2001 년 한국정보처리학회 추계 학술발표논문집 제 8 권 제 2 호

획기반 필기한글 문자분할

김호연*, 김두식*, 남윤석*
*한국전자통신연구원
e-mail : hoyon@etri.re.kr

Handwritten Hangul Character Segmentation Based on Stroke Extraction

Ho-Yon Kim*, Doo-Sik Kim*, Yun-Seok Nam*
*Electronics and Telecommunications Research Institute

요 약

본 논문에서는 획기반 필기한글 문자분할 방법을 제안하고 이를 한글단어인식에 적용하였다. 제안된 방법에서는 획 단위의 문자분할을 시도함으로써 불필요한 분할점을 줄일 수 있었을 뿐 아니라 문자간 획의 접촉이나 겹침을 해결할 수 있었다. 실험에서는 이를 단어인식에 적용하여 비교적 높은 인식률을 얻음으로써 제안된 방법의 가능성을 입증하였다. 실험에서 이용한 문자인식기의 성능이 낮음에도 불구하고 비교적 높은 단어인식률을 얻을 수 있었던 것은 의미 있는 획 단위의 문자분할을 통해 불필요한 분할 가능성을 줄였고, 단어사전을 이용함으로써 사전정보를 충분히 활용할 수 있었기 때문이다.

1. 서론

문자인식은 패턴인식분야의 오랜 연구주제로서 많은 연구가 수행되었으며, 특히 필기문자인식은 최근까지도 꾸준히 연구되고 있다. 필기한글인식에 관한 연구도 발표되고 있으나 영문자나 숫자인식에 비해서는 논문 수가 매우 적을 뿐 아니라 주로 독립된 개별 문자인식에 관한 연구이어서 현실성이 다소 떨어졌던 것이 사실이다. 한글 개별문자 인식에 대한 연구결과를 실제 문제에 적용하기 어려운 이유는 연속적으로 필기된 단어나 문장에서 개별 문자를 추출해내기가 어렵기 때문이다. 문자인식 결과를 단어인식에 적용하기 위해서는 단어를 문자로 분할해야 하는데 일반적으로 문자간의 접촉이 있는 단어 내에서 각각의 문자를 성공적으로 분할하는 문제는 인식문제 만큼의 난이도를 갖고 있다고 해도 과언이 아닐 만큼 어려운 문제이다.

문자분할문제를 피하기 위해서는 직접적인 문자분할을 시도하지 않고 단어를 매칭하면서 문자분할 결과를 동시에 얻을 수 있도록 하는 사전기반 단어인식 방법을 이용하면 된다. 최근 발표된 한글 단어인식에 관한 연구로서 자모결합유형을 이용한 필기한글단어인식에 관한 연구[1]와 최적분할조합을 통한 사전기반 필기한글인식방법[2]은 모두 여기에 속한다. 첫번째 방법은 한글의 자음 및 모음 모델과 자모의 결합유형 모델을 이용하여 단어 내의 자모를 순서대로 정합하면서 단어를 인식하는 방법으로 인식 속도는 느리지만 39 단어사전에서 98.54%라는 비교적 높은 인식률을 보였다. 이 방법은 자모인식기와 단어인식기가 밀접하게 연관되어 있으므로 다른 방식으로 개발된 한글인식기를 이용할 수 없고 사전이 커질 경우 시간이 많이 걸린다는 단점이 있다. 두 번째 방법은 한글단어 영상을 일정간격으로 수직 분할하고 이를 조합하여 다양한 문자분할조합을 생성한 후에 각각을 인식하여 주어진 사전 내에서 가장 높은 매칭확률을 갖는 단어를 선택하는 방법으로 313 단어사전을 이용했을 때 91.96%라는 높은 인식률을 보였다. 이 방법은 단어 내에서 이웃한 문자가 수직선상에 겹쳐있을 경우 자연스러운 분할이 이루어지지 못하기 때문에 인식기가 이를 인식하지 못하면 단어인식률이 떨어진다는 것과 단어를 잘게 분할하면 인식해야 할 문자의 수가 늘어나기 때문에 인식시간이 많이 걸린다는 단점이 있다. 두 방법 모두 인식 이전의 문자분할을 피하고 문자의 분할과 인식을 동시에 처리하도록 함으로써 문자분할 오류를 줄이는 방법이다.

