한글 문서 인식을 위한 문서 영상에서의 문자와 그림의 분리 추출

o
이 인 동, 강 태 호, 권 오 석, 김 태 균

충남대학교 공과대학 전자계산기공학과

A study on the segmentation and extraction of the pictures and characters in korean document

In Dong Lee, Kang Tae Ho, Oh Seok Kwon, Tae Kyun Kim

* Dept. of Computer Eng., College of Eng., Chungnam National Univ.

요 약

한글 문서를 인식하기 위하여 문서 영상에서 문자와 그림을 분리 추출하기 위한 방법에 대하여 논하였다. 분리 추출 방법으로는 실시간으로 입력되는 영상 데이타로부터 문자와 그림의 경계 위치를 알아내는 방법을 사용하였다. 한글, 영문, 한자, 기호 등의 문자와 그림이 혼합된 A4 크기의 문서 영상을 300 DPI의 해상도로 입력 받아 실험하였다. 단 한번의 주사만으로 모든 문자와 그림이 정보 흐름의 순서에 따라 분리 추출되었다. 실험 결과 본 방법은 최소한의 시간과 최소한의 기억 용량으로 완벽한 분리 추출이 가능함을 보였다.

1. 서 론

산업이 다양화되고 전문화되면서 정보의 양이 급증하고 있다. 그러나 컴퓨터에 정보를 입력하는데는 대부분 키보드에 의존하고 있다. 이는 급속하게 증가하고 있는 정보를 처리하는데 한계가 있으며, 영상 입력 장치를 사용한 문서 자동 입력 장치의 개발이 절실히 요구되고 있다. 문서 자동 입력 장치를 개발함에 있어서 문서 영상에서 문자 영역과 그림 영역을 분리한 다음, 분리된 문자 영역에서 개별 문자를 분리 추출하는 연구는 문서 인식의 전처리 단계로서 반드시 선행되야 할 중요한 연구에 해당된다. 하지만 [1,2,3,4,5] 등의 외국 논문과 [6,7,8]의 국내 논문이 발표됐을뿐 관련 연구가 매우 미약하며, 결과 또한 처리 방법 및 처리 속도상의 문제로 문서를 인식하는데 여러가지 제약이 있어 실용화하는데 어려움이 많다.

본 연구에서는 실시간으로 입력되는 문서 영상 데이타로부터 문자와 그림의 경계 위치를 알아내어 문자와 그림을 동시에 분리 추출할 수 있는 방법을 제안하였다. 본 방법의 유효성을 확인하기 위하여 한글, 영문, 한자, 기호 등의 인쇄체 문자와 그림이 혼합되어 가로체로 작성된 A4 크기의 문서 영상을 SCANNER를 통하여 300 DPI의 해상도로 입력 받아 실험하였다. 실험결과 모든 문자와 그림이 정보 흐름의 순서에 따라 분리 추출되었으며, 실시간으로 입력되는 영상 데이타에 대하여 단 한번의 주사로 모든 문자와 그림을 분리 추출할 수 있기 때문에 최소한의 기억용량과 최소한의 처리 시간에 완벽한 분리 추출이 가능함을 확인하였다.

2. 문자와 그림의 분리 추출

실시간으로 입력되는 한글, 영문, 한자, 기호 등의 문자와 그림이 혼합된 문서의 영상 데이타로부터 문자와 그림을 분리 추출하기 위하여 매 주사열에 흑화소의 포함 여부를 나타내는 L_CHK, 개별 문자의 좌우 경계 위치 결정을 위한 배열 CHAR(i) (i=1...n, n=한 열의 화소수), 그림의 좌우 경계 위치 결정을 위한 배열 GRAP(i), 주사열에 포함된 개별 문자수 C_NUM, 주사열에 포함된 그림수 G_NUM, 문자열의 상측 경계 위치 CY_ST, 그림의 상측 경계 위치 GY_ST, 문자열의 하측 경계 위치 CY_ED, 그림의 하측 경계 위치 GY_ED, 개별 문자의 좌측 경계 위치 CX_ST, 그림의 좌측 경계 위치 GX_ST, 개별 문자의 우측 경계 위치 CX_ED, 그림의 우측 경계 위치 GX_ED를 설정하였다. (그림 1)은 문자와 그림의 분리 추출 과정을 보인 것이다.

