문서영상에서 표 구성 직선과 데이터 추출

장대근* · 김의정**

The Extraction of Table Lines and Data in Document Image

Dae-geun Jang* · Eui-jeong Kim**

요 약

문서영상에서 표 영역을 분류하고 구조를 파악하려면 표를 구성하는 직선과 데이터를 추출할 수 있어야 한다. 그러나 영상 입력 장치의 오차나 영상축소로 인해 표를 구성하는 직선이 끊어지거나 길이가 변하며 직선에 노이즈나 문자가 붙어 표로부터 직선과 데이터의 정확한 추출이 어렵다.

본 논문에서는 1차원 메디안 필터를 이용하여 표를 구성하는 수평선과 수직선을 추출한다. 1차원 메디안 필터는 필터링 방향의 직선을 추출하는 과정에서 노이즈와 필터링 방향에 수직한 직선을 제거할 뿐 아니라 직선의 끊어진 부분이 필터 탭 길이보다 짧은 경우 끊어진 부분을 연결한다. 또한 수직선을 추출하는 과정에서 직선에 붙어 있던 문자들을 분리함으로써 상용제품을 포함한 기존의 방법에 비해 표 영역 분류 및 구조 분석을 위한 직선과 데이터 추출이 우수한 방법을 제안한다.

ABSTRACT

We should extract lines and data which consist of the table in order to classify the table region and analyze its structure in document image. But it is difficult to extract lines and data exactly because the lines are cut and their lengths are changed, or characters or noises are merged to the table lines. These problems result from the error of image input device or image reduction.

In this paper, we propose the better method of extracting lines and data for table region classification and structure analysis than the previous ones including commercial softwares. The proposed method extracts horizontal and vertical lines which consist of the table by the use of one dimensional median filter. This filter not only eliminates the noises which attach to the line and the lines which are orthogonal to the filtering direction, but also connects the cut line of which the gap is shorter than the length of the filter tap in the process of extracting lines to the filtering direction. Furthermore, texts attached to the line are separated in the process of extracting vertical lines. This is an example of ABSTRACT format.

키워드

문자인식, 문서영상 왜곡보정, 영상처리

character recognition, distortion correction of document image, image processing

* 특허청 전지전자심사본부 전기심사팀

** 공주대학교 컴퓨터교육과 (교신저자)

접수일자 : 2005. 10. 6

Ⅰ. 서 론

정보화와 더불어 전자문서의 사용이 증가함에 따라 인쇄문서의 사용은 감소할 것이라는 예상과는 달리 프린터와 같은 컴퓨터를 이용한 출력장치들의 개발로 인해 오히려 예전보다 인쇄문서의 양은 더욱 늘어나고 있는 추세다. 따라서 대량의 문서를 자동으로 처리하기 위해 인쇄문서를 직접 손으로 입력하지 않고 편집 가능한 전자문서로 자동 전환이 필요하다. 이를 위해서는 문서내의 문자, 그림 ,표 등의 분류를 정확히 해야 하며 특히 표는 문서에서 중요한 정보를 가지고 있는 경우가 많아 정확한 추출 및 해석이 필요하다.

문서영상에서 표를 추출하려면 그 구조 분석이 이루어져야 하며 표를 구성하는 선들 사이의 연결점들을 찾아 구조를 파악하는 것이 일반적인 방법이다 [1]-[2]. 그리고 연결점 파악을 위해서는 표를 구성하는 직선부분을 정확히 추출하는 것이 선행되어야 하며 추출방법으로는 연결요소 기반의 상향식 영역분할을 이용하는 방법이 있다 [3]. 이 방법은 표를 구성하는 직선부분을 하나의 연결요소로 추출할 수 있어 직선들 사이의 연결점들을 찾기에 적합한 방법이다. 그러나 영상 입력장치의 오차나 영상 축소로 인해 표를 구성하는 직선이 끊어지거나 길이가 변화된 경우 또는 직선에 문자나 노이즈가 붙어 있는 경우 연결점들을 정확하게 파악하기 어렵다. 이와 같은 경우 끊어진 직선을 복구하는 방법 [4]와 직선에 붙어 있는 문자를 분리하기 위한 방법 [5]를 이용할 수도 있으나 이러한 방법들을 추가해도 언급한 문제들을 완벽하게 해결하지 못하며 계산량의 증가를 가져오는 문제점이 있다.

