제21회 한국정보처리학회 춘계학술발표대회 논문집 제11권 제1호 (2004. 5)

낱자 특징 기반 자소 인식기를 이용한 인쇄체 한글 인식방법

장승익, 남윤석
한국전자통신연구원 우정기술연구센터
e-mail : {sijang, ysnam}@etri.re.kr

A Method of Machine-Printed Hangul Recognition using Grapheme Recognizer

SeungIck Jang, Youn-Seok Nam
Postal Technology Research Center, ETRI

요 약

본 논문에서는 낱자에서 추출한 특징을 입력으로 사용하는 자소 인식기를 이용한 저해상도 인쇄체 한글 영상의 인식 방법을 제안하였다. 제안한 방법에서는 입력 문자를 한글 6 형식과 기타 형식의 문자, 총 7 종으로 분류한 뒤, 입력 문자를 인식 대상 문자의 수와 자소 복잡도에 따라 하나 또는 두 개의 인식 단위로 구분하여 인식한다. 각 HRU 는 낱자에서 추출한 방향각 특징을 입력으로 사용하는 다층 신경망 인식기를 이용하여 인식한다. 다음으로, 각 다층 신경망 인식기의 신뢰도를 조합하여 최종 인식 결과를 도출한다. 제안한 방법을 사용한 실험에서 98.99%의 인식률을 얻을 수 있었으며, 이는 기존 방법에 비해 15.83%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 300dpi 로 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 낱자에서 추출한 특징을 입력으로 사용하는 자소 인식기를 이용한 저해상도 인쇄체 한글 영상의 인식 방법을 제안한다. 제안한 방법에서는 입력 문자영상을 한글 6 형식과 기타 형식의 문자, 총 7 종으로 분류한 뒤 각각의 유형별 다층 신경망 인식기를 이용하여 입력 문자영상을 인식한다. 입력된 한글 문자는 형식에 따른 인식 대상 문자의 수와 자소 조합 복잡도에 따라 하나 또는 두 개의 인식 단위로 구분하여 인식한다. 제안한 방법을 사용하여 우편영상에서 추출한 200dpi 의 해상도를 가지는 233,932 자의 한글 테스트 영상에 대해 실험한 결과 98.99%의 인식률을 얻을 수 있었다

2. 문자인식 시스템

2.1 시스템 흐름

저해상도 인쇄체 한글 영상 인식 시스템의 전체 흐름도는 그림 1 과 같다. 입력 문자영상에서 문자영상의 방향각 특징을 추출하고, 이를 다층 신경망으로 구성된 유형 분류기를 통해 입력 문자영상의 유형을 분류하게 된다. 이때 한글은 그림 2 에서 보여지는 것과 같이 초성, 중성, 종성의 배치 형태에 따라 1 형식에서 6 형식까지 분류하도록 하였다. 한글 이외의 문자인

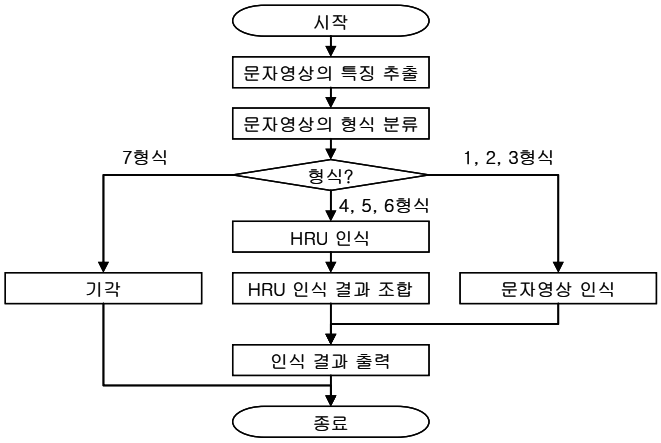
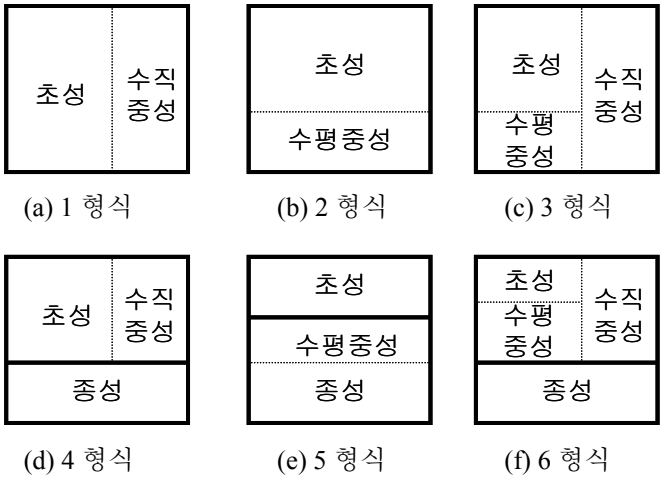


