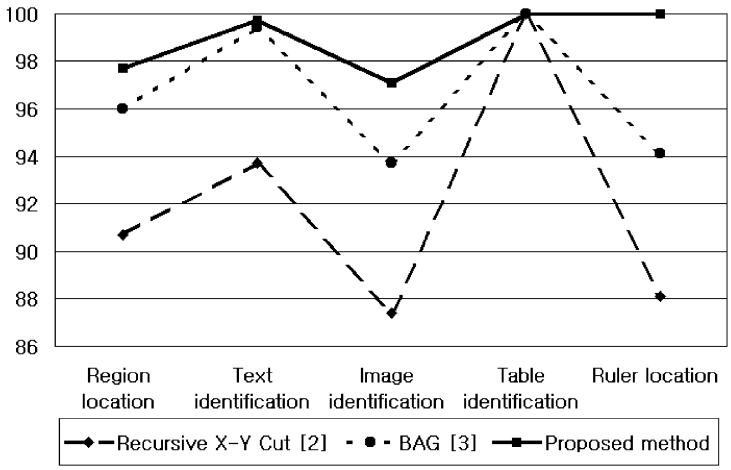



Fig. 14. The comparison of previous methods and the proposed method experimented on images from the UWDB.

A comparison of the results between our method and the BAG method [11] is shown in Fig. 16. We marked wrong segmentation or identification results with thick circles in following figures. In the upper images of Fig. 16, when text line spaces are large in a text block, the BAG method could not segment it as one text block and has split it into several text lines. However, with the proposed method, large text line spaces did not have a bad influence if text lines are regularly aligned. Another problem with the BAG method in those images is that it could not split touched or overlapped regions. However, this problem was solved by the confirmation procedure (Section 3.4) which used the texture information on limited cases. In the lower images of Fig. 16, large font size of titles made it difficult to segment with the BAG method, while it did not matter with the proposed method.

Other experimental results comparing our method with the commercial software Armi 5.0 [15] and OmniPage Pro 9.0 [16] are given in Figs. 17, 18, and 19. In Fig. 17, Armi 5.0 gave incorrect segmentation results because of large font sizes in the title and OmniPage Pro 9.0 could not split a large character at the beginning of the first paragraph.

Fig. 18 shows the result of experimentation on the images from the UWDB. In this figure, Armi 5.0 could not make one text block in the title and subtitle. It seems that it is not important to understand the structure of a document image, but if we consider logical layout analysis [19] such as assigning the labels of the text regions, it could become vague what the title is and what the subtitle is and we cannot get accurate semantic information for them. In the case of OmniPage Pro 9.0 in Fig. 18, several paragraphs are merged and an image region is split into several regions incorrectly.

Fig. 19 gives the experimental results for the images from the MediaTeam DB. In this figure, Armi 5.0 yields an incorrect identification result or a single text line at the top of the image is identified as an image region, while OmniPage Pro 9.0 merged several neighbored text regions into one.

## 5 Conclusions and Further Research

We have presented a parameter-free geometric document layout analysis method that segments the document images into maximal homogeneous regions and identifies them as texts, images, tables, and ruling lines, with a pyramidal quadtree structure, the proposed periodicity measure, and some texture analysis in limited cases. The proposed method was experimented with the images from the University of Washington Database and images from MediaTeam Document Database. Experimental results have shown that the proposed method provides more accurate results than the previous ones.

Compared with other methods relevant to geometric document layout analysis and page segmentation, the major contributions of the proposed method are summarized as

TABLE 2
Performance Evaluation for the Segmentation and Identification

| | UWDB [17] | | | MediaTeam DB [18] | | |
|---|---|---|---|---|---|---|
| | $N_{total}$ | $N_{correct}$ | Correct rate(%) | $N_{total}$ | $N_{correct}$ | Correct rate(%) |
| Region location | 1559 | 1523 | 97.7 | 3018 | 2906 | 96.3 |
| Text identification | 1473 | 1469 | 99.7 | 2693 | 2668 | 99.1 |
| Image identification | 139 | 135 | 97.1 | 365 | 344 | 94.2 |
| Table identification | 4 | 4 | 100 | 20 | 19 | 95.0 |
| Ruler location | 101 | 101 | 100 | 76 | 74 | 97.4 |

Authorized licensed use limited to: Korea University. Downloaded on October 25, 2008 at 01:45 from IEEE Xplore. Restrictions apply.

