

무제약 필기체 한글분할을 위한 가상 분할경로 생성

박성호, 조범준
전남도립 남도대학 컴퓨터정보통신과
조선대학교 컴퓨터공학과

Virtual Segmentation Path Construction for Unconstrained Handwritten Hangeul Segmentation

Sung-Ho Park, Beom-Joon Cho

Dept. of Computer Information Communication, Namdo Provincial College
Dept. of Computer Engineering, Chosun University

요 약

현재까지도 대부분의 필기체 한글 인식은 완벽한 분할을 가정하여 개발되고 있으며 한글 분할 그 자체에 대한 관심은 매우 낮은 편이다. 또한 한글은 자소의 모아쓰기 형태를 갖고 있어 문자 내에 여백이 존재할 뿐만 아니라 문자와 문자 사이의 접촉되는 특징이 흔히 발생되기 때문에 기존의 영문자에 사용된 방법을 그대로 직접 적용하기에는 많은 어려움이 있다. 본 논문에서는 무제약 필기체 한글을 분할하기 위하여 가상 분할경로를 이용하는 방법을 제안한다. 제안된 방법은 기존의 분할 방법들이 갖는 직선화된 분할경로 뿐만 아니라 문자내에 존재하는 여백의 정보를 활용하여 꺾은선 분할경로를 주로 이용하는데 특징이 있다. 분할과정은 크게 2단계로 구성되는데 첫 번째 단계는 입력받은 한글 문자열에 대한 전처리 과정과 수직투영을 통해 직선화된 분할경로를 찾고 이에 따라 쉽게 분할 가능한 문자와 중첩되거나 접촉된 문자를 구분하여 다음 단계의 분할대상으로 결정한다. 두 번째 단계에서는 다양한 가상의 꺾은선 분할경로를 생성하기 위하여 각 문자에서 특징점을 추출하고, 인접한 특징점들 간의 무게 중심점을 찾아서 더 세밀한 경로 생성을 함으로서 가장 최상의 분할경로를 결정한다.

1. 서론

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.

기존의 많은 분할 방법들은 주로 영문 문자열[6,7]이나 디지털 숫자열을 대상으로 치중되어 제안되었으며[4,9], 필기체 한글을 대상으로 하는 연구[1,2,3]는

아직까지 미미한 실정이다. 몇 가지 연구를 보면 먼저 Berrin[5]은 단어에서 가능한 문자 분할을 위하여 비용합수를 이용하는 알고리즘을 제시하였고, Zhao[8]은 무제약 중국문자를 분할하기 위하여 2단계로 이루어진 알고리즘을 제시하였다. Elnagar[4]은 연결된 필기체 숫자열에 대해 적절한 template를 사용하여 공간적인 특성을 이용한 방법을 사용하는 분할과정을 Pal[9]은 접촉된 숫자열에서 water reservoir 개념으로부터 얻어진 특징에 근거하여 최상의 분할 경로를 결정하는 방법을 제시하였다.

한글을 대상으로 하는 연구에서는 김경환[1]이 전처리와 인식단위 분할에 응용할 수 있는, 한글의 구조적 특성을 반영한 기술기 보정 알고리즘을 제안하고 필기자들의 일반적인 필기 습관과 한글이 갖는 2차원

구조의 특성을 반영하면서 문자의 접촉점을 찾아내기 위한 기초 함수들과 접촉점들의 분할방법을 제시하였다. 정선화[2]는 필기 영문 문자열의 단어 분할을 위해 제안된 기존의 세 가지 거리 척도를 채택하고 군집화에 기반한 세 가지 분류방법을 적용하여 한글 문자열의 단어 분리를 위한 최적의 조합을 선정하였다. 황순자[3]는 자소 클래스의 인식에 기반한 문자 분할 알고리즘을 제안 하였는데, 수직 투영에서 검출한 획으로부터 한글의 구조적 특성을 이용하여 자소 클래스 영역을 검출하였으며 한글의 자소 구성 법칙과 경험적 지식을 기반으로 하였다.

