

중심점을 이용한 무제약 필기체 한글 분할

박성호[†] 조범준

[†]전남도립 남도대학 컴퓨터정보통신과

조선대학교 컴퓨터공학과

shpark87@hanmir.com, bjcho@chosun.ac.kr

Segmentation of Unconstrained Handwritten Korean Characters using the Center of Gravity

Sungho Park[†] Beom-joon Cho

[†]Dept.. of Computer Information Communication, Namdo Provincial College

Dept.. of Computer Science Engineering, Chosun University

요 약

현재까지도 대부분의 필기체 한글 인식은 완벽한 분할을 가정하여 개발되고 있으며 한글 분할 그 자체에 대한 관심은 매우 낮은 편이다. 본 논문에서는 무제약 필기체 한글을 분할하기 위하여 중심점을 이용한 분할방법을 제안한다. 제안된 방법은 기존의 분할 방법들이 갖는 직선화된 분할경로 뿐만 아니라, 문자내에 존재하는 여백의 정보를 활용하여 꺾은선 형태의 분할경로도 찾을 수 있다. 이를 위해 먼저, 한글 문자열에 대한 전처리 과정과 수직투영을 통해 분할대상을 결정한다. 그리고 문자에서 특징점을 찾고 인접한 특징점들 간의 중심점을 찾아서 가상의 분할 경로를 생성한 뒤 최적 분할 경로를 결정한다.

1. 서 론

현재 우리는 주변에 존재하는 많은 자료를 컴퓨터에 정보화하여 기록함으로써 과거에 비해 시간적 경제적 비용 절감을 하고자 한다. 그 중에 한가지로 문서인식기술을 들 수 있는데, 우수한 문서인식을 수행하기 위해서는 먼저 처리 대상이 되는 문서의 문자열 영상이 정확하게 분리되어야 한다. 그러나 대부분의 한글 인식기는 미리 완벽한 분할을 가정하고 개발되고 있음에도 불구하고 한글 분할 그 자체에 대한 관심은 매우 낮은 편이다. 본 논문에서는 무제약 필기체 한글을 분할하기 위한 방법을 제안한다.

기존의 많은 분할 방법들은 주로 영문 문자열 [5,6,7,8]이나 디지털 숫자열을 대상으로 치중되어 제안되었으며 [4,9], 필기체 한글을 대상으로 하는 연구 [1,2,3]는 아직까지 미미한 실정이다. 문자열에서 단어 분할방법은 크게 두 가지로 구분될 수 있는데, 하나는 단어 분할 과정에서는 높은 분할 신뢰도를 갖는 단 하나의 분할 결과를 생성하는 방법이고 또 하나는 여러 분할 후보를 인식 과정으로 넘겨줌으로써 인식과 함께 단어 분할 결과를 정하도록 하는 방법이 있다. 전자는 외적분할(External Segmentation) 방법이라고 하며 처리속도를 향상시키지만 잘못된 단어 분할 결과를 수정

할 수 없다는 단점이 있다. 후자는 내적분할(Internal Segmentation) 방법으로 여러 단어 분할 결과를 수용하므로 단어 인식률은 좋지만 넘겨온 분할 결과만큼 여러 번의 인식을 수행해야 함으로 처리속도에 대한 부담이 크다.

일반적으로 무제약 필기체 한글인식의 경우는 분할 문제를 최소화하기 위하여 문자 영상의 수집과정에서 인위적인 분할을 위한 장치를 마련하거나, 영어로 대표되는 유럽언어권 문자와 한글의 구조적 차이점이 명확함에도 불구하고 지금까지 연구된 한글 분할방법은 한글의 특징을 충분히 반영하지 못한 채 영문자에 대한 분할 기술을 그대로 적용하는 경우가 많았다. 이는 결과적으로 분할된 인식 단위의 모호성을 가중시켜 시스템의 전체적인 성능을 저하시키는 주요 원인이 되었다.

본 논문에서는 무제약 필기체 한글 인식과정에서 필수적으로 선행되어야 하는 분할에 있어서, 높은 분할 신뢰도를 갖는 단 하나의 분할 결과를 생성하면서도 처리속도가 빠른 방법을 제안하고자 한다. 제안된 방법은 많은 분할 방법들이 사용하는 직선화된 분할 경로 보다는 자음과 모음의 조합으로 구성되는 한글의 여백 공간을 활용하는 꺾은선 형태의 분할 경로를 주로 사용함으로써 신뢰도를 높이고자 한다. 또한 크게 2단계로 분할과정을 구분하여 세밀한 분할이 요구되는 대상만을 선택하고, 선택된 대상에서도 최소한의 특징점만을 고려하면

서도 다양한 분할 경로를 찾을 수 있게 함으로서 처리 속도를 향상 시킨다.