그러나, 문자분할이 인식 없이 이루어지기는 어렵더라도 모든 분할 가능성을 고려하기 보다는 분할 가능성이 높은 문자분할 후보를 미리 만들어 냄으로써 인식기의 부담을 줄이는 것은 필요하다. 특히 필기한글단어의 경우 이웃 문자와 수직 방향으로 겹치는 경우가 간혹 발생하기 때문에 이러한 문제를 해결할 수 있는 분할기를 개발한다면 기존에 연구된 한글인식기를 단어인식에 보다 쉽게 활용할 수 있을 것이다.

본 논문에서는 이러한 관점에서 개발한 필기한글 문자분할방법을 소개하고자 한다. 또한 기존에 개발된 필기한글 인식기와 문자분할기를 실제 단어인식에 적용함으로써 제안된 방법론의 유용성을 보일 것이다.

2. 필기한글문자분할

2.1 필기한글의 접촉 유형

필기한글은 그 기울어짐이나 획의 삐침 등으로 이웃 문자와 겹치는 경우가 생긴다. 필기한글의 문자간 접촉 유형은 크게 아래와 같이 구분할 수 있다.

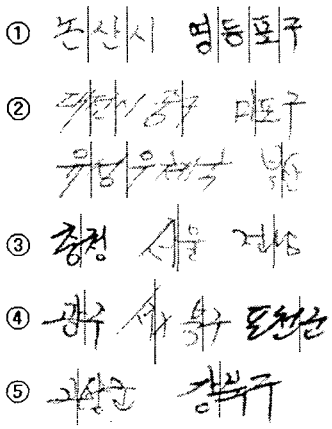

그림 1: 이웃 문자간의 접촉 유형

1) 수직 방향으로 겹치지 않는 경우
2) 수직 방향에는 겹치나 접촉되어 있지 않은 경우
3) 접촉되었으나 수직 분할이 가능한 경우
4) 접촉되었고 곡선분할이 가능한 경우
5) 교차(cross)되어서 곡선분할도 불가능한 경우

그림 1 에서 알 수 있듯이 단순한 투영 방법으로는 1) 의 경우만 해결 가능하며, 수직선에 의한 분할을 적용할 경우 1)과 3)의 경우만 해결 가능하다. 연결요소분석의 방법을 이용하면 1)과 2)는 분할 가능하나 3)과 4)는 불가능하다. 따라서, 1) 2) 3)을 모두 해결하기 위해서는 수직 투영, 접촉점 분할, 연결요소 분석과 같은 방법들을 복합적으로 적용해야 할 것이다. 그러나, 4)와 같이 접촉되었고 수직분할도 불가능한 경우를 해결하기 위해서는 연결모양 분석을 통한 분할점 선택이 필요하다. 본 논문에서는 위의 네 가지 경우를 분할할 수 있는 알고리즘을 개발하는 것을 목표로 하며 5)와 같은 경우는 4)와 동일한 경우로 간주하고 처리한다.

2.2 획기반 한글문자분할

문자분할 알고리즘은 화소들을 X 축에 투영하는 간단한 방법에서부터 문자간 접촉면을 분석하는 복잡한 방법에 이르기까지 다양하다. 본 논문에서 사용한 필기한글문자분할 방법은 획기반 분할 방법이다. 획을 추출하기 위한 여러 방법이 있을 수 있으나 여기서는 세선화를 이용하였다. 획 추출에는 런길이를 이용한 방법과 같이 처리시간 면에서 유리한 다른 방법이 있지만 획 추출의 편이성을 위해 세선화를 사용하였다.

문자 영상에서 문자를 이루는 기본 단위는 펜의 궤적을 통하여 생성된 획이다. 두 획의 끝이 동일한 각도로 자연스럽게 연결된 경우를 제외하면, 문자를 이루는 화소의 대부분은 문자 별로 분할 가능한 획으로 그룹핑될 수 있다. 대부분의 접촉된 문자의 분할점은 획의 끝점이나 접촉점에서 찾을 수 있다. 그러므로, 획 단위로 분할점을 검색하면 매우 효과적으로 분할점을 찾을 수 있게 된다. 그림 2 에서 볼 수 있듯이 세선화 후에는 획의 끝점이나 분할점을 쉽게 찾을 수 있다. 문자의 분할은 이러한 분할점을 이용하여 획들을 그룹핑함으로써 이루어진다.