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 방법의 가장 큰 특징은 첫째, 많은 분할 방법들이 사용하는 직선화된 분할 경로 보다는 자음과 모음의 조합으로 구성되는 한글 구조적 특성에 의한 여백 공간을 활용하는 꺾은선 형태의 분할경로를 제시함으로써 특히 문자가 중첩된 경우에 대하여 신뢰도를 높이고자 하였다. 둘째, 분할과정을 크게 2단계로 구분하여 전체 문자열 중에서 보다 더 세밀한 분할이 요구되는 문자열만을 선택하였고, 선택된 문자열에 대해서도 최소한의 특징점만을 고려하면서도 다양한 가상의 분할 경로를 탐색할 수 있도록 제안함으로써 전체적인 처리속도를 높이고자 하였다.

2장에서는 제안된 분할 방법의 전체적인 절차를 소개하고, 전처리 과정과 수직투영 방법을 통하여 전체 문자열에서 보다 더 세밀한 분할이 요구되는 분할대상을 선택하는 과정을 설명한다. 3장에서는 선택된 분할대상 문자열에 대해 다양한 가상의 분할경로를 생성하는 과정을 설명한다. 이를 위하여 문자열에서 간단한 특징점을 선택하고 그것에 대하여 그룹화과정을 통해 가상경로의 생성, 그리고 더 세밀한 분할경로를 생성하기 위한 중심점 생성등에 대해 설명한다. 마지막으로 현재 진행 중인 비교실험과 추후 보완연구에 대한 내용을 4장에서 설명한다.

2. 제안된 분할 알고리즘

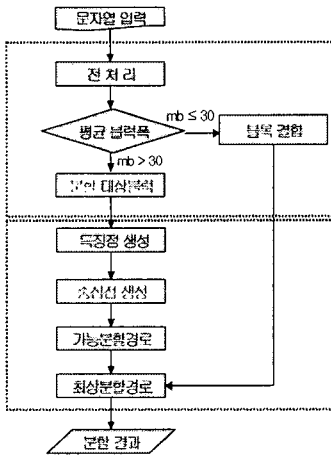
2.1 분할방법의 절차

본 논문에서 제안하는 분할 알고리즘의 전체 과정은 <그림 1>에 나타내었고, 그림에서 볼 수 있듯이 전체 과정은 크게 2단계로 구성되어 있다.

첫 번째 단계는 전체 문자열에서 세밀한 분할이 요

구되는 문자열만을 선택하는 과정으로서 입력받은 이미지에 대하여 수직투영을 통한 직선 분할경로가 생성된다. 본 논문에서는 100명의 서로 다른 필기자로부터 2~5개의 문자로 구성된 무제약 필기체 한글의 문자열을 수집 하였으며, 입력 이미지를 먼저 이진화 처리하여 gray 모드로 변환한 뒤 세선화 과정을 수행하였다. 이진화 처리된 gray 모드에 대하여 수직투영을 통하여 문자간의 중첩 및 접촉되지 않고 별개로 쉽게 구분 가능한 문자들을 찾아내었다. 즉 수직투영의 결과만으로 분할이 이루어지는 문자열은 첫 번째 단계에서 결정하고 나머지 문자열에 대해서만 다음 단계에서 고려된다. 또한 한글이 자음과 모음으로 구성되는 특성에서 하나의 문자이지만 수직투영에서 분리된 것처럼 판단되는 것을 방지하고자, 수집된 자료를 분석하여 평균 문자폭인 30 pixel 이하인 수직투영 결과는 하나의 문자를 구성하는 것으로 판단하여 바로 이전 블록과 결합해 줌으로서 하나의 문자가 구성되게 하였다.