Fig. 15. An example to show the processes of the proposed method: (a) level 4 (50x66), (b) level 3 (100x131), (c) level 2 (199x261), (d) level 1 (398x522), and (e) level 0 (796x1,043).

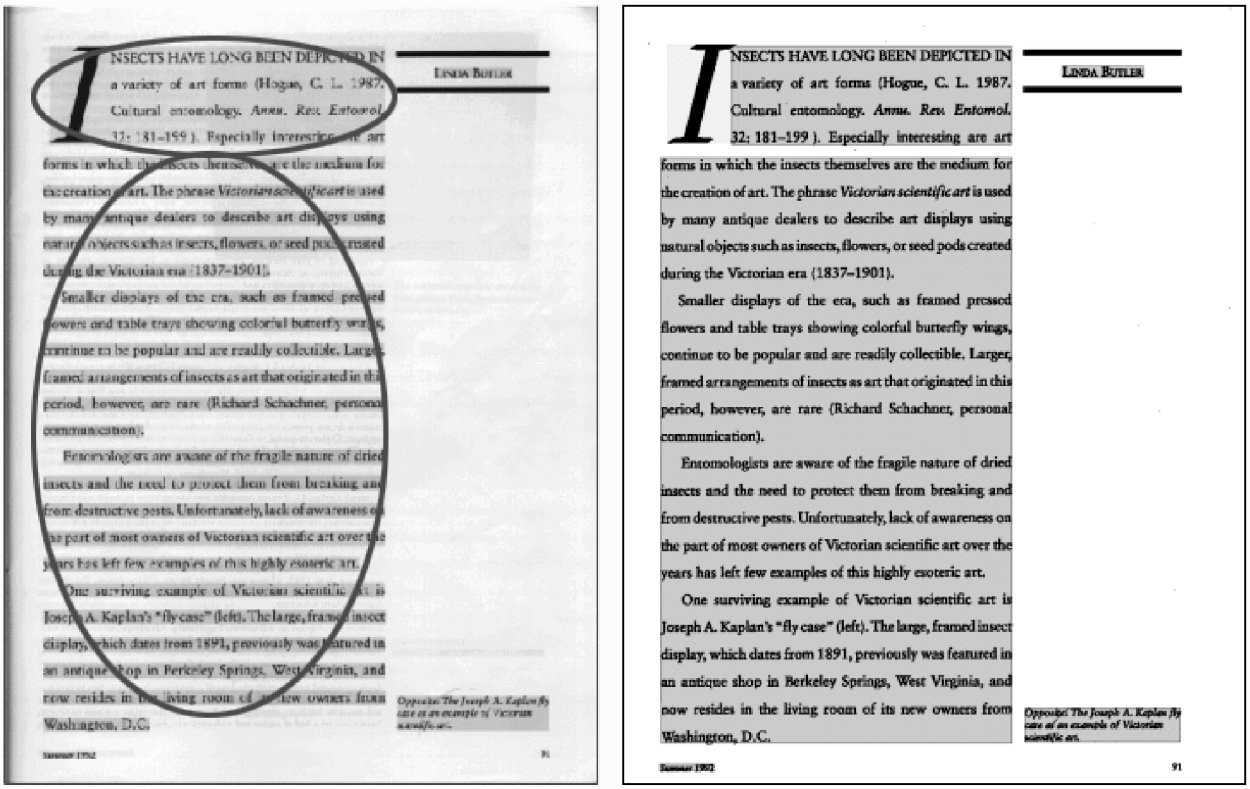
follows: 1) independence on any parameters to segment the document images having various character font sizes, text line spacing, and document layout structures, 2) effective extraction of periodical features for page segmentation by using the proposed periodicity measure, and 3) robust page segmentation obtained with the confirmation procedure using some texture analysis on ambiguous regions.