제안한 방법에서는 연결요소 기반의 상향식 영역분할을 이용하여 표를 구성하는 직선들을 하나의 연결요소 형태로 추출하며 1차원 메디안 필터를 이용하여 추출한 연결요소로부터 수평선과 수직선을 분리한다. 메디안 필터는 자신과 주변의 화소값 가운데 중간값을 택하는 필터링으로 impulse noise를 제거한다. 따라서 적절한 탭 크기를 갖는 1차원 메디안 필터를 표 영역에 적용하면 직선에 붙어 있는 노이즈뿐 아니라 필터링 방향과 수직 방향의 직선은 제거되고 같은 방향의 직선은 추출되며 끊어진 부분이 필터 탭 크기보다 짧은 직선은 중간값을 택하는 필터링의 특성으로 인해 연결된다. 따라서 메디안 필터를 이용하면 언급한 문제점들을 해결하기 위한 부가적 처리과정이 필요 없이 표를 구성하는 수평선과 수직선을 분리할 수 있다. 또한 1차원 메디안 필터를 이용하여 수직선을 추출하는 과정에서 직선에 붙어 있던 문자들의 분리도 가능하다. 그리고 메디안 필터로 추출한 직선을 그대로 이용하지 않고 5단계로 구성한 검사과정에서 표를 구성하는 직선에 관계된 조건들을 적용함으로써 오판된 직선을 제거하고 직선의 길이를 표 구조에 맞게 조정함으로써 표 영역 분류 및 구조 분석을 위한 직선과 데이터의 추출 성능이 상용제품을 포함한 기존의 방법에 비해 우수하다[1].

Ⅱ. 제안한 방법의 전체 구성

제안한 표 영역 분류 및 구조 분석을 위한 직선과 데이터 추출 방법의 전체 구성을 그림1에 나타내었다. 구성은 크게 입력영상을 축소하는 전처리와 문자와 표 후보 영역을 추출하기 위한 영역분할, 표 영역 및 구성성분 추출의 3단계 과정으로 나누어진다. 영역분할은 연결요소 기반의 상향식 분할을 고속화한 방법으로 수평방향 라인 단위로 연결요소를 생성하는 과정과 이웃한 라인간 연결요소들을 결합하여 완성된 형태를 생성하는 2단계 과정으로 구성한다[3]. 표 영역 및 구성성분 추출은 문자 추출 및 최대 문자 크기 결정에서 데이터 추출까지 6단계의 처리과정을 거친다. 이중 문자 추출에서는 2가지 검사과정을 이

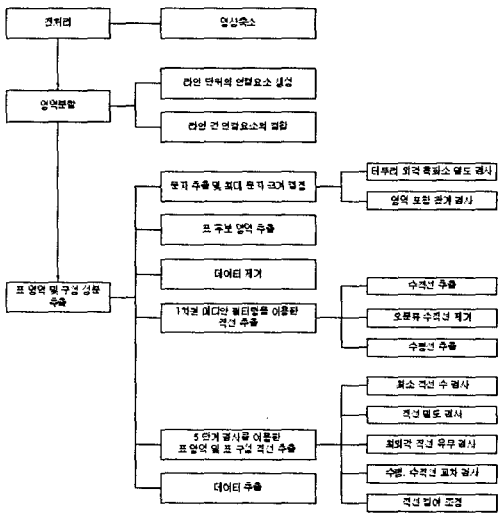


그림 1. 표 영상에서 직선과 데이터 추출을 위한 제안한 방법의 전체 구성도

Fig. 1 Block diagram of proposed method for line and data extraction in table image

용하며, 1차원 메디안 필터를 이용한 직선 추출에서는 수직선 추출에서 수평선 추출까지 3단계 과정을 수행한다. 5단계 검사를 이용한 표 영역 및 표 구성 직선 추출과정은 전단계에서 메디안 필터를 이용하여 추출한 직선들을 대상으로 표와 표를 구성하는 직선이 갖추어야 할 조건들을 적용하여 검사하는 과정이다.