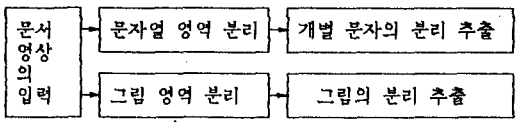

그림1. 문자와 그림의 분리 추출 과정

2.1 문자와 그림의 분리 추출 알고리즘

스케너를 통하여 실시간으로 입력되는 문서 영상 데이타로부터 다음과 같은 과정을 거쳐 문자와 그림을 분리 추출하였다.

1) 주사열의 흑화소 포함 여부 결정

문서 영상의 매열을 주사할 때마다 주사열의 상태를 조사하여 흑화소를 포함하고 있으면 L_CHK의 값에 1을, 주사열 전체에 단 1개의 흑화소도 포함되지 않았을 경우는 L_CHK의 값에 0을 설정한다.

2) 개별 문자의 좌우 경계 위치 결정용 배열의 상태값 설정

실시간으로 입력되는 문서 영상 데이타의 매 열을 주사하면서 흑화소를 만날 때마다 개별 문자의 좌우 경계 위치 결정용 배열 CHAR의 상태 값을 다음과 같이 설정한다.

CHAR(i)=CHAR(i) OR DATA(i), (i=1...n, n=한 열의 화소수)

3) 주사열에 포함된 문자수 결정

문서 영상의 매열을 주사할 때마다 배열 CHAR의 상태를 주사하여 흑화소로 연결된 부분을 개별 문자의 위치로 보고, 백화소에서 흑화소로 교번된 경우의 수를 카운트하여 주사열에 포함된 개별 문자수 C_NUM의 값으로 결정한다.

4) 주사열에 포함된 그림수 결정

문서 영상의 매열을 주사할 때마다 배열 CHAR의 상태와 주사열의 상태를 비교하여 배열 CHAR에서 흑화소로 연결된 부분과 일치하는 주사열의 해당 부분에 흑화소를 포함하고 있는 경우의 수를 카운트하여 주사열에 포함된 그림수 G_NUM의 값으로 결정한다. 즉 G_NUM는 C_NUM의 값에 배열 CHAR에서 흑화소로 연결된 부분과 일치하는 주사열의 해당부분에 흑화소를 포함하지 않은 경우의 수를 감한 값이 된다.

5) 문자열 영역과 그림 영역의 상측 경계 위치 결정

문서 영상의 매 열을 주사하여 처음 흑화소를 포함하고 있는 열을 문자열 영역의 상측 경계 위치 CY_ST와 그림 영역의 상측 경계 위치 GY_ST의 값으로 결정한다.

6) 문자열 영역과 그림 영역의 하측 경계 위치 결정

문서 영상의 매 열을 주사하여 CY_ST와 GY_ST의 값이 결정되어 있고, 흑화소를 포함하고 있지 않은 열(L_CHK의 값이 0인 열)의 바로 전 열을 문자열 영역의 하측 경계 위치 CY_ED와 그림 영역의 하측 경계 위치 GY_ED의 값으로 결정한다.

7) 문자와 그림이 혼합된 경우 문자열의 하측 경계 위치 결정

문서 영상의 매 열을 주사할 때마다 C_NUM과 G_NUM의 값을 조사하여 CY_ST의 값이 결정되어 있고, C_NUM의 값이 10보다 크면서, G_NUM의 값이 1의 경우를 만나면 G_NUM의 값이 1보다 큰 값이 될때까지 주사하여 바로 전 열을 문자열과 그림이 혼합된 경우의 문자열 영역의 하측 경계 위치 CY_ED의 값으로 결정한다.

8) 그림 영역의 좌우 경계 위치 결정용 배열의 상태값 설정

문서 영상을 주사하여 CY_ST, CY_ED의 값이 결정되면, CY_ED로 결정된 열의 상태와 배열 CHAR의 상태를 비교하여 CHAR에서 흑화소로 연결된 부분과 일치하는 CY_ED열의 해당 부분에 흑화소를 포함하는 부분은 문자열과 그림이 혼합된 경우에서의 그림 부분으로보고 GRAP의 해당 부분의 값에 CHAR의 해당 부분의 값으로 다음에서와 같이 설정한다.