This study was confined to geometric layout analysis. However, for transforming the scanned document images into semantic electronic version, a logical layout analysis [19] is also required for determining the type of page, assigning labels to texts (title, subtitle, caption, footnote, etc.), determining the relationships among each region, and ordering the text regions according to their reading sequences.

## ACKNOWLEDGMENTS

This research was supported by National Creative Research Initiatives Program of the Korean Ministry of Science and Technology.

Authorized licensed use limited to: Korea University. Downloaded on October 25, 2008 at 01:45 from IEEE Xplore. Restrictions apply.



INSECTS HAVE LONG BEEN DEPICTED IN a variety of art forms (Hogue, C. L. 1987. Cultural entomology. *Annu. Rev. Entomol.* 32: 181–199). Especially interesting are art forms in which the insects themselves are the medium for the creation of art. The phrase *Victorian scientific art* is used by many antique dealers to describe art displays using natural objects such as insects, flowers, or seed pods created during the Victorian era (1837–1901).

Smaller displays of the era, such as framed pressed flowers and table trays showing colorful butterfly wings, continue to be popular and are readily collectible. Larger, framed arrangements of insects as art that originated in this period, however, are rare (Richard Schachner, personal communication).

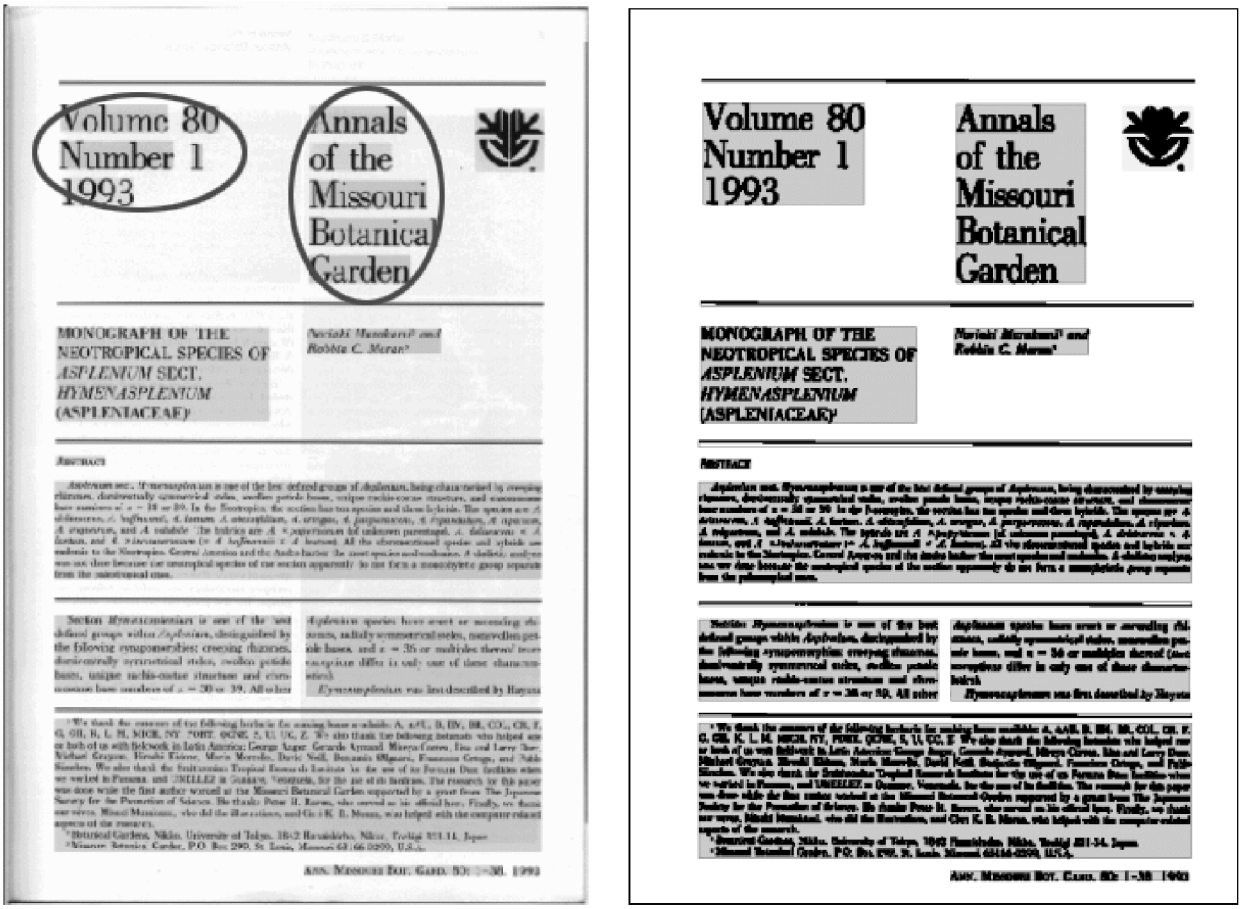
Entomologists are aware of the fragile nature of dried insects and the need to protect them from breaking and from destructive pests. Unfortunately, lack of awareness on the part of most owners of Victorian scientific art over the years has left few examples of this highly esoteric art.

