신문 자동인식 시스템을 위한 문자의 분류에 관한 연구

°이 승형, 전 종익, 조 용주, 남궁 재찬
광운대학교 전자계산기공학과

A Study on the Classify of Character for Newspaper Automatic Recognition System

S. H. Lee, J. I. Cheon, Y. J. Cho, J. C. Namkung
Dept. of Computer Eng. Kwang Woon Univ.

요 약

본 논문에서는 신문자동인식을 위한 신문문자의 분류에 관한 연구를 하였다. 먼저, 문서의 문자를 추출하기 위하여 블럭화를 행한다. 블럭화는 문자열을 찾아 절과절, 단어와 단어 사이를 찾아 분리구간을 정한다음 블럭을 합성 및 분리를 하였다. 다음으로 블럭화된 문자의 종류를 알기위한 각 문자에 대하여 6 형식 분류를 하여 특성을 조사함으로써 문자분류를 행하였다. 본 연구에서는 실험을 통하여 블럭화는 충실하게 추출이 되어졌고 한글의 모아쓰기 특성과 한문과의 유사한 형식특성 때문에 분류에 어려움이 있었으나 비교적 충실하게 추출하였다.

I. 서 론

현대사회의 급속한 정보량의 요구와, 어디에나 산재해 있는 사회생활의 유용한 정보는 컴퓨터의 발전으로 인하여 실용성, 경제성, 신속성이 날로 증가하고 있다. 컴퓨터를 이용한 정보의 전달은 다른 어느 매체보다도 매우 우수하다. 하지만 이미 기록 되었거나 내장된 정보가 아닌 정보는 적절한 정보의 형태로 변환이 되어져야 한다. 사무 자동화나 여러 응용 분야에 있어서 이러한 요구는 갈수록 높아지고 있다.

또한, 정보량의 증가는 여러가지의 통신 수단으로 전송이나 전달이 되고 있고 이러한 점을 자동화하기 위해서 문서영상의 인식에 많은 연구가 진행 되어지고 있다.[3] 문서(Document)는 다양한 종류의 문서가 존재하는데 대표적으로는 각종신문, 잡지, 사무용 문서, 전표등 다양한 종류의 문서가 존재한다. 본 논문에서는 이러한 문서중에서 신문영상에 대하여 한글과 한자를 구별하고 한글에 대해서는 한글이 가지는 6 가지형식을 추출하여 인식에 용이한 방법을 제안하였다.

이러한 연구는 방대한 글자의 자수를 가지는 한글의 인식을 자음과 모음으로 각각의 문자를 끊어내어 자음과 모음을 인식하여 각 자소에 대한 조합법칙을 적용하여 인식을 하므로서 현저한 메모리나 시간의 감소를 가져올 수 있게 하였다.

문자인식에 있어서 인식의 범위가 문서단위로 넘어감에 따라 문서인식에서 가장 큰 문제는 방대한 기억용량과 처리시간이 요구된다는 점이다. 기존의 문자인식의 방법은 결정론적 방법(dicision method)과 구문론적 방법(syntactic method)로 대별이 되고 있는데 메모리나 시간적인 측면에서 모두 장단점을 가지고 있다. 하지만 처리의 단위가 문서단위라면 방대한 처리량을 고려하면 속도의 개선측면에서는 새로운 방법의 출현이 매우 필요한 실정이다. 본 논문에서 제안한 한글의 형식 분류에 의한

자소의 구조별 분리는 초성과, 중성, 종성에 있는 51개의 자소만을 인식하므로 빠른 인식의 결과를 얻을 수 있다.

그동안의 문서영상의 연구상황을 보면 김진형[1]등은 한글 신문에 대하여 신문의 구조적 분석과 문자의 부분을 블럭화하여 신문기사의 추출에 관한 연구를 하였고 남궁연[2]등은 그림과 문자가 혼합되어 있는 문서영상에서 그림과 기사를 추출하는 연구를 하였다. 또한 오인권[7]등은 복합된 문서에서 문자를 블럭화 하는 down-up알고리즘을 제안하여 영문이나 특수문자를 구별하는 연구를 하였다. 그리고 한글 자체에 대한 구조적 분석을 위한 연구를 보면 이주근[4]은 한글을 6가지의 형식으로 분류하였고 남궁재찬[5][6]등은 한글의 자소를 분리 추출하는 Indexed-window 알고리즘을 제안하여 인식을 용이하게 하였다. 또한 남궁재찬[8]은 한글이 가지는 특성의 분석에 관한 연구도 하였다.