그림 1. 인식 시스템의 전체 흐름

영문자, 숫자, 기호 등은 7 형식으로 분류를 하였다. 본 논문에서는 7 형식 문자는 인식 대상에서 제외하였다.

유형 분류기의 결과가 1~6 형식으로 나올 경우, 문자영상의 인식 단위를 그림 2 에서 보여지는 것과 같이 구분한 뒤 인식을 수행한다. 여기서, 한글 문자의 기본 인식 단위를 HRU(Hangul Recognition Unit)로 정의한다. 그림 2 에서 굵은 실선으로 둘러진 사각형이 하나의 HRU 를 나타내며, 점선은 HRU 를 구성하는 자소의 경계를 나타내고 있다. 인식기의 입력으로는 형식분류를 수행하기 위해 전체영상에서 추출한 방향각 특징을 그대로 사용한다. 입력 문자영상이 1, 2, 3 형식인 경우, HRU 가 입력된 문자영상과 동일하기 때문에 전형적인 낱자 구분기와 차이가 없다. 다음으로 입력 문자영상이 4, 5, 6 형식인 경우, 입력영상은 각각 2 개의 HRU 로 구분된다. 이 경우 각각의 HRU 인식기에 전체영상에서 추출한 방향각 특징을 그대로 입력으로 사용하고, 각 인식기의 출력 결과를 조합하여 입력 영상에 대한 결과를 도출하게 된다. 문자영상의 인식 결과는 HRU 인식기의 신뢰도를 바탕으로 결정하게 된다. 1, 2, 3 형식 문자의 경우, 낱자 인식기의 신뢰도를 사용하여 신뢰도가 가장 큰 결과를 최종결과로 선택한다. 4, 5, 6 형식 문자의 경우 자소 조합 인식기의 신뢰도의 곱이 가장 큰 결과를 최종결과로 선택한다.



(a) 1 형식 (b) 2 형식 (c) 3 형식

(d) 4 형식 (e) 5 형식 (f) 6 형식

그림 2. 한글의 6 형식 및 HRU 분할 방법

2.2 HRU 인식기

한글을 6 형식으로 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 분할에 의한 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, 분할에 따른 인식 오류의 중첩으로 인식률이 저하될 수 있다. 또한 분할을 실패하는 경우도 발생할 수 있다.

본 논문에서 제안하는 방법은 앞에서 언급한 두 가지 방법의 장점만을 취하는 방법이다. 먼저, 1, 2, 3 형식과 같이 간단한 형태의 문자는 분할을 시도하지 않고, 입력 문자영상을 그대로 인식함으로써 분할 오류를 제거하였다. 다음으로, 4, 5, 6 형식과 같이 복잡한 형태의 문자는 초성 또는 종성만을 인식하는 인식기와 나머지 부분을 인식하는 인식기 두 개를 조합하여 인식함으로써 1, 2, 3 형식의 인식기와 비슷한 복잡도를 가질 수 있다. 각 형식의 인식 방법을 자세히 설명하면 다음과 같다.

유형별 문자 인식기는 입력의 유형에 따라서 낱자 및 자소 조합 인식기로 구성되며, 식 (1)과 같이 표현된다. 식 (1)에서 HRU^{Tn} 은 유형 T 의 n 번째 HRU 를 인식하는 인식기며, N^T 는 유형 T 의 HRU 의 개수이다. 1~6 형식 인식기는 CR^1~CR^6 으로 표현하고, 유형 분류기는 CR^0 로 표현한다.