One surviving example of Victorian scientific art is Joseph A. Kaplan's "fly case" (left). The large, framed insect display, which dates from 1891, previously was featured in an antique shop in Berkeley Springs, West Virginia, and now resides in the living room of its new owners from Washington, D.C.

LINDA BUTLER

*Opposite: The Joseph A. Kaplan fly case as an example of Victorian scientific art.*

Summer 1992 91



Volume 80
Number 1
1993

Annals of the Missouri Botanical Garden

MONOGRAPH OF THE NEOTROPICAL SPECIES OF *ASPLENIUM* SECT. *HYMENASPLENIUM* (ASPLENIACEAE)

*Noriaki Murakami[1] and Robbin C. Moran[2]*

ABSTRACT

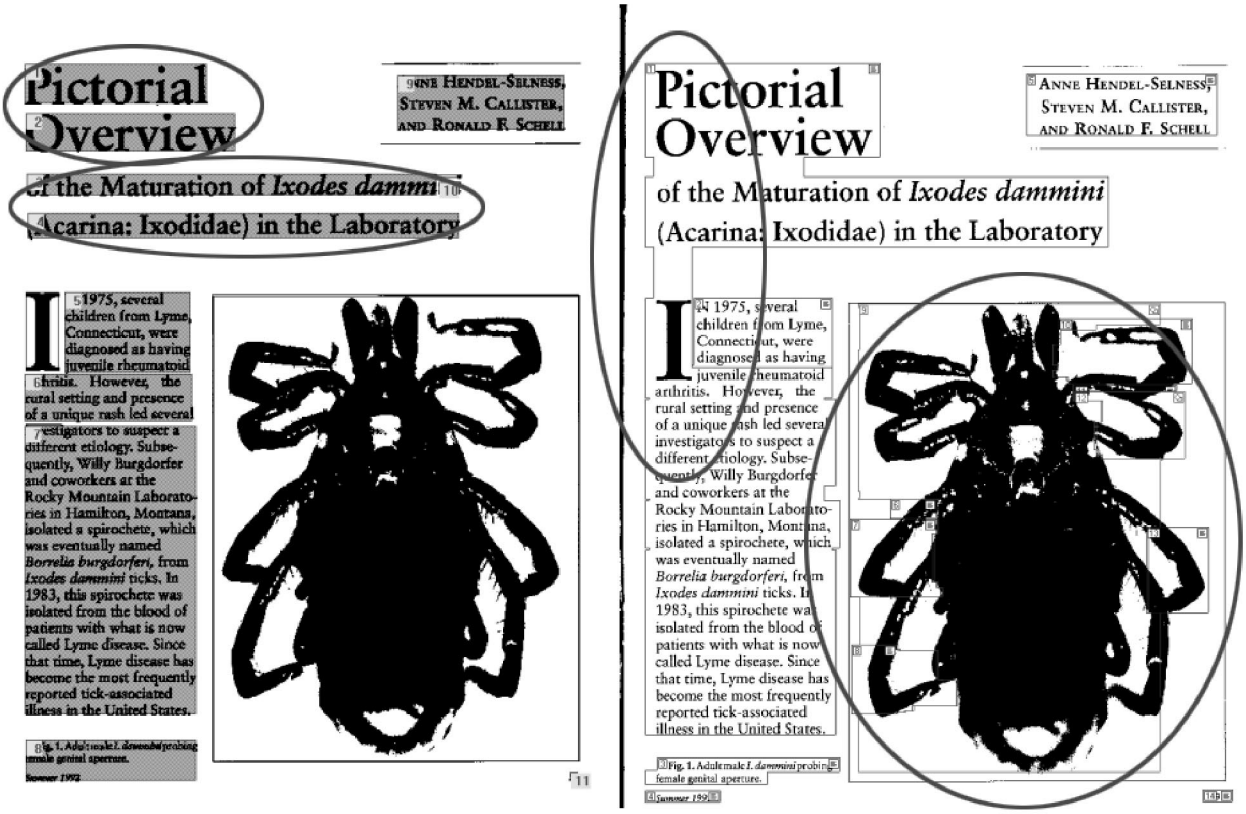
ANN. MISSOURI BOT. GARD. 80: 1–38. 1993

(a) (b)

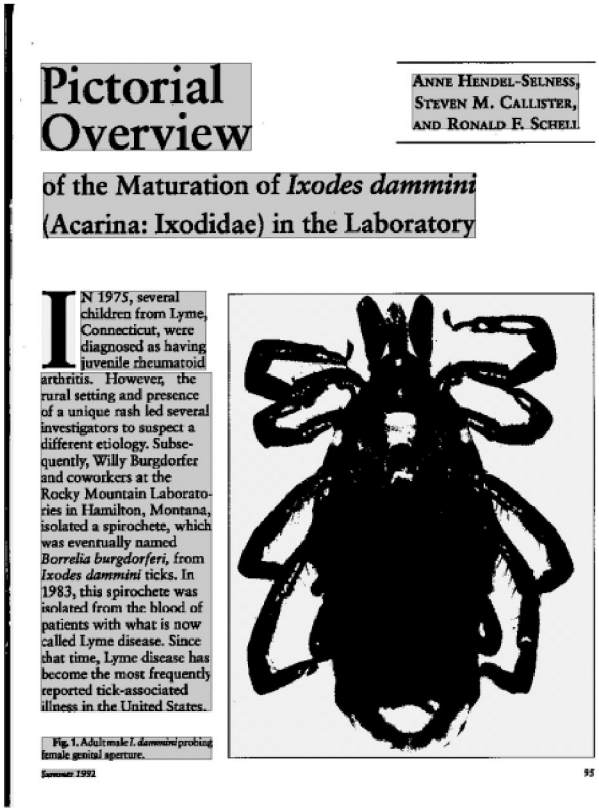
(c)

Fig. 17. Experimental results: (a) Armi 5.0, (b) OmniPage Pro 9.0, and (c) the proposed method.

Authorized licensed use limited to: Korea University. Downloaded on October 25, 2008 at 01:45 from IEEE Xplore. Restrictions apply.



