한글 인쇄체 문자의 형식 분류 및 비선형적 자소 분리에 관한 연구

박용민* · 김도현* · 차의영*

*부산대학교

A Study on Korean Printed Character Type Classification And Nonlinear Grapheme Segmentation

Yong-min Park* · Do-hyeon Kim* · Eui-young Cha*

*Pusan National University

E-mail : nohungry@gmail.com

요 약

본 논문에서는 한글 인쇄체 문자의 자소를 비선형적으로 분리하는 방법을 제안한다. 자소 분리 대상 문자는 자소의 조합 방식에 따라 6개의 형식으로 분류한다. 인쇄체 한글의 6형식 분류를 위해 그레이 레벨의 문자 이미지로부터 망 특성과 수직 · 수평 투영 기법을 이용해 특징을 추출하고, 오류역전파 기법을 이용하여 분류를 시도 한다. 분류된 문자 형식을 기반으로 분리 후보 영역을 지정하고, 이 영역을 기반으로 다단식 그래프 탐색 알고리즘을 이용하여 최적의 비선형적 자소 분리 경로를 찾아낸다. 실험 결과, 제안한 방법은 한글의 6형식 분류에 적합하였으며, 자소가 서로 붙어 선형적으로 분리가 어려운 문자의 자소 분리에 좋은 성능을 나타내었다.

ABSTRACT

In this paper, we propose a method for nonlinear grapheme segmentation in Korean printed character type classification. The characters are subdivided into six types based on character type information. The feature vector is consist of mesh features, vertical projection features and horizontal projection features which are extracted from gray-level images. We classify characters into 6 types using Back propagation. Character segmentation regions are determined based on character type information. Then, an optimal nonlinear grapheme segmentation path is found using multi-stage graph search algorithm. As the result, a proposed methodology is proper to classify character type and to find nonlinear char segmentation paths.

키워드

한글 인쇄체, 자소 분리, 오류역전파 신경망, 형식 분류

Ⅰ. 서 론

컴퓨터의 발전과 함께 급속히 성장하는 정보화 사회 속에서 사회 전반적으로 정보의 가공과 인식에 관한 관심이 높아지면서 은행 전표, 택배 전표, 주문서 등의 자동 입력, 주소 인식을 통한 우편물 자동화, 펜 컴퓨터, PDA, 온라인 서명 인식 등 다양한 분야에 한글 인식의 필요성이 증대하게 되었다.

한글 인식에 관한 범주는 인식 대상의 환경적 요인을 기준으로 했을 때 크게 온라인과 오프라인으로 경계를 구분 지을 수 있으며, 인식 대상 문자의 서체를 기준으로 했을 때는 필기체와 인쇄체로 분류할 수 있다.

인쇄체 문자는 필기체에 비해 형태의 변형이 작기 때문에 상대적으로 인식이 용이하여 문자 인식 연구 초기부터 관심의 대상이 되어 왔으며 근래에는 상용화 제품도 많이 나오고 있으나 한글의 경우는 사용자가 만족할 만큼의 높은 인식률을 보여주지 못하고 있다[1]. 이는 한글과 영어의 문자 수의 양을 상대적으로 비교했을 때 영어의 경우 알파벳 대소문자를 합쳐 52자에 불과하지만 한글의 경우 상용조합형 코드만 무려 11,172자이며 KS완성형 코드만 고려해도 2,350자가 된다. 따라서 영어에 비해 인식할 패턴 수가 월등히 많기 때문에 한글 인식이 영어 인식에 비해 쉽지 않음을 알 수 있다. 여기에 다양한 글자체(font)와 크기를 고려하면 문제는 더욱 복잡해진다.

이런 대용량의 한글 인식에 대한 문제를 해결하기 위하여 여러 개의 인식기들을 트리 구조로 결합시킴으로서 하나의 복잡한 문제를 여러 개의 단순한 문제로 나누어 해결하려는 방법이 일찍이 시도되었으며, 이를 위한 접근 방법은 크게 자소 기반 인식 방법과 음절 기반 인식 방법으로 나눌 수 있다[2]. 자소 기반 인식은 한글을 모음 형태에 따라 6가지 형식으로 분류한 후, 각 형식에 대한 자소의 위치가 일정하다는 특성을 이용하여 자소별로 문자를 인식하는 방법이다[3]. 자소 기반 인식은 자소 분리가 선행되어야 하지만 자소 분리는 글자체나 크기에 따른 변형, 자소의 위치에 따른 변형, 그리고 자소 간의 연결에 따라서 변형이 크기 때문에 어려운 문제인 동시에 해결해야할 과제이다.