Ⅲ. 제안한 표 영역 분류 및 구조 분석을 위한 직선과 데이터 추출 방법의 처리과정

1. 전처리

A4 크기의 문서를 300dpi 해상도의 이진 문서영상으로 만들 경우 2000×3000 pixels 이상의 크기를 갖는 문서영상이 된다. 이 정도 크기의 문서영상을 PC를 이용하여 실시간으로 영역분할하기에는 계산량이 너무 많다. 따라서 제안한 방법에서는 전처리로 입력영상을 1000×1000 pixels 보다 작은 크기로 축소하여 사용한다.

영상의 축소 과정은 다음과 같다. 수식 ①을 이용하여 축소비율(r)을 구한 다음 입력영상을 r× r 크기의 영역으로 분할한다. 분할된 각 영역은 내부에 존재하는 모든 화소 값들을 수식 ②와 같이 OR 연산하여 그 값(s)이 1이면 즉 1개 이상의 흑화소가 존재하면 r× r크기의 해당 영역을 1개의 흑화소로 표현하고 나머지 경우는 1개의 백화소로 표현하여 입력 영상을 축소한다.

$$r = \text{Quantize}\left[\frac{MaxLen}{1000}\right] \quad ①$$

$$s = p[sx, sy] \oplus \cdots \oplus p[sx+x, sy+y] \oplus \cdots p[sx+r-1, sy+r-1] \quad ②$$

r : 축소비율
Quantize [α] : α를 반올림하여 정수화
MaxLen : 입력 문서영상의 가로, 세로 길이 중 큰 값(pixel 단위)
p[*x*, *y*] : 좌표 *x*, *y* 에서의 화소값 (0 or 1)
sx, *sy* : 해당 영역의 시작 좌표
⊕ : 논리연산 OR

2. 영역분할

역분할은 문자와 표 후보 영역을 추출하기 위하여 기존의 방법을 사용한다[3]. 이 방법은 연산량을 줄여 빠른 처리가 가능하도록 개선한 연결요소 기반의 상향식 영역분할 방법이다. 처리과정은 그림 1과 같이 연결요소를 라인 단위로 생성하는 과정과 이웃하는 라인간 연결요소들을 결합하여 완성된 형태의 영역을 생성하는 과정으로 구성된다.

1) 라인 단위의 연결요소 생성

연결요소란 임의의 흑화소 또는 백화소에 대하여 8방향 또는 4방향의 인접 화소가 같을 경우 이 화소들을 모두 연결하여 얻은 화소의 집합이다. 따라서 같은 화소값을 가진 점들을 연결하기 위하여 주변 화소들과 많은 비교를 수행해야 하므로 계산량이 많아 실시간 처리가 어렵다.

X. Li의 방법 [3]에서는 주변 화소들과의 비교 횟수를 줄이기 위하여 가로방향으로만 같은 화소값을 가진 점들을 연결하여 연결요소를 생성한다.

2) 라인간 연결요소의 결합

제안한 방법에서는 글자를 구성하는 흑화소 집합인 연결요소들을 찾기위하여 연산량이 많은 4 또는 8방향 탐색 대신, 가로방향 라인 단위로 흑화소의 연결요소들을 생성하고 인접한 라인간의 연결요소들을 결합하는 방법을 사용한다. 그 과정은, 첫 번째 라인부터 차례로 기준라인(base line)과 그 다음라인을 비교라인(comparative line)으로 설정하고, 그림 2의 예에서와 같이 수식 ③을 만족하는 즉 서로 연결 관계가 있는 두 라인간 연결요소들을 결합함으로써 완성된 형태를 생성한다. 그림 2의 경우 2개 라인의 인접 라인의 연결요소들과 수평방향으로 교차되는

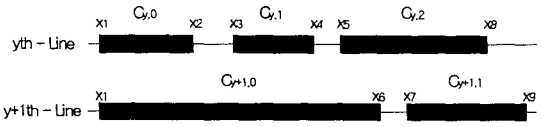


그림 2. 인접한 2개의 라인에서 연결요소 예
Fig. 2 Example of connected components between neigbored lines

$$\min[\max(c_{y,m}), \max(c_{y+1,n})] \geq \max[\min(c_{y,m}), \min(c_{y+1,n})] \quad ③$$

$C_{y,i}$: y번째 라인의 i+1 번째 연결요소

TABLE 2
VALUES OF S, Z, M, C AND SIMILARITY RANKS FOR THE STAR DEMICAME OF FIG. 3 FOR TEST IMAGES ROTATED BY α DEGREES.