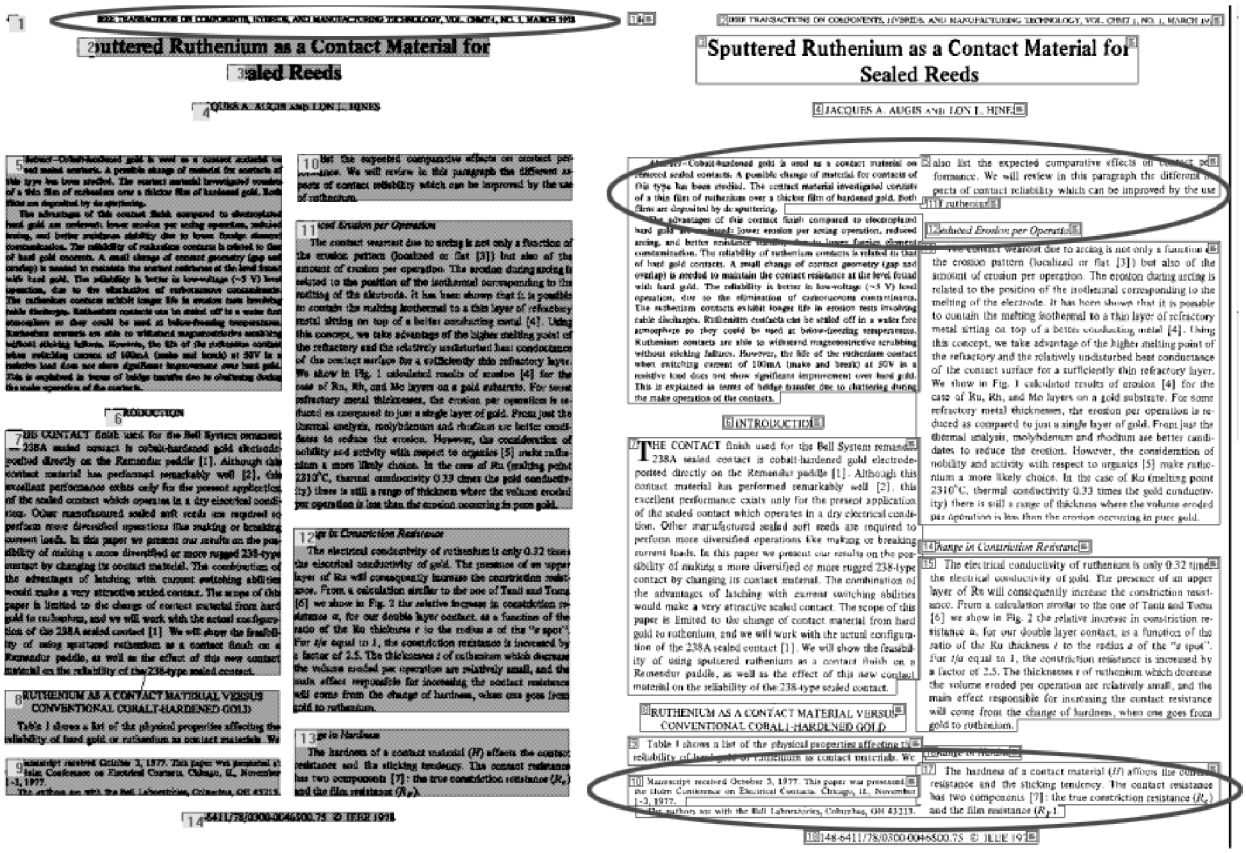

(a) (b)



