낱자 인식기와 자소 조합 인식기를 혼용한 인쇄체 한글 인식방법

장승익[O] 임길택 김호연 정선화 남윤석
한국전자통신연구원 우정기술연구센터
{sijang[O], ktlim, hoyon, sh-jeong, ysnam}@etri.re.kr

A Method of Machine-Printed Hangul Recognition using Character and Combined-Grapheme Recognizers

Seunglck Jang[O], Kil-Taek Lim, Ho-Yon Kim, Seon-Hwa Jeong, Yun-Seok Nam
Postal Technology Research Center, ETRI

요 약

본 논문에서는 낱자 인식기와 자소 조합 인식기를 혼용한 저품질 인쇄체 한글의 고성능 인식 방법을 제안하였다. 제안한 방법에서는 입력 문자를 한글 6형식과 기타 형식의 문자, 총 7종으로 분류한 뒤, 입력 문자를 인식 대상 문자의 수와 자소 복잡도에 따라 하나 또는 두 개의 인식 단위(HRU: Hangul recognition unit)로 분리하여 인식한다. 각 인식 단위 영상에서 추출한 방향각 특징을 다층신경망 인식기를 이용하여 인식한다. 다음으로, 각 다층신경망 인식기의 신뢰도를 조합하여 최종 인식 결과를 도출한다. 제안한 방법을 사용한 실험에서 98.80%의 인식률을 얻을 수 있었으며, 이는 기존 방법에 비해 23.61%의 오류가 감소한 것이다.

1. 서론

문서 자동화 처리분야에서 문자인식 기술은 필수적인 요소이며, 문자인식 기술과 관련한 많은 연구가 있어왔다 [1-3]. 하지만, 다량의 우편물을 자동으로 구분하기 위해 기존의 인쇄체 문자인식 기술을 적용할 경우 인식 성능이 매우 저하된다. 이는 다량의 우편물을 제한된 시간 내에 처리해야 하는 시간적 제약이 발생하기 때문이며, 이로 인해 다량의 우편물 인식에 사용하는 영상은 일반적인 한글 문자인식에서 사용하는 300dpi 이상의 고해상도 영상에 비해서 낮은 200dpi 정도의 저해상도 영상으로 제한될 수 밖에 없다. 이러한 영상의 평균 면적은 300dpi로 획득한 영상의 평균 면적의 44% 수준이며, 이로 인해 인식에 필요한 정보의 소실이 필수적으로 발생한다. 또한, 시간적 제약에 의해서 시간이 많이 소요되는 복잡한 알고리즘을 적용하기가 어렵다.

본 논문에서는 낱자 인식기와 자소 조합 인식기를 이용한 저품질 인쇄체 한글의 고성능 문자인식 방법을 제안한다. 제안한 방법에서는 입력 문자영상을 한글 6형식과 기타 형식의 문자, 총 7종으로 분류한 뒤 각각의 유형별 다층신경망 인식기를 이용하여 입력 문자영상을 인식한다. 입력된 한글 문자는 형식에 따른 인식 대상 문자의 수와 자소 조합 복잡도에 따라 하나 또는 두 개의 인식 단위(HRU: Hangul recognition unit)로 분리하여 인식한다. 제안한 방법을 사용하여 우편영상에서 추출한 200dpi의 해상도를 가지는 233,932자의 한글 테스트 영상에 대해 실험한 결과 98.80%의 높은 인식률을 얻을 수 있었다

2. 문자인식 시스템

2.1 시스템 흐름

저품질 인쇄체 한글 인식 시스템의 전체 흐름도는 그림 1과 같다. 입력 문자영상에서 문자영상의 방향각 특징을 추출하고, 이를 다층신경망으로 구현된 유형 분류기를 통해 입력 문자영상의 유형을 분류하게 된다. 한글은 그림 2와 같이 자소 조합의 형태에 따라 1형식에서 6형식으로 분류된다. 한글 이외의 문자인 영어, 숫자, 기호 등은 인식

