전산 원고지를 이용한 한글 문자 인식

하 지 용° 조 동 섭

이화여자대학교 전자계산학과

A Recognition of The Korean Character using the Presegmented Line

Ji-Yong Ha Dong-Sub Cho

Department of Computer Science, Ewha Womans University

< ABSTRACT >

This paper proposes the recognition of Korean character by using the given pre-segmented area. For higher processing speed, we introduce the techniques for pruning the unused segments. And, new thinning algorithm is used for finding skeleton of each segment.

1. 서론

한글 자료 처리가 기계화됨에 따라서 대두되어진 한글 자료 입출력에 관한 문제점을 해결하기위해 많은 연구가 계속되어져 왔다.

종래에는 인간과 computer 사이의 정보교환은 주로 keyboard나 단말기, 혹은 천공 card에 의한 입출력에 의존하였다. 근래에는 좀더 편리한 기능의 입출력 장치가 사용되고 있지만 아직까지는 주로 keyboard에 의한 자료입력이 많이 사용된다. 이는 자판의 문자 배열의 문제점, 모아쓰기를 한 자료를 직접 입력으로 사용할 수 없다는 제한점을 갖고 있다.

이와같은 문제점을 해결하고 모아쓴 한글을 직접 입력으로 사용하여 한글 자료 처리시 좀더 효율을 높이기 위해 전산원고지 사용에 대한 방법론을 제안한다.

전산원고지를 사용하는 경우 관습대로 모아쓰기를 한 한글자료를 입력으로 사용할 수 있으므로 인식 단계에 이르러 입력된 자료의 각 자소별 분리가 빠른 시간내에 이루어 진다는 장점을 가지고 있다. 본 논문에서는 각 자소별 인식후 다시 모아쓰기를 하여 출력하는 방식이 도입되었다.

2. 전산원고지 설계

2.1 기본선

한글은 사용자의 쓰는 습관에 따라서, 혹은 한글 문자와 유사성으로 인해 인식이 잘못되는 경우가 종종 발생하므로 인식의 정확성을 기하기 위해서 기본선을 두는것이 바람직하다. 이때, 전산원고지의 설계시 고려해야 될 점은 여러사람이 편리하게 사용할 수 있도록 각 영역의 치수가 결정되어져야 하며, 또한 전산원고지를 이용한 글씨체가 되도록이면 어색하지 않도록 설계되어져야 한다. 일반적으로 한글은 초성,중성,종성에 따라 그림 2.1와 같이 6가지 형태로 구별된다.

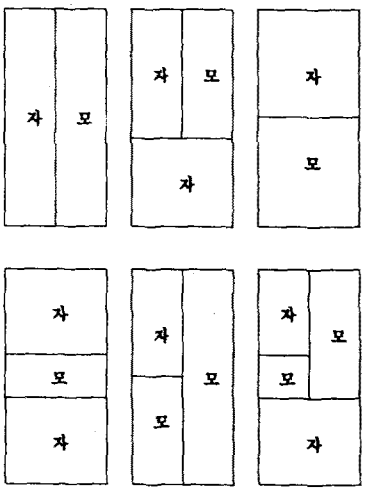

그림 2.1 한글의 여섯가지 구성형태

본 논문에서는 이 여섯가지 형태를 근거로 하여 모아쓰기를 한 한글을 그대로 입력으로 사용할 수 있도록 기본선을 제시하였다. 인식단계의 전처리 과정인segmentation시 좀더 효율을 높일 수 있도록 중성영역을 쌍자음과 복합자음을 위해, 횡모음 영역을 불필요한 부분의 탐색을 제거하기위해 세분화하였다. 그림 2.2에 그 기본선을 제시하였다.

위에서 제시한 기본선은 segmentation시 모든 영역이 아닌 1,2,4,6,7영역만이 사용되어지고 단지 3,5영역은 사용자의 편리를 도모하기 위해 사용되어진다.

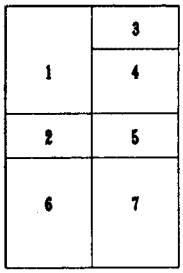

그림 2.2 기본선

2.2 전산원고지 영역의 치수

전산원고지 각 영역의 크기는 자모의 크기에 따라 결정된다. 한글 자모중 최대 폭은 'ㅛ'이며 최대 높이는 'ㅖ'이다. 즉 초성영역(1)의 최대폭은 'ㅛ'에의해서 결정되며 중성영역(2)는 'ㅛ'에의해서, 중성영역(4)는 'ㅖ'에의해서 결정된다. 다만 'ㅖ'의 위,아래 부분은 3,5,영역에 쓰이게 되고 4영역에는 모음의 주요정보가 기입되게된다. 중성영역(6,7)은 복합자음과 쌍자음을 위해 세분화하였으므로 초성영역과 차이가 나지 않는다. 각 영역의 폭과 높이의 비율은 아래와 같다.

초성영역(1) 5 : 4

중성영역(2) 5 : 2

중성영역(4) 4 : 2.5

중성영역(6) 5 : 4

중성영역(7) 4 : 4

그러므로, 전산원고지의 폭과 높이의 비율은 9:10이 되며 위에서 설계한 전산원고지와 사용예가 그림 2.3과 그림 2.4에 각각 소개되었다.