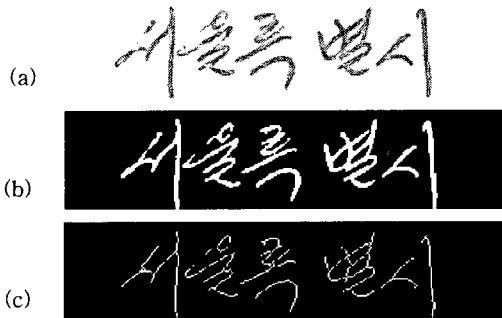
두 번째 단계는 주로 중첩되거나 접촉된 문자열에 대하여 다양한 가상의 분할경로를 생성하여 탐색함으로써 꺾은선 분할경로를 결과로 사용한다. 수직투영을 통하여 분할 대상블럭이 결정되면 가상의 분할경로를 생성하기 위하여 먼저 문자열에서 각 문자에서 끝점들을 특징점으로 고려하는데, 이는 네트워크에 비유할 때 각각의 노드에 해당된다고 볼 수 있다. 그리고 정해진 특징점들에 대하여 서로간의 경로를 생성하기 위하여 본 논문에서는 특징점을 삼각형 요소로 분할할 경우에 가장 질이 좋은 세변의 길이가 비슷한 요소로 분할되는 방법인 딜라우닝 방법을 사용하였다. 또한 보다 더 세밀한 분할경로를 생성하기 위하여 특징점이 연결된 삼각형에 대하여 각각의 무게 중심점을 고려하였고 이것을 삼각형의 세 특징점과 서로 연결하여 가상의 분할경로를 생성하였다. 마지막으로 연결된 모든 특징점과 무게 중심점들로 이루어진 가상의 분할경로를 탐색하였고, 문자 이미지를 손상시키지 않으면서 문자열을 가장 최상으로 분할할 수 있는 경로를 결정하였다.



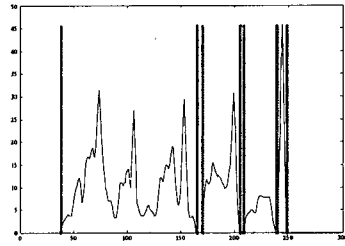
<그림 1> 제안된 분할방법

2.2 직선 분할경로

입력 이미지에 대하여 첫 번째 단계의 분할 알고리즘을 적용하여 전체 문자열에서 분할대상을 선택하는 과정을 살펴보면 다음과 같다. <그림 2(a)>는 한명의 필기자에 의해 총 5개의 문자로 쓰여진 원본 이미지의 한 예를 보여주고 있다. <그림 2(b)>와 같이 이진화 처리를 수행하여 gray 모드로 변환된 이미지를 보면 전체 문자열 중에서 '별'과 '시'는 서로 분리되어 쉽게 분할 될 수 있음을 알 수 있다. 이것은 수직투영 방법을 적용한 <그림 3>의 결과에서도 알 수 있다. 그런데 문제는 '시'에서 'ㅅ'과 'ㅣ'은 떨어진 별개의 문자인 것처럼 나타나는데 이는 평균 문자폭의 기준 이하이므로 연결된 문자로 판단되어야 하며 바로 이전의 문자 블록과 결합하여 온전한 하나의 문자가 되도록 한다.



<그림 2> (a)원본, (b) 이진화, (c) 세션화 이미지



<그림 3> <그림 2(b)>에 수직투영 한 결과

그러나 문자 폭에 크게 의존하는 수직투영 방법을 문자 폭의 변화가 큰 무제약 필기체에 직접 적용하기는 어렵고, 필기체 문자열에서 문자들이 흔히 중첩되거나 접촉되기 때문에 직선화 된 분할경로로는 이들 문자들을 정확하게 분할 할 수가 없다.

결국 전체 문자열에서 중첩되거나 접촉되어 수직투영을 통한 직선 분할경로를 생성할 수 없는 '서울특'에 대해서만 두 번째 분할과정에서 더 세밀하게 다루어질 필요가 있다. 이것은 <그림 4>에 표현되어 있다.