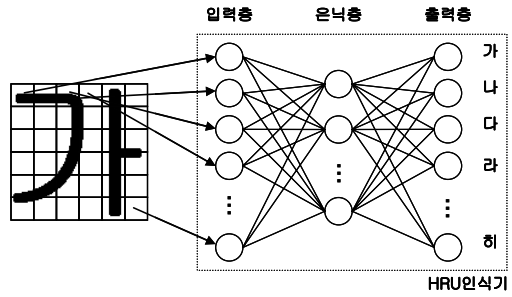
$$CR^T = \{HRU^{T1}, HRU^{T2}, \cdots, HRU^{Tn}\},\ 1 \le n \le N_T \qquad \text{식 (1)}$$

각 문자 인식기의 구조는 표 1 과 같으며, 모두 다층 신경망 기반의 인식기이다. 최종 결과는 CR^1~CR^3 의 경우, 낱자 인식기의 출력층에서 신뢰도가 가장 높은 노드가 되며, CR^4~CR^6 은 HRU^{Tn} 의 신뢰도의 곱이 가장 높은 노드가 최종결과로 선택된다.

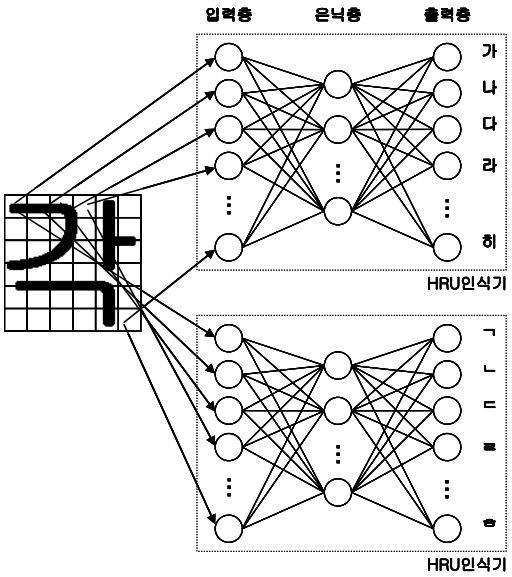
본 논문에서 제안한 HRU 인식 방법의 개념은 그림 3 과 같다. 먼저, 1, 2, 3 형식의 한글은 그림 3 의 (a)와 같이 입력 영상에서 6 by 6 의 비선형 격자를 씌워 추

