필기체한글 글자단위 분할에 관한 연구

박아람, 조범준
조선대학교 컴퓨터공학과

A study of Character segmentation of Handwritten Hangul

Ah-ram Park, Beom-joon Cho
Dept. of Computer Science Engineering, Chosun Univ.

요 약

본 연구에서는 무제약으로 쓰여진 필기체 한글단어를 글자단위로 분할하는 새로운 방법을 제안한다. 이 방법은 글자와 글자사이 혹은 자소사이에 존재하는 배경(Background)정보를 세선화(Thinning) 처리하여 얻은 패스(Path)를 이용하여 글자와 글자사이를 지나는 패스를 결정하는 방법이다. 특히, 이 방법은 분할에 대한 판단을 인식기로 넘기지 않는 외적분할 방법으로 빠른 처리시간을 얻을 수 있고, 외적분할 방법의 단점인 정확도를 다른 외적분할 방법에 비해서 높일 수 있었다. 제안한 방법은 필기체 한글에서 많이 발생할 수 있는 중첩(Overlap)글자와 연결(Touched)글자를 분할하는데 효과적인 성능을 보였다. 중첩글자의 경우, 세선화에 의해 생성된 패스가 자연스럽게 중첩된 부분의 사이를 지나가면서 생성되기 때문에 매우 정확한 패스를 얻을 수 있었고, 연결 글자의 경우는 연결된 부분을 판단하고, 후보영역을 선정하여 연결된 부분을 분리해내는 방법을 사용하였다.

1. 서론

지난 수 십년에 걸친 연구자들의 노력으로 한글의 문자인식 기술은 많은 발전을 이루었다. 특히 인쇄체 문자인식의 경우는 실생활에 이용하는데 전혀 손색이 없는 인식률을 자랑한다. 반면, 필기체 문자인식은 아직까지 실생활에 응용하는데 무리가 있는 것이 사실이다. 필기체 문자인식은 크게 2가지 경우로 분류되는데, 온라인 필기체 인식과 오프라인 필기체 인식이 그것이다. 온라인 인식의 경우 펜의 압력, 획의 시간적 정보 등 유용한 정보를 많이 얻을 수 있지만, 오프라인 인식의 경우 입력된 글자의 공간적 정보만을 가지고 인식해야 하기 때문에 매우 어려운 실정이다. 이러한 이유 때문에 아직도 많은 연구자들은 오프라인 필기체 인식에 대한 연구를 하고 있는 것이다[1][2][3].

문자를 인식하는데 중요한 전처리 과정중의 하나가 바로 단어단위의 데이터에서 문자단위의 데이터로 분리를 해주는 문자분할(Character Segmentation) 과정이다[4]. 하지만 대부분의 문자인식 연구가 대상 글자들이 완벽하게 분할되었다는 가정하에 이루어지고 있어 문자분할에 대한 연구가 매우 미흡한 실정이다. 인쇄체 문자의 분할의 경우 인쇄된 글자의 일정한 모양, 크기, 간격 등 때문에 수월하게 진행할 수 있으나, 필기체 문자의 분할은 필자의 다양한 서체와 특성 때문에 어려움이 많이 있다. 또한, 지금까지 거의 대부분의 분할에 관련된 연구는 한글의 구성원리를 반영하지 못하고, 영문과 숫자에 중심을 두었기 때문에 특히, 한글 필기체 분할에 대한 연구는 미흡

했다. 이에 본 논문에서는 한글의 구성원리를 반영하여 필기체 글자를 분할하고자 한다.

본 논문에서는 글자자체의 특성이 아닌 글자의 외부에 구성되어 있는 배경(Background)의 특성을 이용하여 오프라인 필기체 한글의 글자를 외적으로 분할하는 방법에 대해서 제안한다. 먼저, 2 장에서는 분할과정에 대해서 설명하고, 3 장에서는 제안한 배경 세선화(Background thinning) 과정에 대해서 설명하고, 4 장에서는 실험 및 고찰을, 5 장에서는 결론에 대해서 논한다.

2. 한글에서 글자분할

지금까지 한글의 분할에 관한 연구의 대부분은 필기체 한글에서 매우 빈번하게 나타나는 접촉(Touched)이나 겹침(Overlapped)의 문제를 해결하지 못하고, 한글의 6형식(그림 1)에 따른 영역이 매우 일정하리라는 가정하에 인식과정에서 분할점을 찾는 방법을 주로 택해왔다. 그리고 각 위치에 따라 정해져 있는 자소의 후보(표 1)를 대상으로 인식을 행하여 왔다.