α	S	Z	M	C	Similarity rank
0	1.65	0.59	0.94	0.998	0.942
3	1.98	0.60	0.94	0.998	0.940
5	2.75	0.66	0.95	0.996	0.938
9	3.92	0.75	0.94	0.985	0.920
12	5.08	0.81	0.92	0.995	0.889
15	6.50	0.91	0.89	0.995	0.810
18	8.66	0.98	0.86	0.995	0.621
21	10.61	1.06	0.82	0.995	0.398
24	12.64	1.64	0.59	0.978	0.005
27	14.88	1.90	0.61	0.970	0.002

그림 3. 영역분할 수행 결과 예
Fig. 3 Example of result of region segmentation

관계에 있으므로 모든 연결요소들은 하나로 결합되는 결과가 된다. 그림 3은 1)과 2)의 과정을 거쳐 영역분할을 수행한 결과 이다. 즉 글자를 구성하는 연결된 흑화소의 덩어리인 자소가 하나의 완성된 형태의 연결요소를 추출됨을 확인할 수 있고, 이들을 검은색 사각형 테두리로 둘러싼 영역으로 나타내었다.

3. 표 영역 추출

표의 형태는 다양하므로 인간의 시각으로 표를 판별하는 경우에도 구분이 어려운 경우가 있다. 따라서 표를 구분하기 위해서는 먼저 표의 범주를 명확히 정의하여야 할 필요가 있으며 제안한 방법에서는 표 전체를 관통하는 수평선과 수직선이 각각 3개 이상, 최외각을 구성하는 4개의 직선 중 3개 이상 존재하는 경우만 표로 정의하고 나머지는 차트나 그래프로 분류한다. 그림 1에서와 같이 표 영역 및 구성성분 추출은 6단계의 과정으로 구성되며 각각의 기능은 아래와 같다.

1) 문자 추출

2에서 추출한 영역 중 문자 영역을 분류하기 위하여 영역 테두리에 가장 인접한 외각 부분의 백화소 밀도(do)와 영역 내부에 다른 영역이 포함되어 있는지를 검사하여 문자 영역을 판단한다.

(1) 영역 테두리 외각 흑화소 밀도 검사

그림 3의 예를 보면 이진화된 문서영상에서 글자를 구성하는 흑화소 주변은 백화소로 구성되어 있다. 따라서 글자를 둘러싸는 테두리 영역인 R 영역(dotted regions)의 백화소 밀도(do)를 수식 ④를 이용하여 계산하여 do > 0.98인 영역을 문자 영역이 될 수 있는 후보로 추출한다.

흑화소 밀도 검사 구간(R)
그림 4. 흑화소 밀도 검사 구간 표시 예
Fig. 4 Example of the region for investigation of density of black pixel

$$d_0 = \frac{n_w(R)}{n_p(R)} \quad ④$$

$n_w(R)$: R 영역 백화소 수, $n_p(R)$: R 영역 총화소 수

(2) 영역 포함 관계 검사

그림, 표, 차트, 그래프는 내부에 문자와 같은 작은 조각 영역들을 포함하고 있으나 문자 영역의 경우 대부분 내부에 문자 영역을 포함하고 있지 않다. 따라서 (1)에서 추출한 후보 영역들 중 내부에 또 다른 영역을 포함하고 있는 것은 제외함으로써 문자 영역을 추출한다.

(3) 최대 문자 크기(L_{maxch}) 결정

최대 문자 크기(L_{maxch})는 표를 구성하는 직선 추출에서 필터 탭 크기를 결정하는 기준이 되며 추출한 문자 영역 중 가장 긴 세로의 길이를 문자의 최대 크기로 지정한다.

2) 표 후보 영역 추출

많은 문서영상을 대상으로 실험한 결과 표 영역 내부의 데이터를 제외한 흑화소 비율은 표 영역 전체 면적의 10 % 이하가 대부분이다. 그리고 음영이 포함된 표의 경우는 25% 이하인 경우가 대부분이다. 따라서 문자 영역을 제외한 나머지 영역 중 내부의 데이터를 제외한 흑화소 비율이 25% 이하인 영역들을 후보 영역으로 추출한다.