G=G+(CHAR(i) AND DATA(i)),

i=CHAR에서 흑화소로 연결된 그룹의 시작에서 끝위치까지,

IF G>=1 THEN GRAP(i)=CHAR(i)

9) 개별 문자의 좌우 경계 위치 결정

문서 영상을 주사하면서 CY_ST, CY_ED의 값이 결정되면, 배열 CHAR의 상태를 주사하여 CX_ST의 값이 초기값(0)이면서 흑화소를 만나면, 추출하고자하는 개별 문자의 좌측 경계 위치 CX_ST의 값으로 결정하며, CX_ST의 값이 결정되어 있으면서 백화소를 만나면 바로 전 행을 추출하고자하는 개별 문자의 우측 경계 위치 CX_ED의 값으로 결정한다. CY_ST, CY_ED, CX_ST, CX_ED의 값이 결정되면 경계 위치 내의 영상을 분리하여 저장한 다음 CX_ST, CX_ED를 초기값(0)으로 재설정한다. 배열 CHAR에서 주사된 부분은 주사후 초기값(0)으로 재설정된다. 배열 CHAR의 끝까지 주사하여 주사열에 포함된 개별 문자가 모두 분리 추출되면, CY_ST, CY_ED, CX_ST, CX_ED의 값을 초기값(0)으로 재설정한다.

10) 그림 영역의 좌우 경계 위치 결정

문서 영상을 주사하면서 GY_ST, GY_ED의 값이 결정되면, 배열 GRAP의 상태를 주사하여 GX_ST의 값이 초기값(0)이면서 흑화소를 만나면, 추출하고자하는 그림의 좌측 경계 위치 GX_ST 의 값으로 결정하며, GX_ST의 값이 결정되어 있으면서 백화소를 만나면 바로 전의 행을 추출하고자하는 그림의 우측 경계 위치 GX_ED의 값으로 결정한다. GY_ST, GY_ED, GX_ST, GX_ED의 값이 결정되면 경계 위치 내의 영상을 분리하여 저장한 다음 GX_ST, GX_ED를 초기값(0)으로 재설정한다. 배열 GRAP에서 주사된 부분은 주사와 동시에 초기값(0)으로 재설정된다. 배열 GRAP의 끝까지 주사하여 주사열에 포함된 그림이 모두 분리 추출되면, GY_ST, GY_ED, GX_ST, GX_ED의 값을 초기값(0)으로 재설정한다.

2.2 분리 추출 알고리즘의 적용

그림이 포함되지 않은 문서 영상(그림2)과 그림이 포함된 문서 영상(그림3)에 대하여 본 분리 추출 알고리즘을 적용하여 영상의 매 열을 주사했을 때 각 결정값이 결정되어 개별 문자와 그림이 분리 추출되는 과정과 문자와 그림이 분리 추출된 결과를 (표1, 2, 3, 4)에 각각 표시하였다.

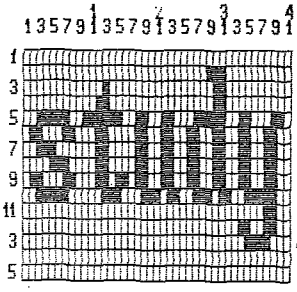

그림2. 그림이 포함되지 않은 문자열 영상의 예

주사열	L_CHK	C_NUM	G_NUM	CY_ST	CY_ED	GY_ST	GY_ED
1	0	0	0	0	0	0	0
2	1	1	1	2	0	2	0
3	1	2	2	2	0	2	0
4	1	2	2	2	0	2	0
5	1	7	7	2	0	2	0
6	1	7	7	2	0	2	0
7	1	7	7	2	0	2	0
8	1	7	7	2	0	2	0
9	1	7	7	2	0	2	0
10	1	5	5	2	0	2	0
11	1	5	5	2	0	2	0
12	1	5	5	2	0	2	0
13	1	5	5	2	0	2	0
14	0	5	0	2	13	2	0
배열 CHAR를 주사 문자의 좌우 경계 위치 결정							
15	0	0	0	0	0	0	0

표1. (그림2)의 영상에서 문자가 분리 추출되는 과정

추출 순서	CY_ST	CY_ED	CX_ST	CX_ED	추출 문자
1	2	19	2	8	"s"
2	2	13	10	16	"t"
3	2	13	18	24	"u"
4	2	13	26	32	"d"
5	2	13	34	40	"y"

