가중치 패턴 클러스터를 이용한 한글 문자 인식

김도형○ 이선화 차의영
부산대학교 일반대학원 전자계산학과
{dhkim, somy77, eycha}@harmony.cs.pusan.ac.kr

The Recognition of The Korean Characters Using The Weighted Pattern Cluster

Do-Hyung Kim○ Seon-Hwa Lee Eui-Young Cha
Dept. of Computer Science, Pusan National University

요 약

본 논문에서는 스캐너로 입력된 한글 문서 영상에서 한글 문자를 인식하는 방법을 제시한다. 입력된 한글 문자를 한글의 구조적 특징에 따라 6개의 유형으로 분리하고, 각 유형에서의 모음의 형태학적 특징에 근거하여 모음을 인식한다. 각 유형에서의 자음의 인식을 위해서 가중치 패턴 클러스터를 생성하고 생성된 클러스터와 원영상간의 유사도 측정을 통해 자음을 인식하게 된다. 오인식 가능성이 있는 자음은 오인식 교정을 위한 세부 유사도 매칭과정을 통해 최종적으로 인식된다. 제안하는 알고리즘을 바탕으로 실험한 결과 스캐너로 입력받은 상용 한글 문자 14,983자에 대해 최종 95.68%의 인식률을 보였으며, 차후 정형화된 한글 문서 인식 시스템에 응용될 수 있을 것이다.

1. 서 론

스캐너나 카메라를 통해 입력받은 문서영상을 처리하여 문자를 인식하는 기술은 새로운 컴퓨터 입력장치로서 여러분야에 있어서 그 필요성이 대두되고 있다. 이러한 문자 인식 기술은 키보드에 의한 문자 입력 방식을 대체하고 있으며, 사무 자동화 및 정보화가 급격히 진행됨에 따라 그 사용 범위가 더욱 광대해지리라 예상된다.

문자인식 기술은 전자펜 등의 문자입력기에 의해 시간적, 공간적 정보를 가짐으로써 획순, 획수, 필기방향이나 속도 등의 부가 정보를 가질 수 있는 온라인 정보 입력에 따른 인식 기법과 종이에 작성된 문서를 스캐너 등에 의해 입력하여 공간적 밝기 정보를 가지는 오프라인 정보 입력에 따른 인식 기법으로 크게 분류할 수 있다[1].

일반적으로 인쇄체 문자가 기계에 의한 제한된 활자체로 구성되어 각 문자가 불변적인 특성을 가지는 반면, 필기체 문자는 더 복잡한 변형을 포함하고 있으므로 복잡한 변형을 흡수해야 된다는 점에서는 필기체 문자 인식이 더 난해하다. 하지만 인쇄체 문자일 경우에도 실제로 입력된 문자를 신뢰성 있게 인식할 수 있는 일반화된 방법의 제시가 어려운데 이는 몇가지 문제점이 존재하기 때문이다. 앞에서 언급한 바와 같이 인쇄체 문자는 온라인 정보를 사용할 수 없다는 문제점이 있고, 또한 프린터와 스캐너에서 발생하는 잡음이 추가되어 영상에 변형이 생길 뿐만 아니라, 글자체를 미적으로 보기 좋게 하고자 하여 발생하는 각 자소간의 접촉과 자소의 크기 및 위치의 변형이 인쇄체 한글 문자의 인식을 어렵게 한다.

이에 본 논문에서는 일반화된 오프라인 인쇄체 한글 문자의 인식 방법을 제한한다. 입력되는 한글 문자의 기본자소의 배치에 따라 6가지 유형으로 분리하고, 각 유형에서의 모음의 형태학적 특징에 근거하여 모음부를 인식하게 된다. 각 유형에서의 자음의 인식은 생성된 가중치 패턴 클러스터와 원영상간의 유사도 측정 과정을 거쳐 수행된다. 또한 형태학적으로 유사한 자음의 오인식 교정을 위한 세부 유사도 매칭 과정을 거쳐 최종적으로 한글 문자를 인식하게 된다. 제안된 방법은 잡음에 강인하며, 글자간의 접촉 및 형태학상의 변형에도 강인한 인식률을 보인다.

본 논문은 다음과 같이 구성된다. 1장 서론에 이어, 2장에서는 전체 시스템 구성을 간략하게 설명하고, 3장에서는 한글의 유형 및 모음 인식에 대하여 기술한다. 4장에서는 가중치 패턴 클러스터를 이용한 자음의 인식 방법을 제안하며, 5장에서는 실험결과를 분석하고 마지막으로 결론을 맺는다.