표 1. Groups of graphemes in Hangul

First Consonant(FC)	ㄱ,ㄴ,ㄷ,ㄹ,ㅁ,ㅂ,ㅅ,ㅇ,ㅈ,ㅊ, ㅋ,ㅌ,ㅍ,ㅎ,ㄲ,ㄸ,ㅃ,ㅆ,ㅉ
Vertical Vowel(VV)	ㅏ,ㅐ,ㅑ,ㅒ,ㅓ,ㅔ,ㅕ,ㅖ,ㅣ
Horizontal Vowel(HV)	ㅗ,ㅛ,ㅜ,ㅠ,ㅡ
Last Consonant(LC)	ㄱ,ㄴ,ㄷ,ㄹ,ㅁ,ㅂ,ㅅ,ㅇ,ㅈ,ㅊ, ㅋ,ㅌ,ㅍ,ㅎ,ㄲ,ㅆ,ㄳ,ㄵ,ㄶ,ㄺ, ㄻ,ㄼ,ㄽ,ㄾ,ㄿ,ㅀ,ㅄ

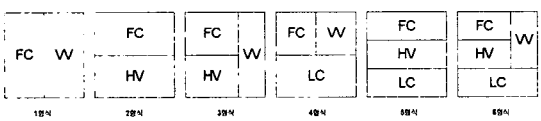


그림 1. 한글의 6형식의 구조

그러나 일반적으로 가정되는 것과는 달리 제약없이 쓰여진 한글의 경우 각 자소그룹이 위치를 보장받지 못하는 경우가 많다. 또한 주기적으로 나타나는 글자와 글자의 간격도 보장받지 못한다. 그래서 수직주사의 방법으로 강제분할 하는 것은 불완전한 인식단위를 생산하게 되고 때문에 인식기의 부담을 증가시키고, 또한 인식성능을 저하시키는 주요원인이 된다.

잘 쓰여지거나 인쇄된 단어라면 수직&수평주사로도 충분이 분할해 낼 수 있다. 그래서 일반적으로 먼저 수직주사로 분할 할 수 있는 글자는 분할을 한 다음 수직분할로 분할할 수 없는 곳, 바로 글자를 분할하는데 가장 어려움을 겪는 부분인 중첩(Overlap)되거나 연결(Touched)된 부분에 대한 처리를 하게 된다. 중첩된 부분은 (그림 2)와 같이 다른 글자의 영역으로 들어가 있는 형태이고, 연결된 부분이란 그림과 같이 두 글자가 서로 붙어 있는 형태이다.

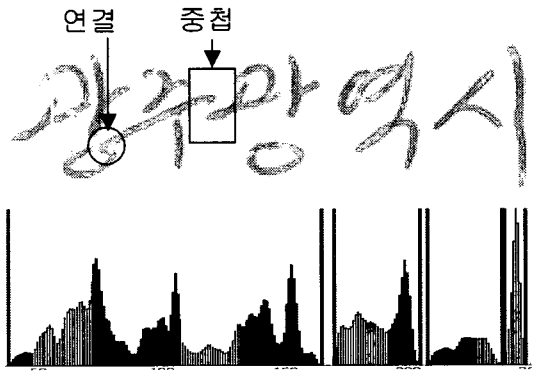


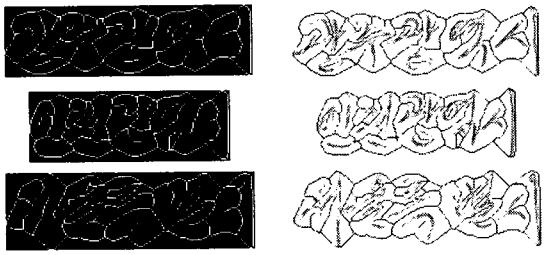
그림 2. 수직주사로 글자를 분할

위의 예에서 볼 수 있듯이 수직분할을 이용하여 글자를 분리를 시도하였으나 '광주광', '역', '시' 의 세 개의 블럭으로 나누어졌다. 만약에 '광주광'을 강제분할을 시도하였다면 인식결과에 많은 영향을 미치게 된다.