본 논문에서는 신문영상에서 문자의 영역을 블럭화 한다음 블럭된 문자에 대하여 문자의 종류를 구별하고 한글에 대해서는 고유의 특성을 적용하여 자소를 분리하기 위한 형식 분류에 관한 연구를 하였다.

II. 신문 문자에 대한 고찰

본 장에서는 현재 통용되고 있는 신문에서의 문자의 농인 형태와 문자의 종류에 대해서 살펴 본다.

현재 통용이 되고 있는 중앙 일간지를 대상으로 살펴보면 문자의 쓰이는 형태는 가로쓰기와 세로쓰기 형태로 되어있고 부분적으로는 가로쓰기와 세로쓰기가 혼용되어 쓰이는 신문도 있다. 이러한 정보는 문자의 블럭화시 문자의 농인 위치가 고려되어 블럭화 되어야 한다.

또한 신문내에 쓰이는 문자체(typeface)는 명조체의 형식을 가지고 있으나 가로쓰기와 세로쓰기의 형태에 따라 문자의 크기가 달라진다. 통상 많이 쓰이는 세로쓰기를 보면 일반적인 교과서체 보다는 문자의 높이(hight)가 많이 축소된점을 알 수 있다. 이렇게 함으로써 많은 문자를 실을 수 있게 한점이 신문 문자의 특징이다.

그리고 신문에서 쓰이는 문자는 한글만이 쓰이는 것이 아니라 다양한 종류의 문자 즉, 한자, 영자, 숫자, 특수문자등이 쓰이는 점을 알 수 있다. 본 논문에서는 이러한 점을 고려하기 위하여 각종 신문의 사설을 대상으로 쓰이는 문자의 종류가 어느정도의 비율을 가지는 가를 조사하여 표 1에 그 결과를 보였다.

표 1. 각 신문의 문자별 조사

	서식형태	총문자수	한 글	한 자	영 어	숫 자	특수기호
국민일보	가 로	1331	1167	22	0	3	39
		(100)	(94.80)	(1.79)	(0)	(0.24)	(3.17)
" "	"	1135	1046	35	0	34	30
		(100)	(92.15)	(3.08)	(0)	(2.11)	(2.64)
한겨레 신문	가 로	854	818	0	0	2	34
		(100)	(95.70)	(0)	(0)	(0.23)	(3.98)
" "	"	784	740	0	0	13	31
		(100)	(94.39)	(0)	(0)	(1.66)	(3.95)
한국경제신문	세 로	917	796	95	0	4	22
		(100)	(86.80)	(10.36)	(0)	(0.44)	(2.40)
" "	"	921	789	55	14	28	37
		(100)	(85.67)	(5.97)	(1.50)	(3.04)	(4.02)
동아일보	세 로	958	889	30	2	3	29
		(100)	(93.28)	(3.15)	(0.21)	(0.31)	(3.04)
" "	"	1209	1131	35	8	4	31
		(100)	(93.55)	(2.89)	(0.66)	(0.33)	(2.56)
중앙일보	세 로	1262	1092	136	0	3	31
		(100)	(86.54)	(10.78)	(0)	(0.24)	(2.26)
" "	"	1015	911	72	0	4	29
		(100)	(89.66)	(7.09)	(0)	(0.39)	(2.86)
조선일보	세 로	1098	1051	12	0	6	29
		(100)	(95.71)	(1.09)	(0)	(0.55)	(2.64)
" "	"	1121	1073	12	0	8	38
		(100)	(95.72)	(1.07)	(0)	(0.71)	(3.39)

표 1 에서 나타난 바와 같이 한글의 사용이 90 퍼센트 이상이 되며, 한문과 영자의 비율이 비교적 낮음을 알 수 있다. 특히, 가로쓰기 형태의 신문은 세로쓰기 형태의 신문에 비하여 한자의 비율이 낮으며 경제신문과 전자시보와 같이 특수 신문의 경우 한자와 영문자의 비율이 높아짐을 알 수 있다. 문자의 형태에 따른 특징은 다음과 같다.