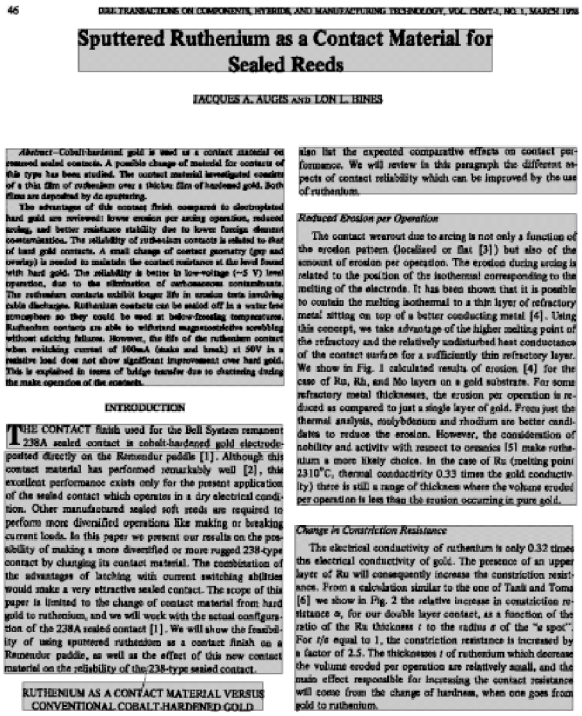

(c)

Fig. 18. Experimental results for the UWDB: (a) Armi 5.0, (b) OmniPage Pro 9.0, and (c) the proposed method.

Authorized licensed use limited to: Korea University. Downloaded on October 25, 2008 at 01:45 from IEEE Xplore. Restrictions apply.



(a) (b)



(c)

Fig. 19. Experimental results for the MediaTeam DB: (a) Armi 5.0, (b) OmniPage Pro 9.0, and (c) the proposed method.

Authorized licensed use limited to: Korea University. Downloaded on October 25, 2008 at 01:45 from IEEE Xplore. Restrictions apply.

## REFERENCES


[1] F. Legourgiois, Z. Bublinski, and H. Emptoz, "A Fast and Efficient Method for Extracting Text Paragraphs and Graphics from Unconstrained Documents," *Proc. 11th Int'l Conf. Pattern Recognition,* pp. 272-276, 1992.

[2] D. Drivas and A. Amin, "Page Segmentation and Classification Utilizing Bottom-Up Approach," *Proc. Third Int'l Conf. Document Analysis and Recognition,* pp. 610-614, 1995.

[3] A. Simon, J. Pret, and A. Johnson, "A Fast Algorithm for Bottom-Up Document Layout Analysis," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 19, pp. 273-276, 1997.

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

[5] J. Ha, R. Haralick, and I. Phillips, "Document Page Decomposition by the Bounding-Box Projection Technique," *Proc. Third Int'l Conf. Document Analysis and Recognition,* pp. 1119-1122, 1995.

[6] A. Jain and Y. Zhong, "Page Segmentation Using Texture Analysis," *Pattern Recognition,* vol. 29, pp. 743-770, 1996.

[7] A. Jain and S. Bhattacharjee, "Text Segmentation Using Gabor Filters for Automatic Document Processing," *Machine Vision and Applications,* vol. 5, pp. 169-184, 1992.

[8] H. Cheng and C. Bouman, "Trainable Context Model for Multiscale Segmentation," *Proc. IEEE Int'l Conf. Image Processing,* vol. 1, pp. 610-614, 1998.

[9] K. Etemad, D. Doermann, and R. Chellappa, "Multiscale Document Page Segmentation Using Soft Decision Integration," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 19, pp. 92-96, 1997.

[10] A. Antonacopoulos, "Page Segmentation Using the Description Background," *Computer Vision and Image Understanding,* vol. 70, pp. 350-369, 1998.

[11] A. Jain and B. Yu, "Document Representation and Its Application to Page Decomposition," *IEEE Trans. Pattern Analysis and Machine Intelligence* vol. 20, pp. 294-308, 1998.

[12] S. Mallat, "A Theory for Multiresolution Signal Decomposition: The Wavelet Representation," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 11, pp. 674-692, 1989.

[13] Y.Y. Tang, H. Ma, J. Liu, B.F. Li, and D. Xi, "Multiresolution Analysis in Extraction of Reference Lines from Documents with Gray Level Background," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 19, pp. 921-926, 1997.