2. 전체 시스템 구성

인쇄체 한글 문자 인식 시스템은 그림 1과 같은 단계로 구성된다. 먼저 스캐너로 입력된 영상의 한글 문자부분을 획득하여 한글의 구조적 특성에 따른 타입 판별 과정을 거친다. 각 타입에 따른 모음의 위치는 대체로 일정한 위치에 존재하므로 사전 지식에 따른 형태학적 분류 방법에 따라 모음을 인식한 후 자음 정규화 과정을 거치고, 미리 생성된 자음 가중치 패턴 클러

스터를 이용한 유사도 검출기법으로 자음 부분을 인식한다. 인식된 자음이 오인식 가능성이 있는 자음일 경우 세부 판별 과정을 거쳐 최종적으로 한글 문자를 인식하게 된다.

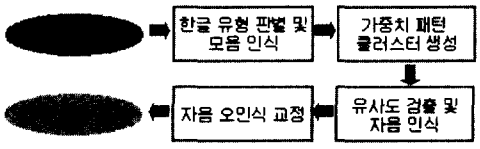

그림 1. 한글 문자 인식 과정

3. 한글 유형 및 모음 인식

3.1 한글 유형

한글은 부류수가 방대하고 글자간 유사성이 매우 높기 때문에 패턴의 구별이 매우 어렵다. 이러한 방대한 문자의 인식을 위해서는 서로 비슷한 특성을 가지는 문자별로 분리하여 각각의 특성에 따라 문자를 인식하는 것이 인식률 향상에 유리하다고 할 수 있다. 한글은 기본 자소의 위치와 종성의 유무에 따라 그림 2와 같이 6가지 유형으로 분류되는 구조적 특징을 가지고 있다[2].

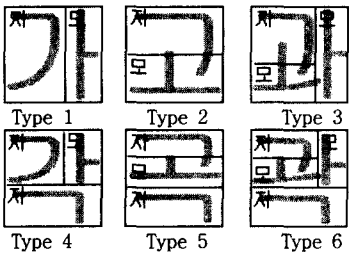

그림 2. 구조적 특징에 따른 한글 유형의 분류

한글의 형태를 분류하기 위한 기존 방법으로 신경망을 이용하여 분류하는 방법[3][4]이 주류를 이루고 있으나 각 타입의 훈련 패턴간의 유사도가 낮아 타입 분류의 정확도가 다소 떨어지는 단점을 보이며, 훈련 패턴의 양이 방대하여 그 훈련 시간에 따른 비용이 매우 크다. 따라서 본 논문에서는 한글 문자의 형태학적 특징에 기반한 방법으로 labeling을 통하여 세로 모음과 가로 모음을 판별한 후, 받침 자음의 유무를 따지는 방법으로 프로세싱의 간편성과 정확성 향상을 도모하였다.

3.2 모음 인식

입력 한글 문자 패턴이 6가지 한글 타입으로 분류되면, 각 타입에 따라 모음의 위치는 대체로 일정하다고 할 수 있다. 따라서 모음이 위치할 만한 관심 영역(ROI:Region of Interest)에 대하여 모음의 형태학적 특징에 따라 주획과 부획을 검색함으로써 최종 모음을 인식하게 된다. 이 때 원영상의 변형을 초래할 가능성이 있는 정규화 과정과 잡음 및 훼손된 영상에 민감한 세선화 과정은 거치지 않고, 전체 문자 영상의 가로, 세로 비율을 통해 모음을 인식하게 된다. 다음은 모음의 세로주획을 검색하는 과정을 나타내며, 가로주획 검색 과정도 이와 유사하다.

1) ROI 영역에서 연결 픽셀수가 임계치 이상인 column 검색
2) 검색된 각 column의 left, right neighbor를 포함한 확장 블럭 생성
3) Labeling 을 통한 생성된 블럭들의 영역 경계 설정
4) 최대 영역 높이를 가지는 블럭을 세로주획으로 인식