<그림 4> 분할대상 선택

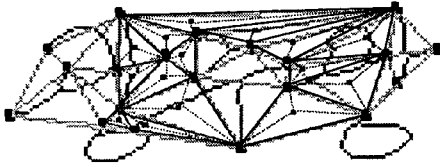
3. 꺾은선 분할경로

3.1 가상 분할경로의 생성

가상의 분할경로를 생성하기 위하여 먼저, 문자열에서 각 문자에 대하여 특징점을 고려하였다. 꺾은선 형태의 분할경로는 문자와 문자사이의 여백을 이용하여 생성되어야 한다. 이를 위해 각 문자의 끝부분이 가상 분할경로 생성 과정의 목적지, 즉 네트워크로 비유할 때 각각의 노드가 되도록 하기 위해 특징점으로 고려되었다. 또한 각 문자의 끝부분에 해당되면서 접촉된 부분에 대해서도 마찬가지로 특징점으로 고려하였다.

다음으로 고려된 특징점을 서로 연결해서 다양한 가상의 분할경로를 생성하기 위하여 몇 가지 방법들이 있지만 본 논문에서는 인접한 3개의 특징점들을 가장 효과적으로 그룹화할 수 있다고 알려진 델라우닝(Delaunay) 방법을 사용하였다. 이 방법에 의해 특징점들은 서로 연결되어 삼각형이 생성되는데, 즉 노드를 연결하는 링크가 만들어지고 이것이 가상 경로가 생성된다.

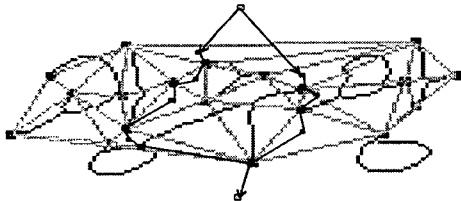
마지막으로 3개의 특징점으로 그룹화 된 삼각형 각각에 대해서 더 세밀한 가상경로를 생성하기 위하여 무게 중심점을 고려하는데, 이는 더 정확한 꺾은선 형태의 분할 결과를 얻기 위해서이다. 이것은 <그림 5>에서 볼 수 있다.



<그림 5> 특징점과 가상경로 생성

3.2 최상 분할경로 탐색

최상 분할경로를 탐색하기 위하여 먼저 <그림 5>에 나타난 가상 경로 중에서 문자 이미지의 foreground를 지나가는 경로들은 분할 경로로는 불가능하므로 제외될 필요가 있다. 그리고 분할 경로를 생성하기 위하여 출발 되는 시작점과 종료되는 끝점을 고려하기 위해 분할 대상 문자열의 위와 아래에 각각의 중앙점을 시작점과 끝점으로 정하고 분할경로를 생성해 나가는 방법을 사용하였고 최상 분할경로에 대한 결과가 <그림 6>과 같다.



<그림 6> 최상 분할경로

4. 결론

본 논문에서는 무제한 필기체 한글을 분할하기 위하여 가상 분할경로를 이용한 분할방법을 제안한다. 제안된 방법의 가장 큰 특징은 첫째, 많은 분할 방법들이 사용하는 직선화된 분할 경로 보다는 한글 구조적 특성에 의한 여백 공간을 활용하는 꺾은선 형태의 분할경로를 제시함으로써 특히 문자가 중첩된 경우에 대하여 신뢰도를 높이고자 하였다. 둘째, 분할과정을 크게 2단계로 구분하여 전체 문자열 중에서 보다 더 세밀한 분할이 요구되는 문자열만을 선택하였고, 선택된 문자열에 대해서도 최소한의 특징점만을 고려하면서도 다양한 가상의 분할경로를 탐색할 수 있도록 제

안함으로서 전체적인 처리속도를 높이고자 하였다.

현재 수집된 필기체 데이터에 대하여 비교실험이 진행 중이며, 특히 문자가 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 필기 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 Alhadj, "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.