2장에서는 제안된 분할 방법을 소개하고, 먼저 분할 대상을 선택하기 위해 요구되는 전처리 과정과 수직투영 방법을 설명한다. 3장에서는 선택된 분할대상에 대해 분할경로를 찾기 위해 사용되는 특징점 선택, 특징점 그룹화 그리고 중심점 결정에 대해 설명한다. 마지막으로 결론과 추후 진행 중인 연구에 대해 4장에서 설명한다.

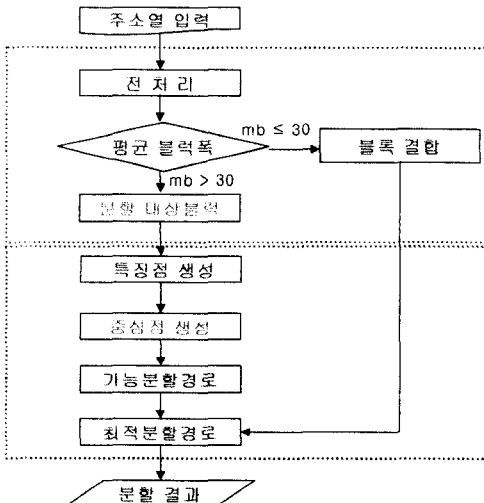
2. 분할대상 선택

2.1 분할방법의 절차

먼저, 입력 이미지에 대한 전처리과정과 수직투영을 통해 분할대상을 구분한다. 이때 평균블럭 폭이 30 pixel 이하이면 하나의 글자를 구성하는 것으로 판단하여 이전 블럭과 결합한다.

분할 대상블럭이 결정되면, 각 글자에서 끝점을 특징점으로 고려하고 인접한 특징점을 델라우닝(delaunay) 방법에 의해 연결한다. 그리고 더 세밀한 분할경로 생성이 가능하도록 하기 위해 연결된 세 특징점에 대해 무게 중심점을 생성한다.

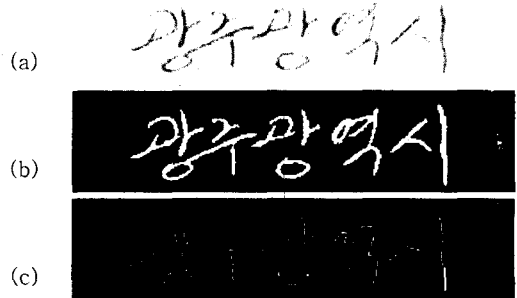
마지막으로 연결된 특징점과 중심점에 의해 가능 분할경로를 탐색하고 이 중에서 가장 최적 분할경로를 결정한다. 이 과정은 <그림 1>과 같은 순서도로 나타낼 수 있다.



<그림 1> 제안된 분할방법의 순서도

2.2 전처리

입력 이미지는 수직투영을 위해 이진화 처리하였고, 특징점을 고려하기 위해 세선화하였다.



<그림 2> (a)원본, (b) 이진화, (c) 세선화 이미지

2.3 분할대상 선택

이진화된 입력 이미지에 대해 수직투영을 수행하면 먼저, '역'은 그 자체를 분할된 결과로 볼 수 있으며 '시'에서 'ㅅ'과 'ㅣ'은 떨어진 것으로 보이지만 평균 블럭폭 이하이므로 연결된 글자로 판단한다. 따라서 '광주광'만이 접촉되거나 중첩되어서 연결된 문자열로 볼 수 있으며 이것이 분할대상이 된다.



<그림 3> 분할대상 선택

3. 분할경로 결정

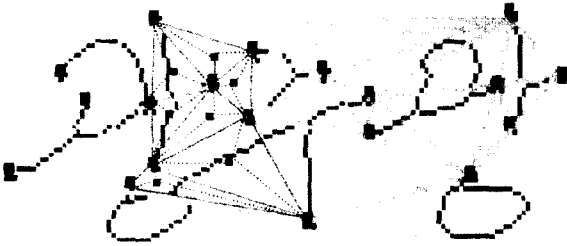
3.1 분할경로의 생성

다양한 분할경로를 생성하기 위하여 먼저, 글자의 특징점을 고려하는데, 이 때는 자음과 모음간의 여백을 분할경로로 활용할 수 있게만 하면 되므로 단지 글자의 끝점만을 특징점으로 고려한다.