(1) 한 글 : 90 퍼센트 이상의 문자가 여기에 속한다.
한글의 6 가지 형식으로 세분될 수 있다.

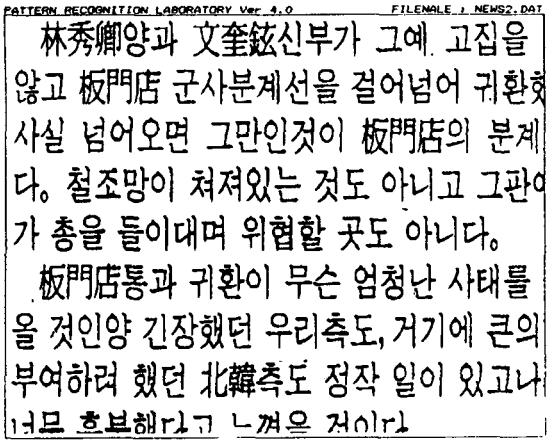
(2) 한 자 : 2자에서 4자까지 연속되는 형태가 대부분이며 획(stroke)이 복잡하고 글자가 군을 이룬다.

(3) 영 자 : 주로 대문자를 사용하며 단어를 이룬다.
kg과 km같은 것은 한개의 문자로 사용하였다.

(4) 숫 자 : 세로쓰기 형식에서는 2-3글자를 한 문자로 사용하고 있다. ex) 3, 15, 125

(5) 특수문자 : 신문사에 따라 제한된 형태를 가진다.

그림 1 에는 본 논문에서 처리한 300 DPI의 해상도를 가진 imege scanner에 입력받은 데이타의 예를 보여주고 있다.



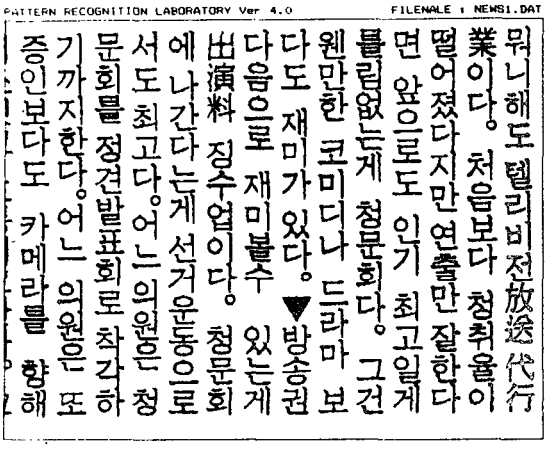


그림 1. 입력 데이타

III. 문자의 블럭화

신문에 쓰이는 문자는 한글자가 추출되어지기 위하여 본 장에서는 각각의 문자를 추출하기 위한 문자의 블럭화를 행한다. 한글 문자의 블럭화에 관한 연구는 지금까지 몇가지의 방법[1][7][2]이 제안 되었으나 본 연구에서는 속도를 향상시킨 새로운 방법을 제안 하였다. 본 연구에서는 행단위로 문자를 추출하고 각 행을 대상으로 블럭화하여 이를 합성 및 분리를 하였다.

1. 잡음제거

입력된 문서는 디지탈화에 따른 문서상의 잡음과 하드웨어상의 잡음을 제거하기 위하여 3X3 마스크를 사용하여 고립점을 제거하였으며 1X3 마스크를 사용하여 문자간에 붙음을 최소한 방지하였다. 그림 2 에 본연구에 쓰인 마스크를 보였다.