4. 자음의 인식

4.1 가중치 패턴 클러스터

한글 문자 인식에서는 일반적으로 자모를 분리하여 각각의 자모에 대해서 정규화 과정을 거친 후 신경망을 통한 문자 인식을 하는 방법들이 주류를 이루고 있다. 하지만 이는 자모가 이상적으로 분리된 비교적 깨끗한 영상에서는 높은 인식률을 보일 수 있을지 모르나, 영상의 훼손 등에 의해 자모의 일부분이 결합된 영상에서는 자모의 분리 자체가 매우 어려운 문제로 제기되어 높은 인식률을 기대하기가 어렵다. 이에 본 논문에서는 가중치 패턴 클러스터를 생성함으로써 자모를 분리하지 않고도 분리하여 인식하는 것과 같은 효과를 나타내는 방법에 대하여 제안한다. 가중치 패턴 클러스터의 생성을 위해 모음이 제거된 자음 부분만으로 구성된 영상을 정규화 한 후, 식(1)의 가중치 패턴 클러스터 생성식을 사용하여 각 타입의 자음 가중치 패턴 클러스터를 만든다. 이 때 '고','구' 와 같이 같은 타입이면서도 모음의 종류에 따라 자음의 위치와 형태가 달라지는 경우가 발생하므로, 모음의 종류에 따라 가중치 패턴 클러스터군을 달리한 집합을 생성한다. 그림 3은 타입 1의 자음에 대하여 생성된 가중치 패턴 클러스터의 일부를 보여주고 있다.

$$C(i,j)^{new} = \frac{f(i,j) + C(i,j)^{old} \| C(i,j)^{old} \|}{\| C(i,j)^{old} \| + 1} \quad (1)$$

$0 \le i < input\_width \ , \quad 0 \le j < input\_height$

$f(i,j)$: *input value of pixel(i,j) in input pattern* (0 *or* 1)

$C(i,j)^{new}$: *new value of pixel(i,j) in Weighted Pattern Cluster*

$C(i,j)^{old}$: *old value of pixel(i,j) in Weighted Pattern Cluster*

$\| C(i,j)^{old} \|$: *old member number of Weighted Pattern Cluster*

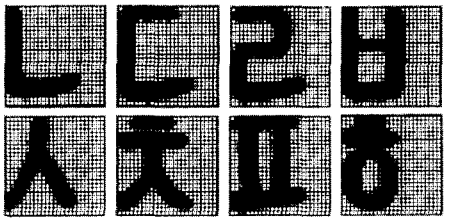

그림 3. 타입 1의 자음 가중치 패턴 클러스터의 예

입력 패턴의 자음 인식은 생성된 가중치 패턴 클러스터와의 유사도를 검출하여 가장 높은 유사도값을 나타내는 승자 패턴 클러스터를 선택함으로써 이루어진다. 유사도 검출 과정중 가중치 패턴 클러스터에서 0이 아닌 값만 유사도 검출에 참여한다. 따라서 입력 영상에서의 모음 부분은 유사도 검출에 영향을 주지 않아, 모음이 제거된 효과를 도모할 수 있다. 유사도 검출식은 식(2)와 같다.

$$SV_p = \sum_{i=0}^{N}\sum_{j=0}^{M}(\omega \cdot f(i,j)\cdot C(i,j)) \qquad \text{if } f(i,j) = Object$$

$$SV_n = \sum_{i=0}^{N}\sum_{j=0}^{M} C(i,j) \qquad \text{if } f(i,j) = Background \qquad (2)$$

$$SV = \frac{SV_p - SV_n}{N\ M}$$

ω : *weight* SV : *Similarity Value* N : *Width*, M : *Height*

제안된 가중치 패턴 클러스터를 이용한 유사도 검출 기법은 자모 분리시 일어나는 오류의 발생 가능성을 없애고, 잡영 및 자모가 결합된 영상에서도 강인하여, 전체 오류 발생률을 최소화 시킬 수 있는 장점을 가진다.

4.2 오인식의 교정

제안된 가중치 패턴 클러스터를 이용한 유사도 검출 기법은 한글 자음의 인식에 높은 인식률을 보인다. 하지만 형태가 유사한 자음일 경우 입력 패턴의 상태에 따라 오인식을 보이기도 한다. 따라서 오인식을 발생시킬 가능성이 있는 자음일 경우 (ㅁ/ㄷ, ㄷ/ㅌ, ㅂ/ㅁ, ㅋ/ㄱ)가중치 패턴 클러스터의 세부영역에 대해서만 식(2)와 같은 유사도 검출식을 사용함으로써 인식률 향상을 도모한다.