따라서 본 연구에서는 그레이 레벨의 인쇄체 한글 문자 이미지로부터 망 특징과 수직, 수평 투영 특징을 추출한 후, 오류역전파 기법을 이용해 분류를 시도하고, 분류된 각 형식에 따라 비선형적 자소 분리 방법을 제안한다.

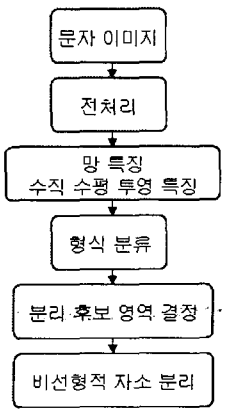

그림 4. 제안한 비선형적 자소 분리 알고리즘의 전체 흐름

II. 본 론

2.1. 특징 추출 및 형식 분류

한글 6형식은 이주근[4]에 의해서 제안된 정의로서, 한글의 자음과 모음의 조합에 따라 한글의 형식을 분류함으로써 형식에 따른 자음과 모음의 위치 등을 알 수 있어 인식 시에 유용한 정보로 활용할 수 있다.

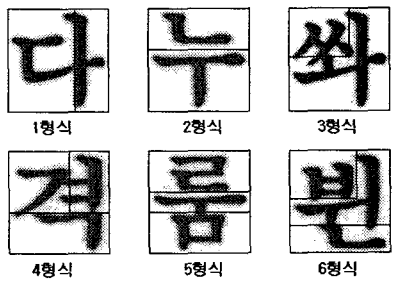

그림 5. 한글 6형식

그림 2에서 알 수 있듯이 각 형식 별로 자음과 모음의 위치는 유사하기 때문에 형식을 분류한 후, 형식 별 자소 분리 영역을 결정할 수 있으며, 이를 기준으로 자소 분리를 시도한다.

본 논문의 형식 분류 대상은 세리프(Serif)가 존재하여 특히 자소의 분리가 어려운 견명조체와 궁서체를 대상으로 했으며 완성형 한글에 의해 정의된 2350자에 대해서 분류를 시도하였다. 형식 분류는 오류역전파 신경망을 이용하였다.

2.2. 자소 분리 후보 영역

표1. 자소 분리 후보 영역

형식	자소 분리 후보 영역
1형식	$(x-gap, 0, x+gap, h)$
2형식	$(0, y-gap, w, y+gap)$
3형식	$(x-gap, 0, x+gap, h)$ $(0, y-gap, w, y+gap)$
4형식	$(x-gap, 0, x+gap, h)$ $(0, y-gap, w, y+gap)$
5형식	$(0, y_1-gap, w, y_1+gap)$ $(0, y_2-gap, w, y_2+gap)$
6형식	$(x-gap, 0, x+gap, h)$ $(0, y_1-gap, w, y_1+gap)$ $(0, y_2-gap, w, y_2+gap)$

각 형식별 자소 분리 후보 영역은 표 1과 같이

정의한다. 각 자소 분리 후보 영역은 Rect의 형태를 띄며 x, y는 자소 분리 영역의 시작 좌표이며, h와 w는 높이와 너비를 의미한다. gap은 탐색 영역의 폭을 조절하는 변수이다. 예를 들면 6형식의 경우 그림 2에서 알 수 있듯이 3개의 분리 영역이 존재하며, 수직 방향의 경우에 길이를 h로 정의하고, 수평 방향의 경우에 길이를 w로 정의하였다.

2.3 비선형적 자소 분리 알고리즘

한글 인쇄체 문자의 인식이 영문이나 숫자에 비해 어려운 이유는 서로 다른 문자의 경우에도 문자를 구성하는 자소 간의 구분이 작은 획 하나에 의해 결정될 정도로 유사성이 높기 때문이다.