[14] B. Gatos, N. Papamarkos, and C. Chamzas, "Skew Detection and Text Line Position Determination in Digitized Documents," *Pattern Recognition,* vol. 30, pp. 1505-1519, 1997.

[15] *Armi Professional for Windows 95/98 and NT Version 5.0,* Hapsan Computer Inc., Seoul, Korea, 1999.

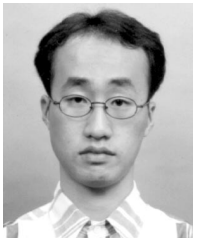
[16] *OmniPage Pro for Windows 95/98 and NT Version 9.0,* Caere Corporation, Los Gatos, Calif., 1998.

[17] I. Phillips, S. Chen, and R. Haralick, "CD-ROM Document Database Standard," *Proc. Second Int'l Conf. Document Analysis and Recognition,* pp. 478-483, 1993.

[18] J. Sauvola and H. Kauniskangas, "MediaTeam Document Database," *A CD-ROM Collection of Document Images,* Univ. of Oulu, Finland, 1999.

[19] R. Haralick, "Document Image Understanding: Geometric and Logical Layout," *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition,* vol. 8, pp. 385-390, 1994.

**Seong-Whan Lee** received the BS degree in computer science and statistics from Seoul National University, Seoul, Korea, in 1984, the MS and PhD degrees in computer science from KAIST in 1986 and 1989, respectively. From February 1989 to February 1995, he was an assistant professor in the Department of Computer Science at Chungbuk National University, Cheongju, Korea. In March 1995, he joined the faculty of the Department of Computer Science and Engineering at Korea University, Seoul, Korea, as an associate professor, and now he is a full professor. Currently, Dr. Lee is the director of National Creative Research Initiative Center for Artificial Vision Research (CAVR) supported by the Korean Ministry of Science and Technology. He was the winner of the Annual Best Paper Award of the Korea Information Science Society in 1986. He obtained the First Outstanding Young Researcher Award at the Second International Conference on Document Analysis and Recognition in 1993 and the First Distinguished Research Professor Award from Chungbuk National University in 1994. He also obtained the Outstanding Research Award from the Korea Information Science Society in 1996. He has been the co-editor-in-chief of the International Journal on Document Analysis and Recognition since 1998 and an associate editor of the *Pattern Recognition Journal*, the *International Journal of Pattern Recognition and Artificial Intelligence*, *International Journal of Image and Graphics*, and the *International Journal of Computer Processing of Oriental Languages.* He was the program co-chair of the Sixth International Workshop on Frontiers in Handwriting Recognition, the Second International Conference on Multimodal Interface, the 17th International Conference on Computer Processing of Oriental Languages, the Fifth International Conference on Document Analysis and Recognition, and the Seventh International Conference on Neural Information Processing. He was the workshop co-chair of the Third International Workshop on Document Analysis Systems and the First IEEE International Workshop on Biologically Motivated Computer Vision. He served on the program committees of several well-known international conferences. He is a fellow of IAPR, a senior member of the IEEE, and a life member of the Korea Information Science Society and the Oriental Languages Computer Society. His research interests include pattern recognition, computer vision, and neural networks. He has more than 150 publications on these areas in international journals and conference proceedings, and authored five books.



**Dae-Seok Ryu** received the BS degree in computer engineering from Seokyeong University, Seoul, Korea, in 1998, and the MS degree in computer science and Engineering from Korea University, Seoul, Korea, in 2000. He is currently working as a research engineer at Hyundai Information Technology, Co., Ltd. in Korea. His research interests include document image analysis, pattern recognition, and image processing.

▷ **For more information on this or any other computing topic, please visit our Digital Library at** http://computer.org/publications/dlib.

Authorized licensed use limited to: Korea University. Downloaded on October 25, 2008 at 01:45 from IEEE Xplore. Restrictions apply.