표2. (그림2)의 영상에서 문자를 분리 추출한 결과

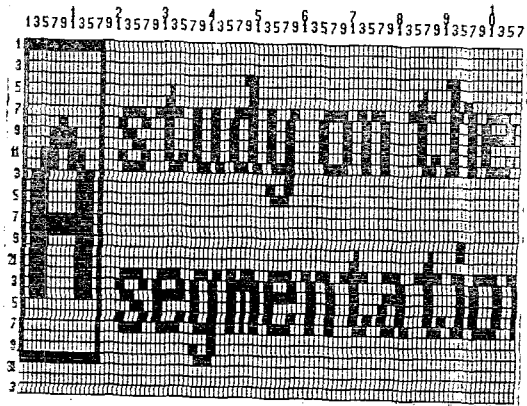

그림3. 그림과 문자열이 혼합된 문서 영상의 예

추사열	L_CHK	C_NUM	G_NUM	CY_ST	CY_ED	GY_ST	GY_ED
1	1	1	1	1	0	1	0
2	1	1	1	1	0	1	0
3	1	1	1	1	0	1	0
4	1	3	3	1	0	1	0
5	1	5	5	1	0	1	0
6	1	6	6	1	0	1	0
7	1	14	14	1	0	1	0
8	1	13	13	1	0	1	0
9	1	13	13	1	0	1	0
10	1	13	13	1	0	1	0
11	1	13	13	1	0	1	0
12	1	11	11	1	0	1	0
13	1	11	2	1	0	1	0
14	1	11	2	1	0	1	0
15	1	11	2	1	0	1	0
16	1	11	1	1	0	1	0
17	1	11	1	1	0	1	0
18	1	11	1	1	0	1	0
19	1	11	2	1	18	1	0
배열 CHAR을 주사 문자의 좌우 경계 위치 결정 배열 GRAP의 그림에 해당하는 부분을 1로 설정							
19	1	2	2	19	0	1	0
20	1	4	4	19	0	1	0
21	1	4	4	19	0	1	0
22	1	16	16	19	0	1	0
23	1	13	13	19	0	1	0
24	1	13	13	19	0	1	0
25	1	13	13	19	0	1	0
26	1	13	13	19	0	1	0
27	1	13	13	19	0	1	0
28	1	13	2	19	0	1	0
29	1	13	2	19	0	1	0
30	1	13	2	19	0	1	0
31	0	13	0	19	30	1	30
배열 CHAR를 주사 문자의 좌우 경계 위치 결정 배열 GRAP를 주사 그림의 좌우 경계 위치 결정							
32	0	0	0	0	0	0	0
33	0	0	0	0	0	0	0

표3. (그림3)의 영상에서 문자가 분리 추출되는 과정

추출순서	CY_ST	CY_ED	CX_ST	CX_ED	추출문자
1	1	18	22	28	"s"
2	1	18	30	36	"t"
3	1	18	38	44	"u"
4	1	18	46	52	"d"
5	1	18	54	60	"y"
6	1	18	65	71	"o"
7	1	18	73	79	"n"
8	1	18	84	90	"t"
9	1	18	92	98	"h"
10	1	18	100	106	"e"
11	19	30	22	28	"s"
12	19	30	30	36	"e"
13	19	30	38	44	"g"
14	19	30	46	52	"m"
15	19	30	54	60	"e"
16	19	30	62	68	"n"
17	19	30	70	76	"t"
18	19	30	78	84	"a"
19	19	30	86	92	"t"
20	19	30	94	96	"i"
21	19	30	98	104	"o"
22	19	30	106	107	"n"
23	1	30	1	18	"그림"

표4. (그림3)의 영상에서 문자를 분리 추출한 결과

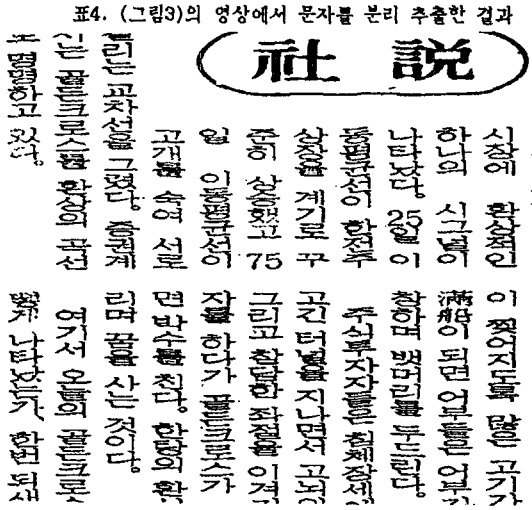

그림4. 분리 추출 실험에 사용된 문서 영상의 예(1)