그림 2: 세선화를 이용한 문자 분할점 선택 방법

획을 그룹핑하는 방법으로 상향식과 하향식 접근법이 있다. 하향식은 접촉된 획을 분할해가면서 분할점을 찾는 것이고 상향식은 모든 획을 따로 떼어 두고 획을 모아가며 그룹핑하는 방법이다. 하향식 방법은 나누어지는 획들의 다양한 경우를 고려해야 하기 때문에 우리는 상향식 방법을 이용하였다. 즉 획들을 합쳐가면서 문자 후보를 생성하는 방식으로 접근하였다. 문자분할 과정을 간략히 설명하면 다음과 같다.

문자분할과정:

1. 세선화 및 획 추출
2. 획 정렬: 추출된 획들의 X 좌표 중점에 따라 순서대로 정렬
3. 분할점 탐색: 획들을 비교하여 하나의 문자로 합쳐질 가능성이 매우 높은 것들을 그룹핑하고 분할점 선정
4. 문자분할 후보생성: 획 그룹들을 모아가며 문자 후보로 적합한 그룹을 선택
5. 문자분할경로 그래프 생성

위의 간략한 알고리즘에서 알 수 있듯이 문자분할 결과는 하나로 결정되는 것이 아니라 다양한 분할경로를 표현할 수 있는 그래프 형태로 나타난다. 여기서 주목할 것은 문자 분할점 탐색과 문자분할 후보 생성

은 다르다는 것이다. 분할알고리즘의 성능은 분할 점을 찾아서 접촉을 분리하는 능력과 불필요한 분할경로를 제거하여 후보 분할경로 수를 줄이는 능력에 달려있다. 각 단계별로 자세히 설명하면 다음과 같다.

1) 세선화 및 획 추출

세선화는 기존에 잘 알려진 방법 중 하나를 사용하였다. 획 추출에는 끝점과 분기점을 시작으로 하여 이웃한 점을 연속적으로 추출하는 방법을 사용했다.

2) 획 정렬

문자 분할점 탐색의 기본 단위로 획을 사용한다. 이 때 임의의 획 조합에 대해 분할 가능성을 고려하면 복잡도가 매우 높아지기 때문에 정렬된 획을 기준으로 분할 가능성을 점검하였다. 이를 위해서 먼저 획을 X 좌표 중점 순서대로 정렬해야 한다. 그림 4 는 추출된 획의 X 좌표를 이용하여 순서대로 번호를 부여한 예이다.

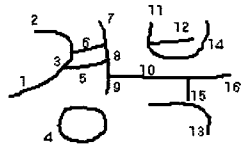

그림 4: X 좌표 중점을 기준으로 정렬된 획의 예

3) 분할점 탐색 (불필요한 분할점 제거)

문자분할을 시도하기에 앞서 문자의 크기(높이)를 추정한다. 추정된 문자의 높이는 분할된 문자의 폭을 예측하는데 이용된다. 일반적으로 문자의 높이는 문자열의 높이보다 작은데, 이는 문자열이 기울어져 있거나 문자의 크기가 일정치 않기 때문이다.

분할점 탐색에 주로 이용되는 정보는 획의 외접 사각형의 위치 정보와 접촉점의 수 이다. 접촉점의 수는 정렬된 획 상의 특정 획에서 분할될 경우 분할된 양쪽 획 그룹들 간에 접촉된 점의 수를 말한다. 예를 들면 그림 4 에서 3 획 이전과 4 획 이후 획간의 접촉점의 수는 2 가 되고, 7 획 이전과 8 획 이후 획간의 접촉점의 수도 2, 그리고 9 획 이전과 10 획 이후 획간의 접촉점의 수는 1 이 된다. 획 그룹들의 외접 사각형 정보와 그룹간의 접촉점 정보를 이용하여 분할점 가능성 지수를 계산하게 되는데 이는 경험적으로 다음 표와 같이 정하였다.

조건	가능성지수
c == 0, and d > 0	1.0
c == 0, and d <= 0 and d > -h/4	0.8
c == 0, and d <= -h/4 and d > -h/2	0.2
c == 1, and d == 0	0.7
c == 1, and d <= 0 and d > -h/4	0.5
그 외의 경우	0.1

h: 추정된 문자의 높이

d: 현재 획 그룹과 다음 획 그룹과의 거리 (다음 그룹의 최소 X 값 – 현재 그룹의 최대 X 값)

c: 접촉점의 수

표 1. 접촉 조건에 따른 획 분할 가능성지수

이 외에도 획 조합에 대한 경우의 수를 줄이기 위해서 매우 짧은 획은 이어진 획과 하나의 그룹으로 만들고, 세로방향의 획들을 하나로 그룹핑하는 등의 과정을 거치게 된다. 그림 4 를 예로 들면 7, 8, 9 번의 획은 같은 방향으로 연결된 세로획이므로 하나로 합쳐지게 된다. 이와 같이 획 간의 연결되는 모양이나 획의 크기 등을 고려하여 분할 가능성을 재조정한다. 분할점 제거 조건을 열거하면 다음과 같다.