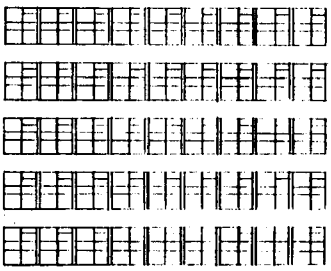

그림 2.3 전산원고지

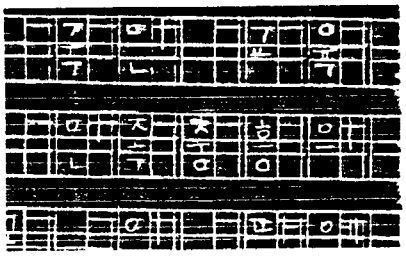

그림 2.4 전산원고지 사용예

3. 한글 인식의 전처리

3.1 segmentation

전산원고지를 사용하여 작성된 한글 문자를 인식하기위해 먼저, 입력한 문자를 각 자소별로 분리하는 과정이 필요하다. 전산원고지는 이미 초성,중성,종성 영역별로 규격화되어있기 때문에 원고지의 각 음절 크기에 맞게 window를 조정하여 빠른시간내에 구현할 수가 있다. 이미 앞에서 언급되었던바와 같이 초성,중성,종성의 중요정보는 1, 2, 4, 6, 7 영역에 위치하게 되므로 이 영역만의 segmentation이 필요하다.

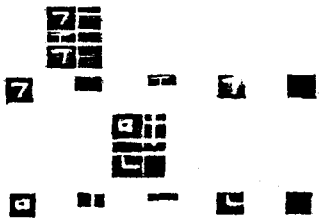

그림 3.1 segmentation 결과

3.2 세선화

한글 입력자료를 각 자소별로 segment한 후 그것들의 인식을 위해 세선화가 요구되어진다. 입력된 자료는 3 x 3 window를 사용하여 한 pixel씩 검색이 일어난다. 중심화소 P(X,Y)에 위치하는 점이 삭제 가능한가 아닌가는 근방화소들과의 연결성을 고려하여 결정한다.

P8 (X-1,Y-1)	P1 (X-1,Y)	P2 (X-1,Y+1)
P7 (X,Y-1)	P (X,Y)	P3 (X,Y+1)
P6 (X+1,Y-1)	P5 (X+1,Y)	P4 (X+1,Y+1)

그림 3.2 화소 P의 8근방화소

4근방화소 : P를 둘러싼 홀수방향에 있는 4개의 근방화소

P1, P3, P5, P7을 의미한다.

4근방화소의 갯수에 따라 화소제거 조건은 다음과 같다.

i) 4근방화소의 갯수가 1인 경우: '01' 패턴이 오직 한번 일어난 경우에만 제거 가능하다.

ii) 4근방화소의 갯수가 2인 경우: '01' 패턴이 한번 일어난 경우 모두 제거 가능하다.

iii) 4근방화소의 갯수가 3인 경우: 현재 window 상태가 그림 3.3인 경우일때는 window를 이동하여 한번씩더 체크하여야 한다. (a)의경우 window를 위로 한줄 옮기고 (b)의경우 왼쪽으로 한줄 옮겨서 이동한 window의 중심화소를 체크한다. 즉 P(X,Y)가 P'(X-1,Y), P'(X,Y-1)이 되도록 이동한 다음 그점이 이미 삭제된 점인가 확인을 한다. 만약 삭제된 점이라면 원래의 중심점 P(X,Y)는 제거되지 않는다. (c),(d)의경우에는 아래로 한줄, 오른쪽으로 한줄 이동하여, 즉 P(X,Y)가 P'(X+1,Y), P'(X,Y+1)이 되도록 이동하여 이동한 점이 앞으로 삭제가능한 점인가 체크를 한다. 만약 삭제 가능하다면 원래의 점 P(X,Y)는 제거되지않는다.

iv) 4근방화소의 갯수가 4인 경우: 이경우는 모두 삭제 불가능하다.

위의 알고리즘을 삭제 가능한점이 발견되지 않을때까지 적용하며 그 결과는 그림 3.4와 같다.

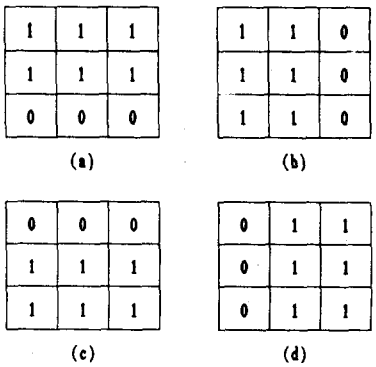

그림 3.3 서브루틴이 필요한 패턴

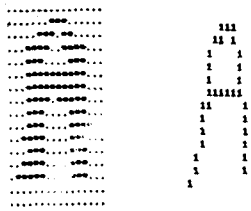

그림 3.4 세선화 결과

4. 결론

본 논문에서는 좀더 효율적으로 한글 자료를 처리하기위한 방법으로 전산원고지의 사용을 제안하였다. 앞으로 소개된 세선화 알고리즘을 한글 인식에 도입하고 또한 각 자소별로 인식한 결과를 다시 모아쓰기를 하여 음절단위로 출력하는 연구가 요구되어진다.

참고 문헌

1. 황종선,백두권,박도순, "광학식 문자인식을 위한 한글수서 문자의 설계", 정보과학회지, Vol.6, No.4, 1988. 8.
2. 황종선, "광학식 문자인식을 위한 한글 수서문자의 치수에 관한 연구", 공업진흥청, 1986. 10.
3. 강순모, "이진영상에 대한 세선화 알고리즘의 병렬처리에 관한 연구", 이화여자대학교 대학원 석사논문, 1988.
4. 한상기, "구조적 방식에 의한 한글인식", KAIST 석사학위 논문, 1984.
5. 이주근,남궁재찬,김영건, "한글 pattern에서 subpattern 분리와 인식에 관한연구",전자공학회지, Vol.18, No.3, 1981. 6.