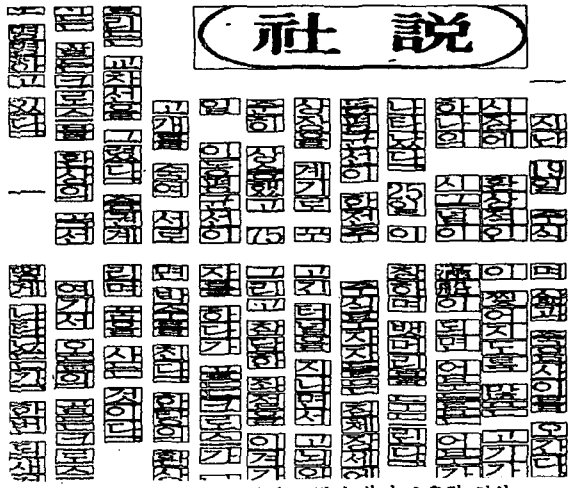

그림5. (그림4)에서 문자와 그림이 분리 추출된 영상

3. 실험 및 고찰

본 논문에서 제안한 알고리즘의 유효성을 검토하기 위하여 A4 크기의 문서를 신문 또는 잡지에서 선별하여 SCANNER를 통하여 300 DPI의 해상도로 입력하면서 문자와 그림의 분리 추출 실험을 행하였다. 인식 실험에 사용된 시스템은 16 bit 급 IBM-PC AT와 호환기종인 대우 PRO-3000을 사용하였고, 실험에 사용된 문서 영상과 문자와 그림이 분리 추출된 영상의 예를 (그림 4,6)와 (그림 5,7)에 각각 나타내었다.

본 실험에서 그림과 개별 문자를 분리 추출하는데 소요된 처리 시간은 A4 크기의 문서 영상 전체를 300 DPI의 해상도로 입력하면서 가로 2400개의 화소, 세로 3300열의 주사선으로 구성되는 문서 영상에서 25분의 매우 짧은 시간에 문자와 그림을 분리 추출할 수 있었다. 또한 그림과 문자의 분리 추출에 있어서 입력된 원래의 영상 자체가 분리할 수 없는 상태에 있던지 문자들이 서로 붙어 있지 않는한 모든 그림과 문자가 정보 흐름의 순서대로 정확히 분리 추출되었으며, 서로 접촉되어 있는 문자에 대하여도 서로 접촉된 대로 분리 추출되기 때문에 인식과정에서 쉽게 분리하여 인식할 수 있는 정도의 양호한 상태로 분리되었다. 또한 기울어져 입력된 문서 영상에 대하여도 문자열과 문자열 사이의 공백 이상으로 기울어져 입력되지 않는한 문자와 그림을 분리 추출하는데 아무런 문제가 발생되지 않았다.

그림6. 분리 추출 실험에 사용된 문서 영상의 예(2)