- 세로방향으로 유사한 각으로 붙은 획
- 가로방향으로 유사한 각으로 붙었을 경우 획의 길이가 짧고 반대쪽이 연결되지 않은 경우
- 획이 아주 짧은 경우 연결성이 큰 획과 결합
- 획이 붙었고 합 친 것의 크기가 아주 작은 경우 (높이가 문자높이의 1/8, 폭이 문자높이의 1/6)
- 획이 붙었을 때, 한쪽이 작고, 작은 쪽의 반대편이 떨어져 있는 경우
- 매우 짧은 획의 경우
- 양쪽을 합친 높이에 비해 한쪽이 매우 작고, 넓은 쪽을 기준으로 봤을 때 많이 겹쳐 있고, 한쪽이라도 떨어져 있는 경우(양쪽 다 따로따로 연결되어 있으면 합치지 않음)
- 큰 것을 기준으로 봤을 때 작은 것이 중간 부근에 있고 반대편이 떨어져 있는 경우
- 붙어있고, 합치면 높이가 많이 커지고, x 축으로도 많이 겹치는 경우

이 단계의 목적은 불필요한 분할점을 가능한 줄이되 꼭 있어야 할 분할점은 남겨두는 것이다. 이러한 원칙을 만족시킬 수 있는 범위에서 다양한 휴리스틱이 동원될 수 있다.

4) 문자분할 후보생성

문자분할 후보점이 선정되면 이를 이용하여 문자분할 후보를 생성하게 된다. 문자분할 후보는 분할점을 순차적으로 결합하면서 그 크기와 위치 등을 고려하여 문자분할 후보가 될 수 있는지 여부를 검증하며 생성한다. 이 때 분할점 통합은 생성될 문자의 폭이 추정된 문자의 높이에 비해 지나치게 크지 않을 때까지 반복한다. 문자분할 후보 생성과정에서 이용된 휴리스틱은 다음과 같다.

- 분할점 5 개 이상은 합치지 않음
- 분할점 2 개 이상부터는 멀리 떨어져 있으면 (글자 높이의 2/3 이상) 합치지 않음
- 3 개 이상에서는 추가되는 폭이 문자 높이의 1/2 을 넘으면 합치지 않음
- 4 개 이상에서는 추가되는 폭이 문자 높이의 1/4 을 넘으면 합치지 않음
- 폭이 너무 넓으면 합치지 않음 (가로가 기준 문자높이보다 크고, 가로 세로의 비율이 1.8 배 이상 가로로 넓을 때)
- 상대적으로 많이 떨어져 있고 합쳤을 때 크기

가 너무 넓어지면 합치지 않음.
- 가로획부터 시작하지 않도록 하기 위해서, 세로가 짧은 획(가로획)이 왼쪽에 있으면서 떨어진 경우 합치지 않음

5) 문자분할경로 그래프 생성

위의 과정을 거친 후에 남는 유효한 문자분할후보의 시작점과 끝점을 연결하여 문자분할경로 그래프를 생성한다.

3. 실험 결과

실험에는 실제 우편영상에서 추출한 100 개의 단어 샘플을 이용하였다. 문자분할기의 성능 평가는 문자분할경로 그래프 상에 있는 후보 중 성공한 문자의 수, 즉 분할 성공률과 전체 후보문자의 수인 생성비율로 나타낼 수 있다. 혹은 낱자 인식기와 연동하여 단어인식에 적용했을 때의 결과를 통해서 분할기의 성능 및 유용성을 확인할 수 있다. 본 논문에서는 우선 기존의 낱자 인식기와 연동하여 단어인식에 활용했을 때의 실험결과를 통하여 논문에서 제안한 문자분할 방법의 가능성을 입증하고자 한다.

단어인식에 이용된 사전은 우편주소에 사용되는 289 개의 시/도/구/군 명칭을 이용하였으며 낱자 인식기는 신경망으로 구성된 인식기를 이용하였다. 본 논문의 목적이 단어인식기 성능에 관한 것이 아니므로 한글 1 순위 문자인식률이 47.6%로 비교적 낮지만 기존의 인식기를 그대로 활용하였다[3].