3) 표 내부 데이터(문자) 제거

표를 구성하는 직선을 추출하는 과정 4)에서는 영역내부의 문자 중 'I', '—'와 같은 획들도 직선으로 분류되므로 내부의 문자를 먼저 제거하는 것이 필요하다. 따라서 표를 구성하는 직선을 추출하기 전에 내부의 데이터를 먼저

제거한다.

4) 1차원 메디안 필터를 이용한 직선 추출

2)에서는 표를 구성하는 직선들을 하나의 연결요소 형태로 추출하였으며 표의 구조를 파악하기 위하여 추출한 연결요소로부터 수평선과 수직선을 분리한다. 제안한 방법에서는 필터 탭 크기가 $\frac{L_{maxch}}{2}$인 1차원 메디안 필터를 수평, 수직방향으로 각각 적용하여 수행 방향의 직선을 추출한다. 메디안 필터를 이용한 직선 추출은 서론에서 언급한 것과 같이 직선에 노이즈나 문자가 붙어있는 경우나 직선이 끊어지거나 길이가 달라진 경우에도 별도의 처리과정이 필요 없이 직선의 추출이 가능하다는 장점이 있다.

그림 6은 선이 끊어지고 길이가 변한 그리고 문자와 노이즈가 선에 붙어있는 표를 대상으로 1차원 메디안 필터를 이용하여 수평, 수직선을 추출한 예이다. (a)는 입력 영상이고 (b)는 영역 내부의 데이터를 제거한 후 표를 구성하는 직선들을 하나의 연결요소 형태로 추출한 결과로 직선에 붙어 있는 문자나 노이즈가 분리되지 않으며 선이 끊어지고 길이가 변한 부분이 그대로 있음을 확인할 수 있다. 수평, 수직선은 그림 6의 (b)의 결과에 1차원 메디안 필터를 적용하여 추출하며 과정은 3단계로 다음과 같다.

(1) 수직선 추출

수직선은 1차원 메디안 필터링을 수직방향으로 수행한 후 2의 영역분할을 해당 영역 부분에만 적용함으로써 각각의 수직선을 사각형 영역으로 추출한다. 그림 6의 (c)는 수직선을 추출한 예로 선의 끊어진 부분이 연결되고 선에 붙어 있던 문자들이 선에서 분리됨을 확인할 수 있다.

(2) 오분류 수직선 제거

(1)에서는 선에 붙어있던 문자들도 수직선으로 추출되므로 각각의 수직선과 수평방향으로 평행선상에 있는 문자들과의 세로방향 길이를 비교함으로써 오판된 수직선들을 제거한다. 그림 5를 예로 보면 'L'은 선에 붙어있던 문자가 (1)의 과정에서 수직선으로 추출된 예로 제거하려는 대상이 된다. 그리고 '가'는 'L'과 평행선상에 있는 문자이다. 먼저 수식 ③을 사용하여 'L'과 '가'가 가로방향으로 평행선상에 있는지를 확인한다. 수식 ③을 그림 5에 적용한 것이 수식 ⑤이고 이 식을 만족하는 문자는 해당 직선과 수평방향으로 평행선상에 있게 된다. 다음은 직선 'L'의 세로길이 dL과 평행선상에 있는 문자의 세로길이 dC를 비교한 수식 ⑥를 만족하면 직선 'L'은 제거한다. 즉 문자가 수직선으로 오분류된 경우이다. 그림 6의 (d)는 선에 붙어있던 문자들이 수직선으로 추출된 경우를 제거한 결과이다.

그림 5. 수직선과 수직방향으로 평행한 문자 예
Fig. 5 Character example which parallels with the vertical line

$$\min[\max(C_{y1}, C_{y2}), \max(L_{y1}, L_{y2})] \geq \max[\min(C_{y1}, C_{y2}), \min(L_{y1}, L_{y2})] \quad ⑤$$