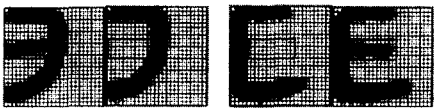

그림 4. 세부 유사도 검출 영역 설정 예(ㅋ/ㄱ, ㄷ/ㅌ)

5. 실험 및 결과분석

제안하는 알고리즘은 Pentium 450MHz, Memory 128Mbyte, Window98 환경에서 Visual C++ 6.0을 사용하여 구현하였다. 가중치 패턴 클러스터를 생성하기 위하여 1521개의 상용 한글을 설정하였다. 설정된 한글 문자를 포함하는 크기 10의 굴림체로 구성된 한글문서를 작성하고, HP ScanJect ADF 스캐너를 이용하여 400dpi로 입력받아 훈련 문자 집합을 생성하였다. 또한 입력 문자의 변형에 효율적으로 반응하기 위하여 좌,우로 회전된 영상 또한 같은 방법으로 총 4563자의 훈련 문자 집합을 생성하였다. 4563자의 훈련 문자 집합을 이용하여 각 타입에 따른 자음 가중치 클러스터를 생성하는데, 이 때 자음의 모양 및 위치는 모음의 종류(단모음,복모음/ㅜ,ㅗ)에 따라 달라짐으로 표 1과 같이 복수개의 클러스터 집합을 생성한다.

표 1. 타입별 가중치 패턴 클러스터집합 개수

타입	1	2	3	4	5	6	합
집합 개수	2	2	4	2	2	4	16

인식 실험에 사용될 인식 문자 집합은 2가지 집합으로 나누어 구성하였다. 인식 문자 집합(A)는 훈련 문자 집합 생성시와 같은 방법으로 굴림, 신명조, 바탕체 등 다양한 폰트와 9, 10, 11의 크기로 총 13,689자를 생성하였고, 인식 문자 집합(B)는 잡지, 논문, 처방전과 같이 실제 한글을 포함하는 다양한 문서에서의 한글 문자 1294자로 구성하였다.

표 2. 타입 분류 및 문자 인식 실험 결과 (Set A)

타입	문자 개수	타입분류 오류개수	문자인식 오류개수	전 체 오류개수	전 체 인식률
1	783	5	4	9	98.85
2	576	4	3	7	98.78
3	747	16	27	43	94.24
4	6552	42	56	98	98.5
5	3960	35	29	64	98.38
6	1071	39	46	85	92.06
합계	13,689	141	165	306	97.76

표 3. 타입 분류 및 문자 인식 실험 결과 (Set B)

타입	문자 개수	타입분류 오류개수	문자인식 오류개수	전 체 오류개수	전 체 인식률
1	323	4	8	12	96.28
2	265	5	5	10	96.23
3	132	9	4	13	90.15
4	302	6	9	15	95.03
5	218	11	9	20	90.82
6	54	7	6	13	75.92
합계	1,294	42	41	83	93.59

표 2,3과 같이 타입 6을 제외한 모든 타입에서 높은 인식률을 보이고 있으며, 전체 인식률은 95.68%였다. 타입 6에서의 인식률 저하는 문자 자체의 복잡성으로 인한 자모의 결합 및 원영상의 훼손등이 원인이라고 할 수 있으며, 향후 연구 대상이다.

6. 결 론

본 논문에서는 스캐너로 입력된 한글 문서 영상에서 한글 문자를 인식하는 방법을 제시하였다. 입력된 한글 문자를 한글의 구조적 특징과 형태학적 특징에 따라 6개의 유형으로 분리하고 모음을 인식하였다. 자음의 인식을 위해서 가중치 패턴 클러스터를 생성하고 생성된 클러스터와 원영상간의 유사도 측정을 통해 자음을 인식하는 방법을 제안하였다. 제안된 방법은 자모를 분리하는 과정을 제거함으로써 자모 분리시 나타날 수 있는 오류를 줄임으로서 전체 인식률 향상을 도모하였다. 인식결과에서도 높은 인식률을 보여 차후 정형화된 한글 문서 인식 시스템에 응용될 수 있을 것이다.

참 고 문 헌

[1] 안 창, 이상범, "한글처리-문자 중심 인식 고찰", 정보처리학회지, 제5권, 제5호, pp.48-53, 1998.

[2] 김명원,"신경 회로망을 이용한 한글 형태분류 및 음소 인식", 부산대학교 석사학위논문, 1990.

[3] 조성배, 김진형, 인쇄체 한글문자의 인식을 위한 계층적 신경망",한국정보과학회논문지,제17권,제3호, pp.306-316, 1990.

[4] 권재욱, 조성배, 김진형, "계층적 신경망을 이용한 다중 크기의 다중활자체 한글문서 인식", 한국정보과학회지, 제19권, 제1호, pp.69-79, 1992.