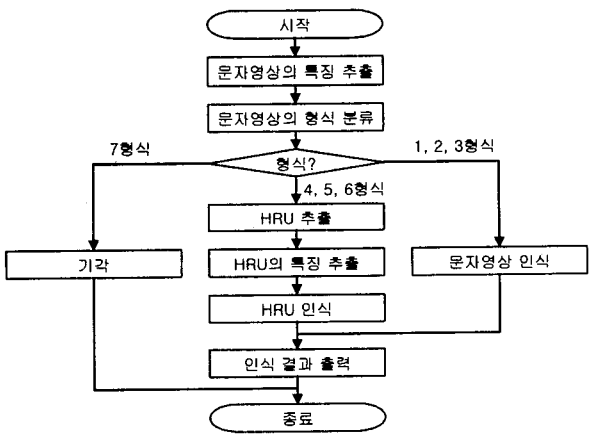
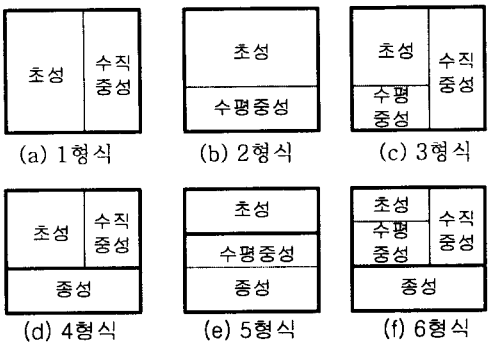


그림 1. 제안한 한글 인식 방법



(a) 1형식 (b) 2형식 (c) 3형식

(d) 4형식 (e) 5형식 (f) 6형식

그림 2. 한글 형식의 종류

대상으로 설정하지 않았으며, 7형식으로 분류하였다.

유형 분류기의 결과가 1~6형식으로 나올 경우, 문자영상을 그림 2에서 보여지는 것과 같이 분할하고, 각각의 인식 단위의 영상으로부터 방향각 특징을 추출하여 인식한다. 여기서, 한글 문자의 기본 인식 단위를 HRU(Hangul Recognition Unit)로 정의한다. 그림 2에서 굵은 실선으로 둘러진 사각형이 하나의 HRU가 되며, 점선은 HRU를 구성하는 자소의 경계를 표현하고 있다. 입력 문자영상이 1, 2, 3형식인 경우, HRU가 문자영상 전체이므로 문자영상의 분할을 시도하지 않으며, 유형 분류를 위해 추출했던 방향각 특징을 그대로 사용하여 문자인식을 수행한다. 다음으로 입력 문자영상이 4, 5, 6형식인 경우, 각각의 영상을 2개의 HRU로 분할한 뒤, 각 HRU로부터 특징을 추출하여 문자인식을 수행한다. 그리고, 본 논문은 1~6형식의 한글문자에 대한 인식 방법에 국한된 것으로 한글이 아닌 7형식 문자에 대한 인식은 논의에서 제외한다.

문자영상의 인식 결과는 HRU 인식기의 신뢰도를 바탕으로 결정하게 된다. 1, 2, 3형식 문자의 경우, 낱자 인식기의 신뢰도를 사용하여 신뢰도가 가장 큰 결과를 최종결과로 선택한다. 4, 5, 6형식 문자의 경우 자소 조합 인식기의 신뢰도의 곱이 가장 큰 결과를 최종결과로 선택한다.

2.2 입력 문자영상 분할

한글을 6형식으로 분류하고 인식하는 방법은 크게 문자 전체를 인식하는 방법과 문자를 분할하여 인식하는 방법으로 나눌 수 있다. 전자의 경우 입력 문자영상을 분할하지 않기 때문에 분할에 의한 속도 저하와 분할 오류가 발생하지 않는다. 하지만, 인식 대상 클래스의 수가 많거나 입력 문자영상이 복잡한 경우 좋지 못한 인식 결과를 얻을 수 있다. 후자의 경우 적은 수의 문자 모델로 많은 수의 클래스로 구성된 입력 문자영상을 인식할 수 있지만, 분할에서 오는 속도 저하와 분할 오류의 위험이 있다.

본 논문에서 제안하는 방법은 앞에서 언급한 두 가지 방법의 장점만을 취하는 방법이다. 먼저, 1, 2, 3형식과 같이

그림 3. 5형식 문자 영상

간단한 형태의 문자는 분할을 시도하지 않고, 입력 문자영상을 그대로 인식함으로써 분할 오류를 하였다. 다음으로, 4, 5, 6형식과 같이 복잡한 형태의 문자는 문자영상에서 초성 또는 중성을 추출하여 1, 2, 3형식의 문자영상과 비슷한 복잡도를 가질 정도로만 분할함으로써 분할 오류를 최소화하여 인식한다. 4, 5, 6형식의 분할 방법에 대해 자세히 설명하면 다음과 같다.