표 1. 문자 인식기 구조

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



(a) 1, 2, 3 형식의 문자 인식기



(b) 4, 5, 6 형식의 문자 인식기

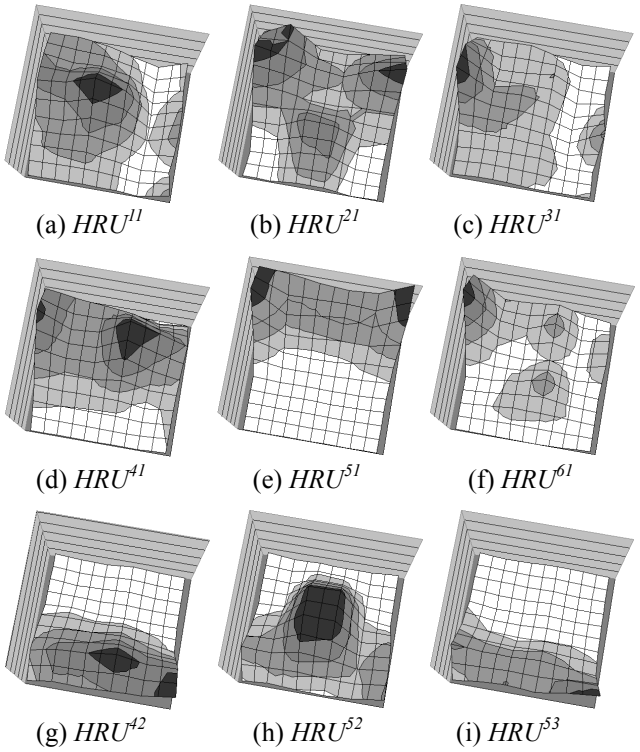
그림 3. 한글 문자 인식기의 구조

출한 방향각 특징을 다층 신경망의 입력으로 사용한다. 출력층의 노드 수는 각 형식에서 조합 가능한 한글의 수이며, 표 1 에서 도시되어 있는 것과 같다. 다음으로 4, 5, 6 형식의 한글은 그림 3 의 (b)와 같이 두 개의 HRU 인식기를 이용하여 인식하게 된다. 다층 신경망의 입력은 1, 2, 3 형식에 사용한 특징과 동일한 특징을 사용하며, 출력층은 자소 또는 자소 조합의 수이다. 입력영상은 분할하지 않은 상태에서 추출한 특징을 각 HRU 인식기의 입력으로 사용한다. 즉, 인식 유형에 차이 없이 모든 HRU 인식기에 동일한 형태의 특징이 사용된다. 하지만 동일한 특징을 사용하더라도 HRU 인식기의 출력층의 구조는 형식에 따라 상이하다.

3. 실험 및 결과

3.1 실험 데이터 및 훈련

본 논문에서 구현된 문자인식 시스템에 대한 성능 실험은 200dpi 의 해상도로 입력된 실제 우편봉투의 영상에서 추출한 인쇄체 60 여 만자 중, 한글 467,868 자에 대해서 수행하였다. 이중 233,936 자는 다층 신경망 학습을 위해 사용하였으며, 233,932 자는 테스트 데이터로 사용하였다. 수집된 우편봉투 영상은 우편물의 재질, 창의 유무, 글자체 등이 매우 다양한 형태이며,



(a) HRU^{11} (b) HRU^{21} (c) HRU^{31}

(d) HRU^{41} (e) HRU^{51} (f) HRU^{61}

(g) HRU^{42} (h) HRU^{52} (i) HRU^{53}

그림 4. 신경망 가중치의 에너지

이진화의 영향에 따라 문자의 획이 끊어지거나 잡영이 존재하는 등 저품질의 문자영상이 많이 존재한다.

각 다층 신경망의 입력층 노드 수는 사용하는 특징에 의존적이다. 본 논문에서는 입력영상을 6 by 6 의 비선형 격자로 분할한 뒤, 각각에서 10 개의 방향각 특징을 추출하여 총 360 차원의 특징을 사용하였다. 은닉층의 노드 수는 70 개로 고정하였으며, 출력층의 수는 표 1 에서 도시된 것과 같이 설정하였다. 모든 다층 신경망의 학습률은 0.1, 관성항은 0.7 로 두어 100 회 반복학습을 하였다. 다층 신경망과 관련된 상수들은 실험적으로 좋은 성능을 보이는 값들이다.

그림 4 는 훈련이 끝난 HRU^{1n}~ HRU^{6n} 의 입력층과 은닉층 사이의 신경망 가중치의 에너지를 표현한 것이다. 에너지는 다층 신경망 가중치의 제곱의 합으로 계산한 뒤, 에너지가 높은 부분은 짙은 색으로, 에너지가 낮은 부분은 옅은 색으로 표현하였다. 그림에서 볼 수 있듯이 문자 인식을 수행함에 있어서 중요하다고 판단되는 부분의 가중치가 높은 것을 알 수 있다. 예를 들어 4 형식 문자에서 초성 및 중성을 인식하는 HRU^{41} 의 경우 그림 4 의 (d)와 같이 상단쪽에 높은 가중치로 학습이 되었음을 알 수 있으며, 종성을 인식하는 HRU^{42} 의 경우 그림 4 의 (g)와 같이 하단쪽이 높은 가중치로 학습이 되었음을 알 수 있다.

3.2 실험결과

입력되는 문자의 형식을 구분하는 CR^0 의 인식률은 99.9% 이상으로 높게 나타난다. 하지만, 본 논문에서는 실험의 편의성을 위해서 CR^0 의 인식률이 100.0% 라고 가정하고 각각의 형식별 문자영상에 대해서 실험을 수행하였다.