그리고 고려된 특징점을 서로 연결하여 가상의 분할경로를 생성하기 위하여 인접한 3개의 특징점들을 그룹화한다. 이를 위해 본 연구에서는 델라우닝(Delaunay) 방법을 사용하였다.

마지막으로 더 세밀한 가상경로를 생성하여 더 정확한 분할 결과를 얻기 위하여 3개씩 그룹화 된 삼각형에 대해 중심점을 생성한다. 이것은 <그림 4>에서 볼 수

있다.

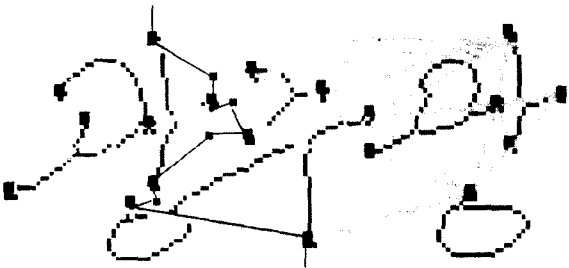


<그림 4> 특징점 그룹화와 중심점 생성

3.2 분할경로 탐색

가능 분할경로를 탐색하기 위하여 <그림 4>에 나타난 가상 경로 중에서 이미지(foreground)를 교차하는 경로는 불가능한 경로이므로 제외한다.

그리고 나서 하향식(Top-Down) 방법으로 가능한 분할경로를 탐색해 나가면서, 글자 이미지를 올바르게 가장 잘 분할할 수 있는 최적 분할경로를 결정한다. 분할 대상 글자 중에서 '광' 과 '주' 사이의 최적 분할경로에 대한 결과가 <그림 5>와 같이 나타난다.



<그림 5> 최적분할경로

4. 결론

본 논문에서는 무제약 필기체 한글을 분할하기 위하여 중심점을 이용한 분할방법을 제안한다. 제안된 방법은 기존의 분할 방법들이 갖는 직선화된 분할경로 뿐만 아니라, 문자내에 존재하는 여백의 정보를 활용하여 격은선 형태의 분할경로를 찾고자 하였다. 이를 위해 먼저, 한글 문자열에 대한 전처리 과정과 수직투영을 통해

분할대상을 결정한다. 그리고 선택된 분할대상에 대해 글자의 여백 활용을 위한 먼저, 간단한 특징점을 고려하였고, 가상의 분할경로를 얻기 위해 특징점을 연결하여 그룹화하였으며 더 세밀한 가상경로를 생성하기 위해 중심점을 생성하여 최적 분할경로를 결정하였다.

현재 수집된 필기체 데이터에 대하여 비교실험이 진행 중이며, 특히 글자가 접촉된 경우에 대해 분할 신뢰도를 높이기 위한 보완 연구가 진행중에 있다.

[참고문헌]

- [1] 김경환, 윤정석, "Run-Length Code를 이용한 제약없이 쓰여진 한글 필기체 주소열 분할", 한국정보과학회 논문지, 제28권 제11호, pp.813-821, 2001.
- [2] 정선화, 김수형, "GAP 군집화에 기반한 필기체 한글 단어 분리", 한국정보과학회 논문지, 제27권 제6호, pp.660-667, 2000.
- [3] 황순자, 김문현, "자소 클래스 인식에 의한 off-line 필기체 한글 문자 분할", 한국정보처리학회 논문지, 제3권 제4호, pp.1002-1013, 1996.
- [4] Ashraf Elnagar, Reda Alhaji, "Segmentation of connected handwritten numeral strings", Pattern Recognition, Vol. 36, pp.625-634, 2003.
- [5] Berrin Yanikoglu, Peter A. Sandon, "Segmentation of off-line cursive handwriting using linear programming", Pattern Recognition, Vol. 31, pp.1825-1833, 1998.
- [6] G. Kim, "Architecture for handwritten text recognition systems", Proc. Sixth International Workshop on Frontiers in Handwritten Recognition, pp.113-122, Taehon, Korea, August 1998.
- [7] G. Dzuba, A. Filatov and A. Volgunin, "Handwritten ZIP code recognition", Proc. Fourth International Conference on Document Analysis and Recognition, pp.766-770, Ulm-Germany, August 1997.
- [8] Shuyan Zhao, Zheru Chi, Penfei Shi and Hong Yan, "Two-stage segmentation of unconstrained handwritten Chinese characters", Pattern Recognition, Vol. 36, pp.145-156, 2001.
- [9] U. Pal, A. Belaid and Ch. Choisy, "Touching numeral segmentation using water reservoir concept", Pattern Recognition Letters, Vol. 24, pp.261-272, 2003.