0	0	0
0	1	0
0	0	0

0	1	0

그림 2. 3X3과 1X3 마스크

2. Write type 조사

본 절에서는 두가지 형태로 쓰여지는 신문영상의 쓰여지는 형식을 조사하여 쓰여지는 형태에 따라 블럭화를 하였는데 위에서 아래로 그리고 좌에서 우로 스캔하면서 가로쓰기와 세로쓰기의 형태를 구분하였다.

3. Write type에 의한 블럭화

Write type을 조사한후 write type에 따라서 각 문자열 별로 시작점과 마지막점을 찾아서 각 문자열에 대하여 문자의 블럭을 구성하였다. 이와같이 처리한 결과를 그림 3 에 보였다.

그림 3. 문자의 블럭화

4. 빈도수에 따른 표준 폰트(font) 크기 결정

신문에 있는 문자는 신문 인쇄종이의 불균일과 선명도 때문에 많은 문자들이 붙어서 추출되어진다. 본 연구에서는 이러한 점을 개선하기 위하여 신문에 쓰이는 문자의 크기를 조사하여 표준적인 문자의 크기를 결정였다.

5. 문자크기에 의한 블럭의 합성 및 분리

결정된 문자의 크기는 제일 많이 사용되는 한글의 문자의 크기와 비슷하므로 이 단계에서 문자의 블럭들중 붙어있는 문자나 자소가 분리된 경우에 한하여 문자의 분리 및 합성을 한다. 다음에 블럭및 합성에 관한 알고리즘과 그림 4 에 분리, 합성된 데이타의 예를 보였다

단계 1. 문자열을 찾는다.

단계 2. 문자열에서 절과 절, 단어와 단어의 사이를 찾아 분리구간을 결정한다.

단계 3. 분리구간 내의 블럭들과 표준으로 삼은 문자 크기를 비교한다.

단계 4. 해당블럭에 조합법칙을 이용하여 분리 및 합성

단계 5. 다음 블럭을 찾아 단계 3에서 단계 5 반복

단계 6. 다음 문자열을 찾아 단계 2에서 단계 5 반복

단계 7. 마지막 문자열에 이를때 까지 단계 1에서 단계 6을 계속해서 반복

그림 4. 분리 및 합성된 데이타

IV. 한글의 형식 분류

한글은 한글 고유의 구조적 특성이 자소들을 각각 모아놓은 조합형 문자라는 것을 감안하여 자소들이 놓이는 위치에 따라 6 가지의 형식[4]으로 분류할 수 있다.

한 자소는 각 형식에 따라 서로 크기, 형태, 위치가 미세하게 달라져 6 가지의 형식[4]을 갖는다. 그러므로 한글 인식을 하기 위한 전처리 단계로서 한글의 형식을 분류하게 되면 각 자소의 인식을 행할때 사전에 있는 것과의 비교는 한 형식에 대하여 행하기 때문에 그만큼 빠른 인식을 행할 수 있는 것이다.

1. 한글의 자소 특성

본절에서는 한글이 가지는 6 가지 특성중 자음과 모음이 가지는 특성에 관하여 논한다. 대상으로 삼은 데이타는 모두 한글로 간주하고 실행을 하였다.

(1) 한글의 자소중 긴횡모음보다 수평 방향으로 큰것은 없다.

(2) 한글의 종성은 수평방향으로 볼때 초성이나 횡모음 보다 앞에 나오지 않는다.

(3) 한글의 자음은 거의 수평, 직선 성분이 있다. 그러나 한글의 "ㅇ"은 직선 성분은 없고 화소가 이어진 폐곡선 정보를 이용하였다.

2. 한글의 6 형식

한글의 6 형식을 그림 5에 나타내었다.

한글은 모음을 중심으로 모음 위치에 따라 자음들이 놓인다는 특성이 있어 본 논문에서는 모음을 중심으로 형식을 분류하였다. 한글의 모음은 크게 횡모음과

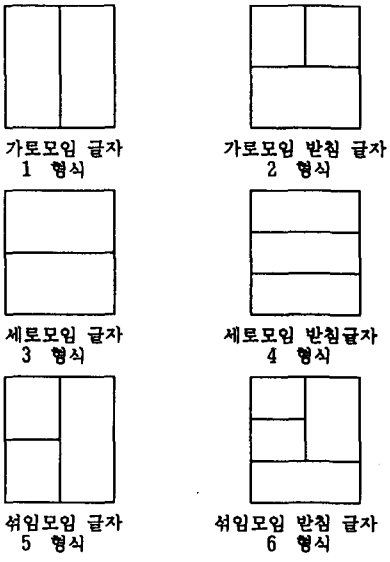


그림 5. 한글의 6 형식

종모음으로 나눌 수 있으며 이를 세분화 하면 긴횡모음, 짧은횡모음, 긴종모음, 짧은종모음으로 나뉘어 진다.