본 논문에서 제안한 방법은 통계치 기반 분할 방법과 휴리스틱 기반 분리방법을 사용한다. 먼저, 4형식과 6형식의 분할 알고리즘은 통계치 기반이며, 아래의 식 (1)의 $Score(y)$의 값이 최고인 지점을 분할위치로 선택한 뒤 HRU를 추출한다. 식 (1)에서, y는 입력 문자영상의 y축 좌표이며, $Histo()$는 문자영상의 y축 기준 히스토그램 값이며, $pdf()$는 훈련 데이터로부터 추출한 분할위치의 확률밀도함수이며, w_1과 w_2는 각각 $Histo(y)$와 $pdf(y)$의 가중치이다. 본 논문에서는 w_1과 w_2는 각각 2.0과 1.0으로 설정하였다.

$$Score(y)=\left(1.0-\frac{Histo(y)}{\max Histo(\,)}\right)^{w_1} pdf(y)\cdot w_2 \qquad \text{식 (1)}$$

5형식 문자영상은 휴리스틱 기반으로 분할위치를 선택하고, HRU를 추출한다. 5형식 문자영상에서도 종성을 분리하면 2형식 문자와 비슷한 복잡도를 가지는 HRU를 얻을 수 있다. 하지만, 그림 3의 예와 같이 5형식 문자에서는 접촉이 존재하여, 종성을 분리하기 쉽지 않다. 따라서, 제안하는 방법은 그림 2의 (e)와 같이 초성을 하나의 HRU로, 중성과 종성을 다른 하나의 HRU로 분리하여 인식한다. 분리 방법은 가장 긴 수평획을 찾고, 수평획 상단의 짧은 수직획의 유무에 따라서 분할위치를 선정한다.

2.3 유형별 문자 인식기

유형별 문자 인식기는 입력의 유형에 따라서 낱자 및 자소 조합 인식기로 구성되며, 식 (2)와 같이 표현된다. 식 (2)에서 HRU^{Tn}은 유형 T의 n번째 HRU를 인식하는 인식기며, N_T는 유형 T의 HRU의 개수이다. 1~6형식 인식기는 CR^1~CR^6으로 표현하고, 유형 분류기는 CR^0로 표현한다.

$$CR^T=\left\{HRU^{T1},HRU^{T2},\cdots,HRU^{Tn}\right\},\ 1\le n\le N_T \qquad \text{식 (2)}$$

각 문자 인식기의 구성은 표 1과 같으며, 모두 다층신경망 기반의 인식기이다. 최종 결과는 CR^0~CR^3의 경우, 낱자 인식기의 출력층에서 신뢰도가 가장 높은 노드가 되며,

표 1. 문자 인식기

형식	인식기	출력층 수
0	$CR^0 = \{HRU^{01}\}$	HRU^{01}= 7
1	$CR^1 = \{HRU^{11}\}$	HRU^{11}= 171
2	$CR^2 = \{HRU^{21}\}$	HRU^{21}= 95
3	$CR^3 = \{HRU^{31}\}$	HRU^{31}= 225
4	$CR^4 = \{HRU^{41}, HRU^{42}\}$	HRU^{41}= 171, HRU^{42}= 25
5	$CR^5 = \{HRU^{51}, HRU^{52}\}$	HRU^{51}= 19, HRU^{52}= 115
6	$CR^6 = \{HRU^{61}, HRU^{62}\}$	HRU^{61}= 210, HRU^{62}= 9

CR^4~CR^6은 HRU^{Tn}의 신뢰도의 곱이 가장 높은 노드가 최종결과로 선택된다.

3. 실험 및 결과

본 논문에서 구현된 문자인식 시스템에 대한 성능 실험은 200dpi의 해상도로 입력된 실제 우편봉투의 영상에서 추출한 인쇄체 60여 만자 중, 한글 467,868자에 대해서 수행하였다. 수집된 우편봉투 영상은 재질, 창의 유무, 글자체 등이 매우 다양한 형태이며, 이진화의 영향에 따라 문자의 획이 끊어지거나 잡영이 존재하는 등 저품질의 문자영상이 많다.