필기체 한글의 경우는 위와같이 수직주사를 사용하기에는 무리가 있다. 그래서 본 논문에서는 한글의 구조적인 특성을 이용하여 문자를 분할하려고 한다. 한글은 표음 문자로서 한 개의 글자가 하나의 음으로 가지고 있어서 보통 글자와 글자사이에 빈 여백 공간이 존재하게 된다. 본 논문에서는 이 여백을 이용하여 글자를 분할하는 방법을 제안하고자 한다.

3. 배경 세선화를 이용한 글자분할

본 논문에서는 한글의 글자 사이에 존재하는 백그라운드의 공간을 글자를 분할하는데 이용하고자 지금까지 흔히 인식과정에서 글자에 적용하였던 세선화를 글자의 백그라운드에 적용하여 백그라운드의 정보를 표현하는 배경 세선화 방법을 제안한다. 배경 세선화의 결과는 (그림 3)에서 확인할 수 있듯이, 글자를 지나지 않고 글자와 글자사이의 공간을 연결하고 있어 명확하게 글자를 분할할 수 있는 방법을 제공한다.



(a) (b)

그림 3. 배경을 세선화한 이미지. (a) 배경세선화의 결과 (b) 원본과 배경세선화를 결합한 이미지

3.1 일반적인 글자의 분할

그림 2에서 보듯이 수직 분할은 일반적으로 잘 분리되어 있는 글자를 분할하는데 효과적이다. 하지만 수직 분할은 한글의 구조적인 특징을 반영하지 못하고 에러를 만들 소지가 많다. (그림 4)에서와 같이 수직 분할로 분할된 이미지는 원래 한글자가 되어야 하는데 단순히 글자의 폭만을 이용하여 분할하다 보니 에러가 만들어진 것이다. 이를 해결하고자 본 논문에서는 일반적으로 잘 분리가 된 글자에 대해서도 배경 세선화를 그대로 적용하여 한글의 구성 규칙 (그림 1)을 이용하여 글자를 분리하였다.

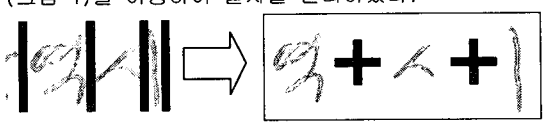


그림 4. 수직주사에 의한 분할 에러

먼저 수직 주사로 수직 분할 후보영역을 선택한 다음, 배경 세선화의 결과의 각각의 폐쇄된 구간영역의 위치를 한글의 6형식의 구성 규칙에 비교하여 형식을 얻게 되면 위와 같은 오류를 수정할 수 있다. 그림 5 에서 그 실행 과정을 묘사하였다.

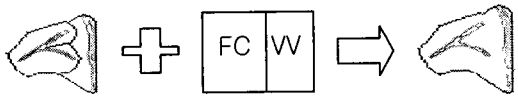


그림 5. 일반적인 경우의 분할

3.2 중첩된 글자의 분할

중첩된 글자의 경우, 글자와 글자 사이 중첩된 부분이라도 중간에 존재하는 배경의 공간이 있기 마련이다. 이 공간으로 경로를 생성해야 하는데 바로 이 배경 세선화 방법은 글자와 글자의 중앙을 지나는 경로를 생성해준다.

먼저 분할 예상 구역을 수직 주사로부터 선정한 다음 배경 세선화의 결과를 따라 경로를 결정하게 되면 아래와 같은 경로를 얻을 수 있다

그림 6. 중첩된 부분의 경로생성

3.3 연결된 글자의 분할

연결된 글자의 분할은 아직까지 많은 연구가 이루어지고 있지만 높은 신뢰도를 갖는 방법은 없는 실정이다. 또한 대부분의 방법들은 수많은 후보영역을 만들어 인식단계로 결정을 넘기는 형식을 취하고 있어서 인식기의 부담을 증가시키고, 결국은 인식 성능의 저하를 유발한다. 본 논문에서는 외적 분할 방법으로 이러한 연결된 부분을 검색하고 찾아내어 분리하는 방법까지 제안하였다.