한글의 자소들 중 긴 횡모음 보다 횡방향으로 큰 직선성분은 없으므로 한 문자의 수평크기에 가까운 연속되는 수평성분이 있으면 3 형식이나 4형식이다. 그리고 긴횡모음 밑에 자음이 있으면 4 형식이고 긴횡모음을 갖지 않은 것은 종모음을 가진 1,2,5,6 형식이 된다.

종모음중 긴종모음을 가진 것은 1형식과 5형식이며 짧은 종모음을 가진것은 2형식과 6 형식이 된다. 긴종모음을 가지며 짧은 횡모음을 가진것은 5 형식이며 짧은 횡모음을 가지지 않은것은 1 형식이 된다. 짧은 종모음을 가진것은 항상 종성이 있기 때문에 단횡모음이 있는 것은 6 형식이고 그렇지 않으면 2 형식이 된다.

또한, 1,2 형식은 모음에서 나타나는 걸출기는 어느 위치 밑으로는 오지 않기 때문에 특정지역 하변에 연속된 수평성분이 있거나 폐곡선 성분이 있으면("0"인 경우) 종성으로 판별하여 2 형식으로 분류 하였다.

3,4 형식에서는 긴횡모음이 있는 위치 이하의 특정지역에서 연속된 수평성분이 임의의 임계치 이상이거나 폐곡선 성분이 있으면 종성으로 판별하여 4 형식으로 분류하였다.

5,6 형식에서는 긴종모음과 짧은횡모음을 포함하고 있으면 3,4 형식에서와 같이 종성을 찾아낸다.

이와같은 한글의 특성을 적용하여 한글의 형식을 분류한 예를 그림 6 에 나타내었다.

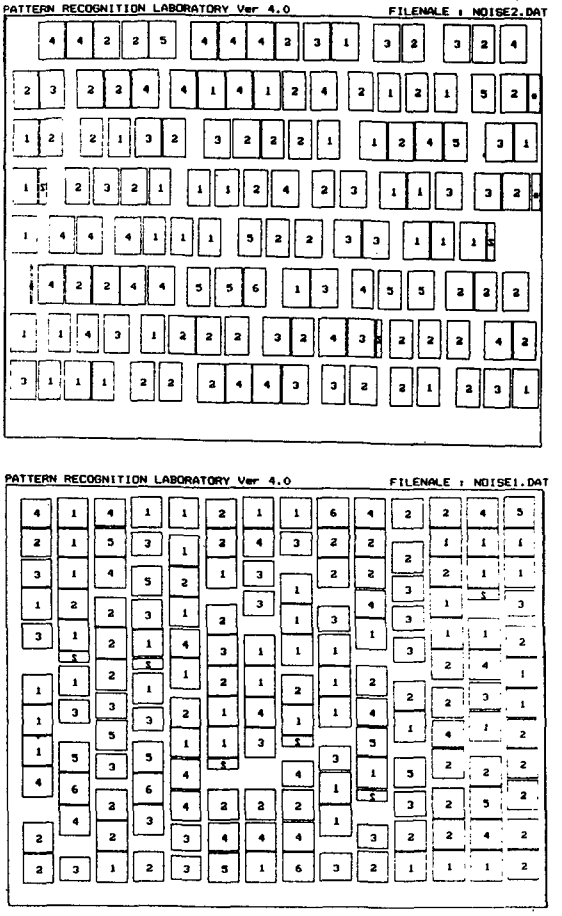


그림 6. 한글의 형식분류

V. 문자 분류 방법의 제안

본 장에서는 블럭화된 문자의 종류가 여러가지로 되고 특히 한글과 한자는 분류에 매우 어려운점이 있으므로 한글과 한자의 특성을 고려하여 문자를 구별하는 방법을 제안한다.