$$d_L \leq (1.2 \times C_{y2} - C_{y1}) \quad ⑥$$

(3) 수평선 추출

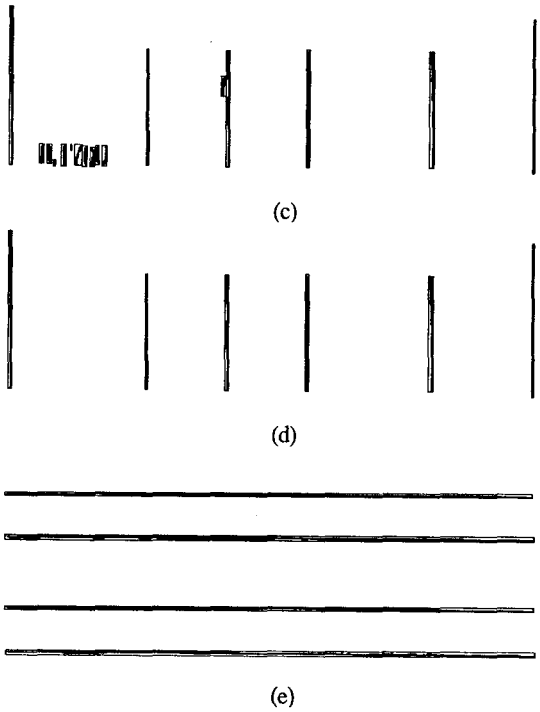
수평선은 1차원 메디안 필터링을 수평방향으로 수행한 후 2의 영역분할을 해당 영역 부분에만 적용함으로써 각각의 수평선을 사각형 영역으로 추출하며 (2)에서 오분류된 수직선 영역을 포함하는 수평선은 추출하지 않는다. 따라서 직선에 붙어 있는 문자는 제외되고 표를 구성하는 수평선만 추출된다. 그림 6의 (e)는 수평선을 추출한 결과이다.

Canonical Discriminant Functions				
Eigenvalue	Pct of	Cum	Canonical	Wilks'
8.1760	100.00	100.00	0.9439	0.1090

(a)

	f			
8.1760				

(b)



(a) 입력 영상
(b) 영역 내부 데이터를 제거한 후 표 구성 직선들을 하나의 연결요소 형태로 추출한 결과
(c) 1차원 메디안 필터와 영역분할을 이용한 수직선 추출 결과
(d) 선에 붙어있던 문자들이 수직선으로 추출된 경우를 제거한 결과
(e) 1차원 메디안 필터와 영역분할을 이용한 수평선 추출 결과

그림 6. 1차원 메디안 필터를 이용한 표 구성 수평, 수직선 추출
Fig.6 The result of horizontal and vertical line extraction using 1-D median filter

5) 5단계 검사를 이용한 표 영역 및 구성 직선 추출

영상 입력장치의 오차와 영상축소, 메디안 필터링으로 인해 직선의 길이가 변화된 경우 발생하므로 4)의 과정에서 추출한 직선들의 길이를 조정할 필요가 있다. 또한 직선을 추출한 표 후보 영역이 차트나 그래프인 경우 표가 아닌 것을 확인할 수 있는 처리과정이 필요하다. 따라서 4)의 과정에서 추출한 직선들을 5단계 검사과정을 이용하여 표를 구성하는 직선과 표의 여부를 결정하며 처리과정은 아래와 같다.

(1) 최소 직선 수 검사

추출한 직선들이 3에서 정의한 표를 구성하기 위한 최소 직선 수 조건의 만족여부를 검사한다. 즉 표 전체 길이의 90% 이상 되는 수평선과 수직선이 각각 3개 이상인지를 검사하여 표를 구분한다. 기준이 90%인 것은 입력 장치의 오차로 인해 선의 길이가 짧아진 경우가 있기 때문이다.

(2) 표 외각 테두리선 검사

상용제품을 포함한 기존 방법에서는 차트나 그래프 중 일부는 구조가 표와 비슷하여 표로 오인식되는 경우가 있다[1]. 따라서 2에서 표의 범주를 제한하여 표를 구성하는 외각 테두리선 4개 중 3개 이상 있는 경우만 표로 정의하였으며 이 조건을 이용하여 표를 구분한다. 외각 테두리선의 추출 방법은 영역의 외각 경계로부터 탐색거리 dL_{ave} 내에 테두리선이 있는지를 검사하며 탐색거리 dL_{ave}는 4)에서 추출한 선들의 두께를 평균한 값에 2를 곱한 값이다.

$$dL_{ave} = 2 \times \frac{\sum_{i=0}^{n(L)-1} dL(i)}{n(L)} \quad ⑦$$

$dL(i)$: i번째 직선 두께, $n(L)$: 총 직선 수

(3) 수평, 수직선 교차 검사

표를 구성하는 수평선은 반드시 2개 이상의 수직선과 교차하며 수직선도 마찬가지다. 이 조건을 이용하여 표 영역에서 오분류된 직선과 표가 아닌 그래프나 차트 영역에서 추출된 직선을 제거한다.