표 1의 각 다층신경망의 입력층 노드 수는 HRU^{Tn}에 따라 120개~360개이다. 중간층 노드 수 역시 HRU^{Tn}에 따라 30개 또는 50개이다. 입력층 및 중간층 노드의 수는 실험적으로 결정된 것으로 최적화된 것은 아니다. 모든 다층신경망의 학습률은 0.1, 관성항은 0.7로 두어 100회 반복학습을 하였다. 학습과 테스트에 사용된 문자영상의 수는 각각 233,936자와 233,932자이다. CR^0의 인식률은 평균 99.9%를 보여주고 있지만, 본 논문의 실험에서는 CR^0의 인식률이 100.0%라고 가정하고 1~6형식의 문자영상에 대해서만 실험을 하였다.

최종 문자인식률 실험 결과는 표 2 및 그림 4와 같다. 기존 방법 [4]에서의 인식률 98.42%에서 98.80%로 상승하였으며, 이는 23.61%의 오류가 감소한 것이다. 그리고, 기존 방법에 비해 1~5형식의 문자 인식률이 상승하였으나, 6형식의 경우는 하락하였다. 하지만, 전체 데이터에서 2.62%만이 6형식 속해 전체 문자 인식률에는 크게 영향을 미치지 못했다.

표 2. 최종 문자인식 실험결과

형식	기존 방법	제안 방법	문자 수
1	98.88%	99.19%	53,292
2	99.41%	99.80%	37,110
3	99.08%	99.43%	12,346
4	98.06%	98.38%	76,583
5	97.59%	98.19%	48,472
6	98.34%	98.01%	6,129
합계	98.42%	98.80%	233,932

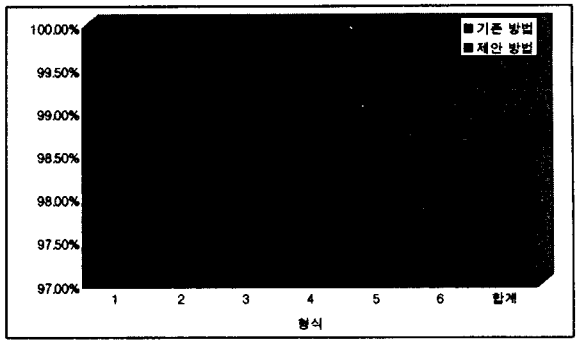


그림 4. 최종 문자 인식률

4. 결론

본 논문에서는 낱자 인식기와 자소 조합 인식기를 병합한 인쇄체 문자인식 방법을 제안하였다. 입력 문자영상의 유형을 7형식으로 분류한 뒤, 한글 문자는 자소의 조합 형태에 따라 유형별로 분류하여 각 유형별로 인식하였다. 1~3형식의 경우 낱자 인식기를 이용하였으며, 4~6형식의 경우 입력 문자영상이 1~3형식 정도의 복잡도를 가지도록 HRU 단위로 분리한 뒤 인식을 수행하였다. 각 HRU마다 다층신경망 인식기를 이용하여, 신뢰도가 가장 높은 노드를 최종 결과로 하였다. 제안한 방법을 200dpi의 우편영상에서 추출한 233,932자의 테스트 데이터에 대해 실험한 결과 98.80%의 높은 인식률을 얻을 수 있었으며, 이는 기존 방법에 비해 23.61%의 오류가 감소한 것이다.

참고문헌

[1] 최동혁, 류성원, 강현철, 박규태, " 계층구조 신경망을 이용한 한글 인식", 대한전자공학회 논문지, 제 28권 B편 제 11호, pp. 1-7, 1991.

[2] 권재욱, 조성배, 김진형, " 계층적 신경망을 이용한 다종 크기의 다중활자체 한글문서 인식", 한국정보과학회 논문지, 제 19권 제1호, pp. 69-79, 1992.

[3] 이진수, 권오준, 방승양, " 개선된 자소 인식 방법을 통한 고인식률 인쇄체 한글 인식", 한국정보과학회 논문지, 제 23권 제 8호, pp. 841-851, 1996.

[4] 임길택, 김호연, 이상호, 송재관, 남윤석, " 우편물 자동구분을 위한 문자인식 시스템", 대한전자공학회 컴퓨터/반도체 소사이어티 추계학술대회, 제 25권 제 2호, pp. 103-106, 2002.