배경 세선화를 이용한 한글 필기체 글자 단위 분할

서원택°, 조범준
조선대학교 컴퓨터공학과
wontagi@ai.chosun.ac.kr, bjcho@chosun.ac.kr

Handwritten Korean Character Segmentation using Background thinning

Wontaek Seo°, Beom-joon Cho
Dept. of Computer science engineering, Chosun university

요 약

본 연구에서는 필기체 한글의 글자단위의 분할을 위해 배경 세선화(Background thinning)라는 방법을 제안한다. 배경 세선화 방법은 글자와 글자 사이에 존재하는 배경의 정보를 세선화 처리하여 필기체 한글에서 많이 발생할 수 있는 중첩(Overlap)글자와 연결(Touched)글자를 서로 분할하는데 효과적인 성능을 보였다. 배경 세선화를 이용하여 글자를 분할하는 방법은 인식과정의 판단을 필요하지 않은 외적분할 방법으로 빠른 속도의 분할 성능을 보였다. 이 방법은 특히, 중첩된 글자의 분할에 탁월한 성능을 보였을 뿐만 아니라, 연결된 글자에 대해서도 좋은 성능을 보였다.

1. 서론

일상생활에서 비교적 많이 사용할 수 있는 패턴인식의 기술중의 하나는 바로 문자인식 기술이라고 할 수 있을 것이다. 지난 수년간에 걸친 연구자들의 노력으로 문자인식의 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. 이러한 이유 때문에 아직도 많은 연구자들은 오프라인 필기체 인식에 대한 연구를 하고 있는 것이다.

문자를 인식하는데 중요한 전처리 과정중의 하나가 바로 단어단위의 데이터에서 문자단위의 데이터로 분리를 해주는 분할(Segmentation) 과정이다[1][2]. 인쇄체 문자의 분할 과정은 인쇄된 글자의 한정된 모양, 크기, 간격 등 때문에 많은 어려움이 없이 진행할 수 있으나, 필기체 문자의 분할은 필자의 다양한 서체와 특성 때문에 어려움이 많이 있었다. 또한, 지금까지 거의 대부분의 분할에 관련된 연구는 영문과 숫자에 중심을 두었기 때문에 한글 필기체 분할에 대한 연구는 미흡했다. 이에 본 논문에서는 한글 필기체 글자의 분할 방법에 대해 연구하였다.

분할의 방법은 크게 두 가지 방법으로 분류해 볼 수 있다. 첫 번째는 외적 분할(External Segmentation) 방법으로 분할 과정에서 인식과정의 도움 없이 바로 신뢰성이 높은 분할결과를 제시하는 방법이다. 두 번째는 내적 분할(Internal Segmentation) 방법으로 일단 분할과정에서 몇 개의 후보를 선정하여 인식과정으로 판단을 보류하여 인식과정에서 최종분할을 결정하는 것이다. 첫 번째 방법은 계산에 대한 부담은 줄어들지만 신뢰성을 정확히 보장하기 힘들고, 두 번째 방법은 계산에 대한 부담은 많아지지만 신뢰성이 첫 번째 방법보다는 높다.

본 논문에서는 배경(Background)의 특성을 이용하여 오프라인 필기체 한글의 글자를 외적으로 분할하는 방

법에 대해서 제안한다. 먼저, 2 장에서는 분할 과정에 대해서 설명하고, 3 장에서는 제안한 배경 세선화 (Background thinning) 과정에 대해서 설명하고, 4 장에서는 실험 및 고찰을, 5 장에서는 결론에 대해서 논한다.

2. 분할과정

다음은 필기체 글자를 분할하는데 가장 흔하게 사용되는 요소들이다.

- 글자의 넓이(Character width)
- 글자의 높이(Character height)
- 글자의 간격(Character gap)
- 위쪽 베이스 라인(top base line)
- 아래쪽 베이스 라인(bottom base line)
- 어휘(Lexicon)
- 수직주사(Vertical projection) 등.

이중에서 가장 쉽고 간단하게 사용할 수 있는 것이 바로 수직주사(Vertical projection) 정도가 될 것이다. 잘 쓰여지거나 인쇄된 단어라면 수직주사로도 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 그림과 같이 다른 글자의 영역으로 돌어가 있는 형태이고, 연결된 부분이란 그림과 같이 두 글자가 서로 붙어 있는 형태이다.

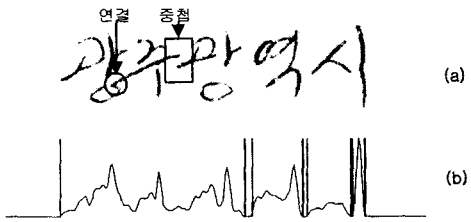

그림 1. (a) 필기체한글의 연결과 중첩의 예.(b) 수직주사로 분할위치 표시

위와 같은 중첩된 부분과 연결된 부분을 분할하는 방법을 제안하였다.