(4) 직선 길이 조정

표를 구성하는 수평선은 수직선과 만나는 점에서 끝나야하고 수직선의 경우도 마찬가지다. 그러나 입력 장치의 오차로 인해 끊어진 선이나 실제 길이보다 길거나 짧은 선이 나타난다. 따라서 추출한 직선의 길이를 표의 구조에 맞게 수평선과 수직선이 만나는 점에서 끝나도록 조정한다.

6) 데이터 추출

데이터는 2의 영역분할 과정에서 추출한 표 내부의 문

자들을 말한다. 그러나 직선에 붙은 문자는 2의 영역분할 과정에서 문자로 추출되지 않으므로 그림 6의 (c)와 같이 직선에 붙은 문자가 수직선으로 추출된 부분에 해당하는 영역을 문자 영역으로 처리하여 데이터로 추출한다. 그림 7은 그림 6의 (a) 영상으로부터 표를 구성하는 데이터를 추출한 결과이다.

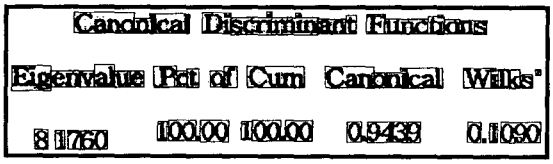


그림 7. 표를 구성하는 데이터 추출 결과 예
Fig. 7 Example of extraction result of data which consist of table

Ⅳ. 실험 및 고찰

제안한 방법은 상용제품을 포함한 기존의 방법과 성능을 비교하였다[1]. 성능평가는 표를 포함하고 있는 문서영상 20장(총 표 수 39개)을 대상으로 표 영역 추출과 표를 구성하는 직선 추출의 정확성을 비교하였다. 직선 추출에서는 추출한 직선이 끊어지거나 표의 구조가 바뀔 정도로 길이가 변한 경우는 오류로 처리하였다. 또한 총 39개의 표 중 끊어진 직선과 직선에 문자가 붙어 있는 경우를 포함하는 표 10개를 대상으로 실험한 결과도 기록하였다.

표 1은 제안한 방법을 기존의 방법 [1]과 상용제품인 P사의 A 6.0 (국내제품), Scansoft사의 Omni page pro 11.0, ABBYY Software House사의 Fine Reader 5.0 Office와 성능을 비교한 결과로 총 표 수는 39개이고, 39개의 표를 구성하는 총 직선 수는 563개 이다.

표 1. 상용제품을 포함한 기존 방법과의 인식 성능 평가 결과[1]
Table.1 The result of comparison of recognition performance with previous methods containing commercial products

구분 \ 제품	인식수(인식률 %)				
	X. Li's Method [1]	A 6.0	Omni page pro 11	Fine Reader 5.0	제안한 방법
표	38(97)	39(100)	38(97)	36(92)	39(100)
직선 (수평선+수직선)	536(95)	528(93)	495(88)	512(91)	557(99)

표 2는 끊어진 직선과 직선에 문자가 붙어 있는 경우를 포함하는 표 10개를 대상으로 실험한 결과이며 10개의 표를 구성하는 총 직선 수는 142개이다.