1. 확률분포에 의한 분류

(1) 각 블럭당 흑화소의 수

흑화소의 수가 단위 블럭에 대하여 한자, 한글, 영자,

숫자, 특수기호의 순으로 많다.

(2) 가로의 불연속 Blank line 수

한글은 2개 이하인 반면 3개 이상이면 모두 한자이다.

예) 는, 言

(3) 세로의 불연속 Blank line 수

한글은 1개이하인 반면 2개 이상이면 모두 한자이다.

예) 가, 川

(4) 첫흑화소의 위치

한자는 중심위주의 문자인 반면 한글은 1,2,5,6 형식일때 중심 오른쪽에 위치 한다.

예) 가, 각, 木

2. 구성성분(primitive)를 이용한 분류

(1) 사선(/,) 성분의 수

한자, 한글, 영문, 특수문자, 숫자의 순으로 많다.

(2) "ㅅ" 성분

중앙의 첫점에서 "ㅅ" 성분이 추출되고 한글 3,4 형식이면 한글로 판정 예) 수, 숙,

(3) "ㅇ" 성분

"ㅇ" 성분이 존재하면 한글로 판정

3. 인식과정에 따른 분류

비교적 빈도수가 적은 숫자, 특수기호, 영자 순으로 인식하며 한글과 한자의 분류는 한글 6 형식으로 분류한 정보와 한글 초성 위치에서의 자음 성분을 조사하여 판정한다.

4. 글자 크기에 의한 분류

각 블럭에 대하여 실제문자의 크기의 문자의 크기를 조사하여 판정하며 그 크기는 한문, 한글, 영문, 숫자, 특수기호의 순으로 크다

이와같은 특성을 조사하여 한글과 한자 그리고 특수기호를 분류한 예를 그림 그림 7 에 보였다.

VI. 실험 및 고찰

1. 실 험

본 연구의 실험에는 신문영상을 300 DPI의 해상도를 가진

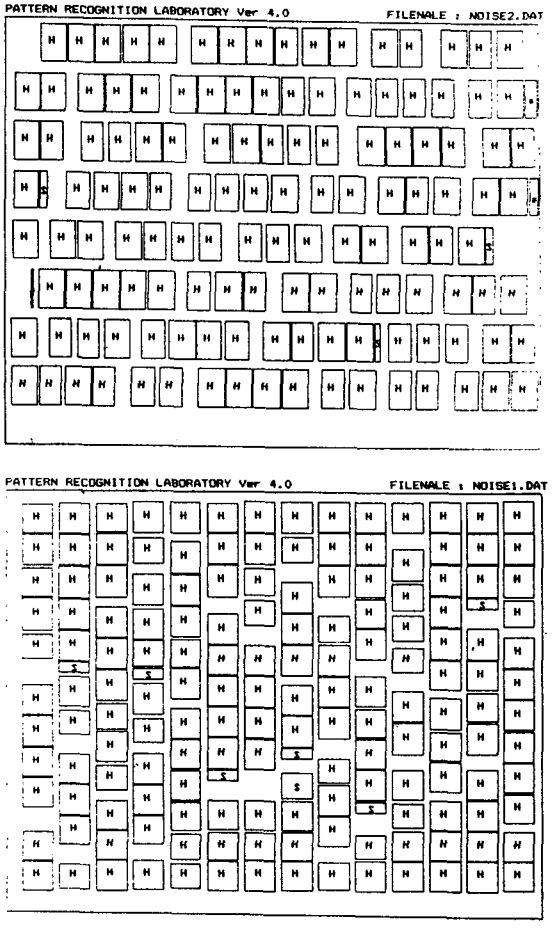


그림 7. 최종 분류된 결과

Image Scanner에 입력을 받아 IBM PC/386 으로 전송하여 처리 하였다. 데이타의 크기는 640X400의 32 KByte 이고 처리에 사용한 언어는 Turbo-C 를 사용하였다 그림 8 에 처리 시스템을 보였다.