표 2. 인식 실험결과

	문자 수	기존방법1	기존방법2	기존방법3	제안 방법
1형식	53,292	98.88%	99.19%	99.19%	99.19%
2형식	37,110	99.41%	99.80%	99.80%	99.80%
3형식	12,346	99.08%	99.43%	99.43%	99.43%
4형식	76,583	98.06%	97.75%	98.38%	98.70%
5형식	48,472	97.59%	97.94%	98.19%	98.54%
6형식	6,129	98.34%	97.41%	98.04%	98.53%
전체	233,932	98.42%	98.52%	98.80%	98.99%

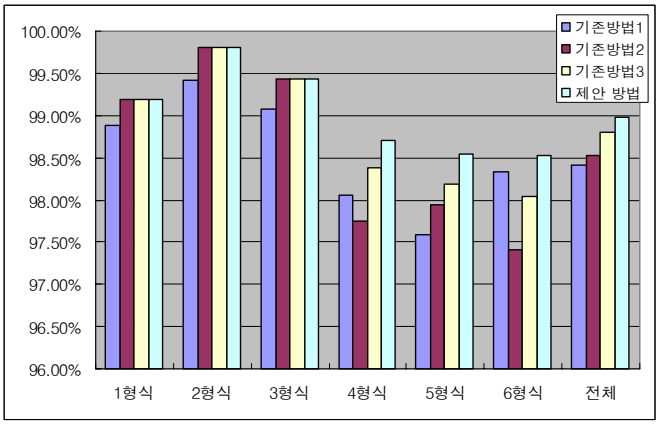


그림 5. 인식 실험결과의 비교

테스트 데이터에 대한 문자 인식률 실험 결과는 표 2 및 그림 5 와 같다. 기존 방법 1 은 [4]에서 제안한 방법이며, 기존 방법 2 는 낱자기반 인식기를 사용한 방법이며, 기존 방법 3 은 [5]에서 제안한 방법이다.

1 형식에서 3 형식까지는 기존 방법 2, 기존 방법 3 과 제안 방법이 동일한 성능을 보여주고 있는데, 이는 이들이 모두 동일한 인식기를 사용하기 때문이다. 나머지 4 형식에서 6 형식까지는 기존 방법들에 비해서 모두 성능 향상이 있었음을 알 수 있다.

본 논문에서 제안한 방법이 테스트 데이터에 대해 98.99%의 높은 인식률을 보이고 있음을 알 수 있다. 이는 기존 방법들 중에서 가장 높은 인식률인 기존 방법 3 의 98.80%의 인식률에 비해 15.83%의 오류가 감소한 것이다.

4. 결론

본 논문에서는 낱자 특징 기반의 자소 인식기와 낱자 인식기를 조합한 인쇄체 문자인식 방법을 제안하였다. 입력 문자영상의 유형을 7 형식으로 분류한 뒤, 한글 문자는 자소의 조합 형태에 따라 유형별로 분류하여 각 유형별로 인식하였다. 1 형식에서 3 형식 한글의 경우 낱자 인식기를 이용하였으며, 4 형식에서 6 형식 한글의 경우 입력 문자영상이 낮은 복잡도를 가지도록 HRU 로 구분한 뒤 인식을 수행하였다.

각 HRU 인식기로는 다층 신경망을 이용하였다. 각 HRU 인식기의 학습 결과, HRU 인식기가 인식하는 부분에 대해서만 높은 가중치를 가지도록 학습이 되었음을 알 수 있었다. 인식 결과는 각 HRU 인식기에서 가장 높은 신뢰도가 출력된 노드를 최종 결과로 확정하였다.

본 논문에서 제안한 방법을 200dpi 의 우편영상에서 추출한 233,932 자의 테스트 데이터에 대해 실험한 결과 98.99%의 높은 인식률을 얻을 수 있었으며, 이는 기존 방법에 비해 15.38%의 오류가 감소한 것이다.

참고문헌

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

[2] 권재욱, 조성배, 김진형, “계층적 신경망을 이용한 다중 크기의 다중활자체 한글문서 인식”, 한국정보과학회 논문지, 제 19 권 제 1 호, pp. 69-79, 1992.

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

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

[5] 장승익, 임길택, 김호연, 정선화, 남윤석, “낱자 인식기와 자소 조합 인식기를 혼용한 인쇄체 한글 인식 방법”, 정보과학회 춘계학술대회, 제 30 권 제 1 호(B), pp. 244-246, 2003.