표 2. 끊어진 직선과 직선에 문자가 붙어 있는 경우를 포함하는 표의 인식 성능 평가 결과
Table. 2 The result of recognition performance in case of existing of broken lines and characters crossing to them

구분 \ 제품	인식수(인식률 %)				
	X. Li's Method [1]	A 6.0	Omni page pro 11	Fine Reader 5.0	제안한 방법
표	8(80)	9(90)	7(70)	7(70)	10(100)
직선 (수평선+수직선)	130(92)	128(90)	123(87)	126(89)	138(97)

Ⅴ. 결 론

본 논문에서는 문서영상에서 표 영역을 분류하고 구조를 분석하기 위한 표를 구성하는 직선과 데이터의 추출 방법을 제안하였다. 표를 구성하는 수평선과 수직선 추출에서 1차원 메디안 필터를 이용함으로써 직선에 노이즈나 문자가 붙어 있는 경우에도 직선 추출이 가능하였으며 끊어진 부분이 있는 직선도 연결된 형태로 추출이 가능하였다. 또한 메디안 필터링으로 추출한 직선을 그대로 이용하지 않고 5단계로 구성한 검사과정에서 표와 표를 구성하는 직선에 관계된 조건들을 적용함으로써 오판된 직선을 제거하고 직선의 길이를 표 구조에 맞게 조정하였으며 표와 구조가 비슷한 차트나 그래프 영역이 표로 오판되는 것을 방지하였다. 그리고 수직선을 추출하는 과정에서 직선에 붙어 있는 문자들을 분리함으로써 표를 구성하는 직선과 데이터의 추출 성능이 상용제품을 포함한 기존의 방법에 비해 우수함을 확인 할 수 있었다[1].

제안한 방법의 성능을 상용제품을 포함한 기존의 방법과 비교 시험한 결과 표 인식률 100%, 표를 구성하는 직선 추출률 97-99%로 비교대상들보다 우수함을 확인 할 수 있었다[1]. 하지만 제안한 시스템이 인간이 수행하는 것과 같은 수준의 성능을 발휘하기 위하여 보완해야 할 점들을 보면 표의 범주를 확대하여 더 다양한 형태의 표를 인식하도록 하는 문제와 표 내부에 다시 표의 구조가 있는 경우 표를 분석하는 문제에 대한 개선책이 필요하다.

참고문헌

[1] X. Li, J. Hong, Z. Zhang and B. Chen, "A Statistical Form Reading System," *Proc. IEEE Region 10 Conf. Computer, Communication, Control and Power Engineering,* vol.2 pp.1062-1065, 1993.

[2] L. A. Pereira and J. Facon, "Methodology of Automatic Extraction of Table-form Cells," *Proc. 8th Brazilian Symp. Computer Graphics and Image Processing,* pp.15-21, 2000.

[3] X. Li, W. Gao, S. Y. Chi, K. A. Moon and H. J. Kim, "An Efficient Method for Page Segmentation," *Proc. ICICS,* vol.2, pp.957-961, 1997.

[4] L. huizhu, G. Agam and I. Dinstein, "Directional Mathematical Mophology Approach for Line Thinning and Extraction of Character Strings from Maps and Line Drawings," *Proc. 3th Int. Conf. Document Analysis and Recognition,* vol.1 pp.257-260, 1995.

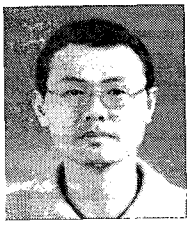
[5] Jain-Shiue Chen and Din-Chang Tseng, "Overlapped Charter Separation and Reconstruction for Table-form Documents," *Proc. Int. Conf. Image Processing,* vol.1 pp.233-236, 1996.

[6] Ren Jean Liou and Mu-Song Chen, "Recognition of Table-form Documents Using High Order Correlation Method," *Proc. Int. Joint Conf. Neural Networks,* vol.3, pp.1851-1856, 1998.

[7] T. Watanabe, Q. Luo and N. Sugie, "Layout Recognition of Multi-Kinds of Table Form Documents," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol.17, no.4, pp.432-445, 1995.

[8] D. Drivas and A. Amin, "Page Segmentation and Classification Utilizing Bottom-up Approach," *Proc. ICDAR,* pp.610-614, 1995.

저자소개



장 대 근(Dae-geun Jang)

2003년 경북대학교 전자공학과 박사
1996년~2005년 ETRI 선임연구원
2005년~현재 특허청 전기전자 심사본부 통신사무관

※관심분야 : 문서영상처리, 영상인식, 로봇비젼

김 의 정(Eui-jeong Kim)

1997년 충남대학교 컴퓨터공학과 박사
1997년~1998년 SERI 연구원
1998년~현재 공주대학교 컴퓨터교육과 교수

※관심분야 : 컴퓨터비젼, 패턴인식, 가상현실, Web-3D