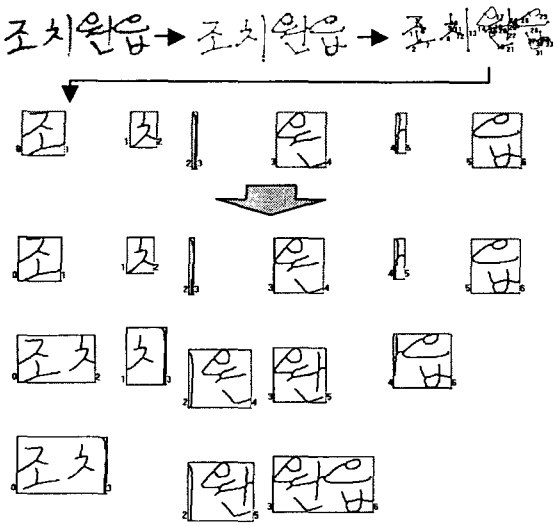

그림 5: 획추출 후 문자분할 후보 생성 과정

그림 5 는 단어영상으로부터 획을 추출한 후 분할점을 선택하고 이를 조합하여 문자분할후보를 생성하는 과정을 나타낸 것이다. 여기에서는 4 글자로 구성된 단어에 대해 5 개의 분할점이 선택되었고 14 개의 분할후보가 생성되었으며 그 중 옳게 분할된 문자가 모두 포함되어 있다.

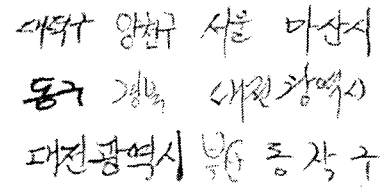

그림 6: 분할과 인식에 성공한 데이터

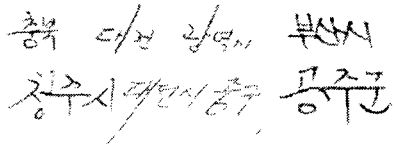

그림 7: 분할에 실패한 데이터

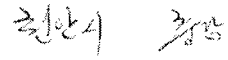

그림 8: 분할 후 오인식된 데이터

실험결과 문자분할과 단어인식을 모두 성공한 인식률이 73%이었다. 오인식된 데이터중 18%는 문자분할 오류 때문이었고 9%의 오류는 문자인식기의 오인식 때문이었다. 비록 인식률이 아주 높지는 않으나 문자인식기의 낮은 성능과 단어사전의 크기를 감안했을 때 상당히 높은 인식률이라고 생각된다.

4. 결론 및 향후 연구 내용

본 논문에서는 획기반 필기한글 문자분할 방법을 제안하고 이를 단어인식에 적용하여 비교적 높은 인식률을 얻음으로써 제안된 방법론의 타당성을 입증하였다. 제안된 방법에서는 획 단위의 문자분할을 시도함으로써 불필요한 분할점을 줄이고 접촉이나 겹침의 문제를 해결하고자 하였다. 특히 문자분할 그래프를 생성하고 이를 단어인식에 적용하여 비교적 높은 인식률을 얻음으로써 제안된 방법의 가능성을 입증하였다. 문자인식기의 성능이 낮음에도 불구하고 비교적 높은 단어인식률을 얻을 수 있었던 것은 의미 있는 획 단위의 문자분할을 통해 불필요한 분할 가능성을 줄일 수 있었고, 단어사전을 이용함으로써 사전정보를 충분히 활용할 수 있었기 때문이다.

참고문헌

[1] 진유호, 김호연, 김인중, 김진형, "자모 결합 유형을 이용한 적은 어휘에서의 필기 한글 단어 인식", 한국정보과학회지 논문지 : 소프트웨어 및 응용, 제 28 권 1 호 pp 52-63, 2001.

[2] S.H. Kim, S. Jeong, C.Y. Suen, "A lexicom-driven approach for optimal segment combination in off-line recognition of unconstrained handwritten Korean words," Pattern Recognition, Vol. 34, pp1437-1447, 2001.

[3] 김호연, 임길택, 남윤석, "필기한글 단어인식에서 사전정보의 효과", 대한전자공학회 학술대회, Vol.22, No.2, pp.1019-1022, 1999.