기존 자소 분리 연구들은 대부분 형식 분류를 통해 입력 문자의 형식을 분류하고, 각 형식에 따른 자음과 모음의 위치 등을 판단하여 선형적으로 분리한다. 하지만 동일한 형식임에도 불구하고 문자를 구성하는 자음과 모음 자체의 특성으로 인해 자음과 모음의 위치가 균일하지 않으며, 그림 4에서 알 수 있듯이 선형적으로 자소를 분리할 경우 때문에 글자체(Font)의 특성에 의해 자소가 인접한 경우가 붙은 경우에 대해서는 분리에 어려움이 따른다.

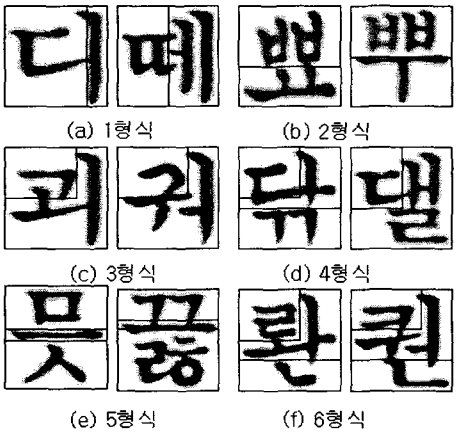

그림 6. 동일 형식에서의 분리 영역의 차이

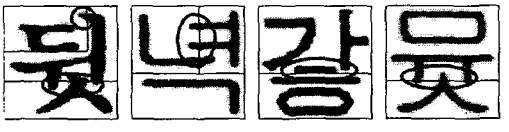

그림 7. 선형적 자소 분리의 한계

따라서 본 논문에서는 이러한 선형적 자소 분리의 한계를 극복하기 위해 다단식 그래프 탐색 알고리즘(Multi-Stage Graph Search Algorithm)을 이용한 비선형적 자소 분리 알고리즘을 제안한다.

다단식 그래프 알고리즘은 K개의 정점과 V개의 단계로 이루어진 다단식 그래프에서 최단 거리를 찾는 문제에 사용된다[5].

2차원 그리드 형태로 표현할 수 있는 문자 이미지의 각 픽셀을 다단식 그래프의 정점(vertex)으로 보면 그림 5와 같이 나타낼 수 있다.

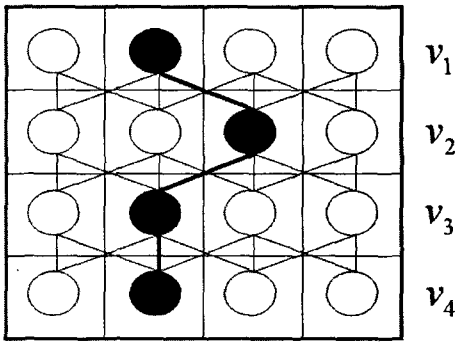

그림 8. 자소 분리 영역의 다단식 그래프로의 표현

그레이 레벨의 이미지에서 픽셀 값의 범위는 0 ~ 255 사이의 값으로 존재하며, 문자부분은 0에 가까운 값들이 분포하며, 배경부분은 255에 가까운 값들이 분포한다. 또한 문자 획의 중심 값이 문자 획의 가장자리 부분보다 작은 값을 가진다는 점으로부터 자소 분리 후보 영역 내의 경로 중에서 누적된 값이 최대를 가지는 경로가 자소 분리 경로라고 할 수 있다.

각 픽셀 간의 연결성은 수직 방향의 경우 임의의 픽셀 (x, y)에 대하여 (x - 1, y - 1), (x, y - 1), (x, y + 1)이 존재하며, 수평 방향의 경우에는 (x + 1, y - 1), (x + 1, y), (x + 1, y + 1)이 존재한다고 제한하며, 수식으로 표현하면 다음과 같다.

$$cost(x,y) = \max_{\substack{l \in V_{i+1} \\ <x,l> \in E}} \{I(x,y) + cost(x,l)\}$$

cost(x, y)는 픽셀 (x, y)에서의 픽셀의 누적된 최대값이며, I(x, y)는 픽셀 (x, y)의 밝기, E는 픽셀 (x, y) 와 연결된 픽셀들의 집합이다. 또한, 자소 분리 후보 영역 내에서 최대값을 가지는 경로 P는 초기 시작 지점으로부터 역추적을 통해서 계산할 수 있으며, 시간 복잡도는 $\theta(e+v)$[6]이다. 여기서 e는 정점의 개수이며, v는 그래프의 Stage 수이다.