연결된 글자의 경우, 배경 세선화로 생성된 구분된 영역중에서 연결된 부분이 있는 영역을 먼저 선택하게 된다. 이는 일반글자의 분할때와 마찬가지로 먼저 수직 주사로부터 분할 후보 영역을 선정한 다음, 각 영역의 폐쇠된 영역을 이용하여 형식을 결정하게 된다. 이때 다른 영역들과는 달리 연결된 글자의 영역의 경우는 두 영역 모두에 걸쳐 단일 영역이 형성되어 있고(그림 7(a)), 또한 영역의 길이도 글자의 평

균 폭보다 크게 된다. 이때 글자의 평균 폭은 수집된 데이터에서 구한 값으로 본 논문에서는 25 픽셀로 하였다. 이렇게 연결부분 후보영역을 검출하게 되면 이 영역내에 있는 끝점을 검출하여 그 중에서 분할 후보지역(아래 그림 8 에서 진하게 칠해진 부분)내에 존재하는 끝점과 만나는 글자부분을 분리점으로 선정하였다. 분리점이 선정되면 박정선[5]외 2인이 제안한 필기 한글 문자의 모양 분해 방법을 응용하여 연결 영역을 분리해 낼 수 있는 경로를 생성하였다. 이 알고리즘에서는 한글 패턴을 T-접점과 B-접점이라는 두 가지 모양 특징을 중심으로 분할할 수 있다는 관찰에 근거하여 꺽인점이나 접점의 분리면을 제시하였다. 아래의 그림 7(b)는 위의 알고리즘을 이용한 분리 후보영역이다. 본 논문에서는 연결 후보영역에 있는 글자영역에만 위의 알고리즘을 적용하여 분리 후보영역을 결정하였다.

(a) (b)

그림 7. 연결되어 있는 영역 추출 및 분할점 예시.(a) 연결영역추출 (b) 분할점 예시

분리점 영역에 있는 분리면을 통과하면서 새로운 패스를 생성하게 된다. 그림 8은 이 과정을 보여주고 있다. 글자영역을 통과하여 맞은편 지점으로 새로운 경계영역을 생성한 다음 중첩된 글자의 경로와 같은 방법으로 분할 경로를 설정한다.

그림 8. 연결된 부분의 분할경로 생성

4. 실험 및 고찰

본 논문에서는 150명으로부터 각각 40개의 주소관련 한글단어를 수집하여 데이터베이스를 구성하여 실험하였다. 각 단어들은 2~5자 사이로 구성이 되어 있고 중간에 기호나 영문, 숫자가 없는 순 한글 단어로만 되어있다. 글자를 쓰는 영역은 주어졌지만 자유스럽게 쓸 수 있도록 유도하였으며, 자연스러운 글씨체로 쓰도록 하였다.

5. 결론

본 연구에서는 필기체 한글의 글자 단위 분할에 대해서 배경 세선화라는 새로운 방법을 제안하였다. 이 방법은 한글의 구조적인 특징과 글자들 사이에 있는 배경의 정보를 이용하여 글자와 글자의 중첩된 곳에서 효과적으로 경로를 생성할 수 있었고 글자들이 서로 연결되어 있는 부분에서도 효과적으로 분리 지점을 추출해 낼 수 있었다.

제안된 방법은 인식기의 도움없이 외적분할만을 이용하여 효과적으로 글자를 분리했다는 것과 실행 속도가 빠르다는데에 의의를 갖는다.

본 연구는 한글 필기체 인식기의 전처리 부분에 결합될 수 있고, 특히, 배경 세선화는 한글의 자소 단위의 분할에도 사용될 수 있을 것으로 판단된다.

[참고문헌]

[1] Luiz S. Oliveira, R. Sabourin, "Automatic Recognition of Handwritten Numerical Strings : A Recognition and Verification Strategy", IEEE PAMI Vol. 24, No. 11, pp.1438-1454, 2002.

[2] Jaehwa Park, "An Adaptive Approach to Offline Handwritten Word Recognition", IEEE PAMI, Vol. 24, No. 7. pp. 920-931. 2002.

[3] S. Zho, Z. Chi, P. Shi, H. Yan, "Two-stage segmentation of unconstrained handwritten Chinese characters", Pattern Recognition, Vol. 36, No. 1, pp. 145-156, 2003.

[4] Sung-Bae cho, Jin H. Kim, "A Hierarchical Organization of Neural Networks for Printed Hangul Character Recognition", Korea Information Science Journal, Vol.17, No.3, pp.306-316, 1990.

[5] Jeong-Sun Park, Ki-Chun Hong, Il-Seok Oh, "Shape Decomposition of Handwritten hangul Characters", Korea Information Science Journal, Vol. 28, no. 7, pp. 511-523, 2001.