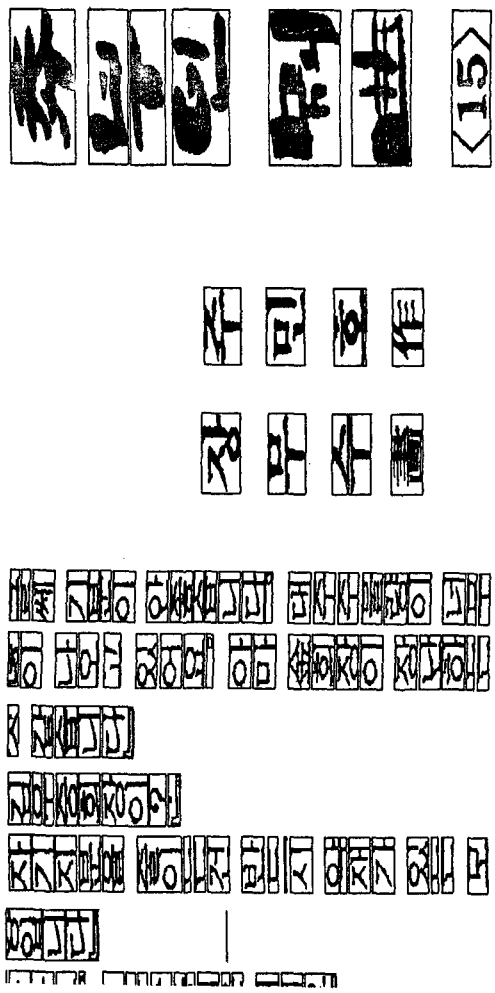

그림7. (그림6)에서 문자와 그림이 분리 추출된 영상

4. 결 론

본 연구에서는 한글, 영문, 한자, 기호 등의 문자와 그림이 혼합된 문서를 인식할 수 있도록 실시간으로 입력되는 문서 영상에서 개별 문자와 그림을 분리 추출하는 알고리즘을 제안하였다.

분리 추출에 사용된 알고리즘은 실시간으로 입력되는 문서 영상으로부터 문자와 그림의 경계 위치를 알아내어 개별 문자와 그림 영역의 영상을 분리 추출하기 때문에 처리할 문서의 영상 데이타가 아무리 많아도 컴퓨터의 기억 용량과는 무관하게 처리된다. 또한 단 한번의 주사로 문자와 그림이 정보 흐름의 순서 대로 모두 분리 추출되어 최소의 시간에 처리가 가능하다. 따라서 본 알고리즘은 입력되는 문서 영상의 데이타 량과무관하게 최소의 기억 용량과 최소의 시간에 문자와 그림이 분리 추출되는 특성이 있다. 또한 문자와 그림의 분리 추출에 있어서도 기존의 어느 방법 이상으로 문자와 그림이 정확히 분리 추출되며, 특히 그림과 문자가 동시에 분리 추출되는 특성을 지니고 있다.

그러나 본 알고리즘은 보통의 소설책이나 보고서 등에서와 같이 가로체 혹은 세로체로 일관되는 문서에만 적용이 가능하며, 신문에서와 같이 가로체와 세로체가 불규칙하게 배열되어 있는 문서 영상에는 적용할 수 없는 단점이 있다. 따라서 신문 구조의 특성을 연구하여 좀더 개선할 필요가 있으며, 약간의 개선만 이루어 진다면, 신문과 같은 복잡한 문서 영상에 대하여도 최소의 기억 용량과 최소의 시간에 정확한 문자와 그림의 분리 추출이 가능할 것으로 기대된다.

본 연구의 결과는 소설책이나 보고서에서와 같이 문자와 그림이 혼합되어 가로체 또는 세로체로 작성되어 있는 문서를 인식하는데 효과적으로 적용될 수 있으리라 생각된다.

참 고 문 헌

[1] Ito Sakatani and Takei, "Parallel segmentation algorithm of document," ICCS, pp.519-523, 1979.

[2] Scherl, Whal and Fuchsberger: "Automatic separation of text, graphic and picture segments in printed material," Pattern recognition in practice, North-holland, pp.213-221, 1980.

[3] Whal, Wong and Casey, "Block Segmentation and Text Extraction in Mixed Text," Computer Graphics and Image Processing, vol.20, pp.375-390, 1982.

[4] Teruo Nakayama, and Isao masuda, "A segmentation method for Document Image without the Knowledge of ocument Formats," 일본전자통신학회 논문지, vol.J66-D, no.1, 1983.

[5] Osamu Nakamura, Norivosh Okamoto, Tsuyoshi Niwata and Toshi Minami, "A symbol Segmentation Algorithm from Free Format Document," 일본전자통신학회논문지, vol. j66-D, no.4, 1983.

[6] 이성환, 조창제, 김진형, "실용적 한글 문서 자동 인식 시스템 개발의 문제점 및 개선 방향," 한국 정보과학회 춘계 학술 발표회 논문집, pp.127-130, 1988년 4월.

[7] 김형훈, 이성환, 김진형, "한국 신문 영상의 구조 분석을 통한 기사의 추출," 한국정보과학회지 제 6권 5호, pp. 392-404, 1988년 10월.

[8] 남궁 재찬, 류황빈, 남궁 윤, "한국어 문서로부터 문자분리 및 도형 추출에 관한 연구," 대한전자공학회 논문지, 제 25권 9호, pp.1091-1100, 1988년 10월