III. 실험 및 결과

3.1 실험 환경

본 논문의 유용성을 검증하기 위하여 Pentium IVPC 2.4GHz, 1GB에서 Visual Studio 7.0을 이용

하여 Windows XP 상에서 형식 분류 및 자소 분리 실험을 하였다. 입력 문자 이미지는 HP ScanJet ADF 스캐너를 사용하여 견명조체와 궁서체를 대상으로 300 ~400 DPI 해상도로 입력 받았으며, 문자의 크기는 8 ~ 10을 대상으로 하였다.

3.2 형식 분류 실험

형식 분류는 견명조와 궁서체 각 각 2350자씩 4세트로 구성하여 9400개의 문자에 대해 실험을 하였다. 형식 분류기의 성능 검증을 위하여 전체 데이터의 2/3는 학습에 사용하고, 학습에 사용하지 않은 1/3의 3133개의 문자에 대해서 실험을 하였으며, 실험 결과 형식 분류 성공율은 99.06%로 나타났다.

3.3 비선형적 자소 분리 실험

자소 분리의 경우 성공률에 대한 정확한 성능 평가를 정량적으로 하기가 힘들다. 그 이유는 자소 분리가 잘 되었는지를 판단하는 명확한 기준이 없기 때문이다. 따라서 본 논문에서는 선형적 분리가 힘들었던 문자들에 대해 기존의 연구와의 비교를 통해 제안된 알고리즘의 유용성을 판단하고자 한다.

그림 7에서 알 수 있듯이 본 논문에서 제안하는 비선형적 자소 분리 알고리즘이 기존의 선형적 자소 알고리즘에 비해 자소 간에 인접하거나 접합된 상황에서 획의 손실이 적으면서도 자소 간의 분리를 잘 할 수 있음을 알 수 있다.

그림 10. 제안된 알고리즘과 선형적 자소 분리의 비교

Ⅳ. 결 론

본 논문에서는 그레이 레벨의 인쇄체 한글 문자를 오류역전파 기법을 이용해 분류를 시도하고, 분류된 각 형식에 따라 비선형적 자소 분리 방법을 제안하였다. 제안된 알고리즘의 유용성을 판단하기 위하여 실험한 결과 제안한 형식 분류 방법이 한글 형식 분류에 탁월한 효과를 나타냈으며, 기존의 연구에서 선형적 자소 분리가 어려운 경우의 문자에 대해서 제안한 비선형적 자소 분리 알고리즘을 적용하였을 때 보다 좋은 결과를 보였음을 알 수 있었다.

향후 연구 과제로는 분리된 자소 각 각에 발생하는 잡음을 제거하는 연구를 통해 보다 인식 성능을 향상시킬 수 있는 방안에 대해 연구할 계획이다.

참고문헌

[1]김도형, "개선된 자소 인식 방법을 이용한 한글 문자 인식기의 구현", 부산대학교 대학원 전자계산학과 이학석사 논문, 2002.

[2]http://netro.ajou.ac.kr/~imagelab/r_circular_pattern.htm

[3]정지호, 김희태, 최태영, "원형패턴벡터를 이용한 인쇄체 한글인식", 제11회 신호처리합동 학술대회, Vol. 11, No. 1, pp. 113 116, 1998.

[4]J. K Lee, "Korean Character Display by Variable Combination and its Recognition By Decompoistion method", Ph.D. Dissertation in Keio Univ. 1972.

[5]Seong-Whan Lee, Dong-June Lee, Hee-Seon Park, "A New Methodology for Gray-Scale Character Segmentation and Recognition", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 18, No. 10, October 1996.