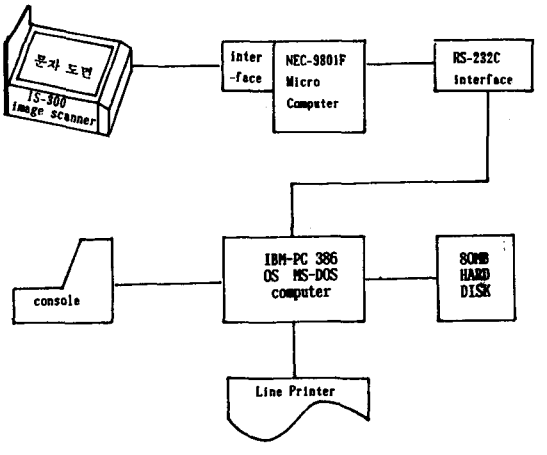


그림 8. 처리 시스템

2. 고 찰

본 연구는 신문영상의 자동인식을 위한 전처리 연구로서 신문에서 표 또는 그림등 다양한 텍스쳐(texture)는 고려하지 않고 문장 만을 입력받아 문자 자체가 가지는 정보를 이용하여 문자의 종류를 구별하고 한글에 대해서는 한글의 형식을 분류하여 인식의 전처리 연구를 하였다.

입력문서에서 문자의 블럭화는 매우 충실하게 추출되었으며 한 데이타의 블럭화에 걸리는 시간은 5초 정도 소요되었다.

또한 한글에서의 형식분류는 한글의 자체가 가지는 자소의 미묘한 변동과 수많은 특성에 의하여 간단한 구조의 문자는 비교적 잘 분리 되었으나 복잡한 형식의 문자의 아직 많은 문제점을 보이고 있다.

VII. 결 론

본 논문에서는 신문영상에서 문자의 추출과 추출된 형식과 구별을 하는 연구를 하였다.

블럭화에 대해서는 기존의 방법과는 새로운 방법을 제안하여 좋은 결과를 얻었으며 문자의 구별과, 한글의 형식분류에 관해서는 일단 한글의 특성을 이용하여 인식을 위한 전처리 연구를 하여 Multi-font, 여러문자를 포함하는 문서에서 인식을 위한 좋은 가능성을 발견하였다. 앞으로 문서인식의 분야에서 많은 연구가 있어야 하겠다.

참고 문헌

[1] 김형훈, 이성환, 김진형, "한국 신문 영상의 구조 분석을 통한 기사의 추출", 정보과학회지, 제 6 권 제 5 호, 1988.

[2] 남궁재찬, 유황빈, 남궁연, "한국어 문서로 부터 문자 분리 및 도형 추출에 관한 연구", 대한 전자공학회 논문지, Vol. 25, No. 9, 1988.

[3] K. INAGAKI, T. KATO, T. HIROSHIMA and T. SAKAI, "MACSYM : A Hierarchical Parallel Image Processing System for Event-driven Pattern Understanding of Documents", Pattern Recognition, Vol. 17, No. 1, pp. 85-108, 1984.

[4] J. K. Lee, "Korean Character Display Variable Combination and its Recognition by Decomposition Method", Ph. D. dissertation in Keio Univ., Japan, 1972

[5] 남궁재찬, "Index-Window 알고리즘에 의한 한글 pattern의 부분분리와 인식에 관한 연구", 인하대학교 박사학위 논문, 1982.

[6] 이주근, 남궁재찬, 김영건, "한글 pattern에서 subpattern 분리와 인식에 관한 연구", 대한 전자 공학 회지, Vol. 18, No. 3, 1983.

[7] 오인권, 남궁재찬, "영문이 혼합된 한글 문서에서의 문자 및 특수문자 추출에 관한 연구", 광운대학교 대학원 석사학위 논문, 1988.

[8] 남궁재찬, "Font 개발을 위한 한글특성 분석에 관한 연구", 광운대학교 논문집, Vol. 18, 1989.