3. 배경 세선화

본 논문에서 제안된 배경 세선화 방법은 지금까지 흔히 인식과정에서 글자에 적용하였던 세선화 방법을 반대로 글자의 배경에 적용시켜 분할 경로를 생성하는 방법이다. 세선화 적용하는것은 Mophology thinning 기법을 이용하였다. 아래의 그림은 배경 세선화의 결과이다.

그림 2. 배경세선화 결과

그림 3. 배경 세선화 결과와 원본 이미지를 결합한 이미지

위의 그림에서 확인할 수 있듯이 글자의 영역을 침범하지 않고 배경의 영역을 분할해서 생성되었다.

3.1 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 선정한 다음 배경 세선화의 결과를 따라 경로를 결정하게 되면 아래와 같은 경로를 얻을 수 있다.

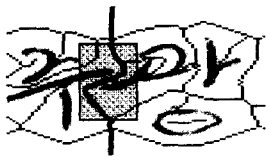

그림 4. 중첩된 글자 사이의 분할 경로

3.2 연결된 글자의 분할

연결된 글자의 경우, 배경 세선화로 생성된 구분된 영역중에서 연결된 부분이 있는 영역을 먼저 선택하게 된다. 이때 영역의 선택은 실험에서 사용한 임계값을 사용하였고, 실제 실험에서는 30으로 하였다. 이 영역 내에 있는 끝점을 검출하여 그중에서 분할 후보영역(아

래 그림에서 진하게 칠해진 부분)내에 존재하는 끝점을 분리점으로 선정하였다. 분리점이 선정되면 글자영역을 통과하여 맞은편 지점으로 새로운 경계영역을 생성한 다음 중첩된 글자의 경로와 같은 방법으로 분할 경로를 설정한다.

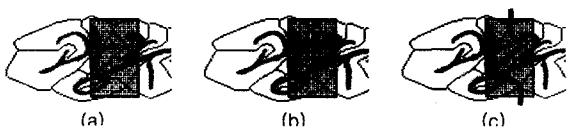

그림 5. 연결된 글자사이의 분할경로 (a)연결 분리점 추출 (b)분리경로 설정 (c)글자 분할경로 생성

4. 실험 및 고찰

본 논문에서는 100명으로부터 각각 40개의 주소관련 한글단어를 수집하여 데이터베이스를 구성하여 실험하였다. 각 단어들은 2~5자 사이로 구성이 되어 있고 중간에 기호나 영문, 숫자가 없는 순 한글 단어로만 되어 있다.

실험은 아래의 Flowchart 순서에 의해서 진행이 되었다.

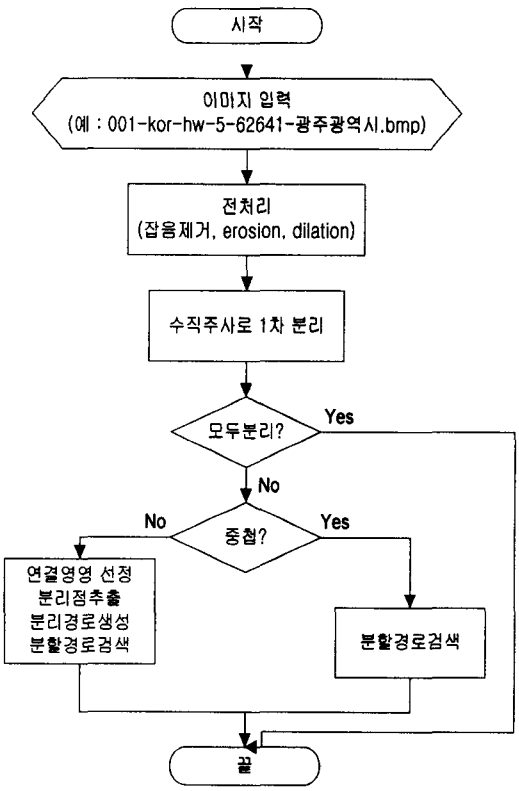

그림 6. 실험 순서도

5. 결론

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

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

[참고문헌]

[1] Yi Lu, "Machine-printed character segmentation", Pattern Recognition, Vol. 28, No. 1, pp.67-80,1995.

[2] Yi Lu, M. Shridhar, "Character segmentation in handwritten words-an overview", Pattern Recognition, Vol. 29, No. 1, pp.77-96,1996.

[3] 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.

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

[5] 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.