한국통신학회 1985년도 추계학술발표회 논문집(1985. 11. 9)

한글 인식에서 자소 추출에 관한 연구

°최 병 만, 김 은 진, 김 정 선
한국항공대학 전자공학과

A Study on Algorithm of Phonemes Extraction in Korean Character Pattern Recognition

Choi, Byeong Man, Kim, Eun Jin, Kim, Jung Sun
Dept. of Electronics Eng., Hankuk Aviation College

ABSTRACT

This paper proposes a algorithm of phonemes extraction in Korean character pattern recognition. The phonemes are classified into the patterns which are separable and connected with each other. The former is extracted by means of pattern matching in consideration of topological structure of phonemes and direction of stroke sequentially. The latter is extracted by means of index and window algorithm which are performed by a 3x3 sequential local operation in the thinned character pattern.

I. 서 론

정보화 사회로의 변환이 급속히 이루어짐에따라 컴퓨터를 이용한 사무자동화(Office Automation:OA)에대한 연구가 활발히 진행되고 있다.

패턴인식은 컴퓨터에 문자 및 picture패턴을 자동입력시켜서 그것을 판독하는 인공지능(Artificial Intelligence:AI)에 관련된 연구의 일환이다. 초기에는 단순한 숫자 및 문자 패턴이 대상이었으나 점차로 지형과 일반 물체에 이르기까지 확장되어 최근에는 전술 병기의 분야까지 이용이 확대되고 있는 실정이다.

문자구조의 연구의 동향은 매우 많으나 대별하여 보면 결정론적 방법과 구문론적 방법의 두 주류를 이루고 있다. 결정론적 방법은 수식모형과 패턴에서 특징을 추출하여 각 패턴을 특징 벡터에 의하여 식별하는 것이 일반적이다. 이 방법은 패턴의 구조와 그들의 관계를 취급하는 통일된 형식이 없다는 것이 결정론적 방법의 결점으로 지적되고 있다. 구문론적 방법은 N. Chomsky의 수학적 모델에 의한 형식 언어 이론에 기초를 두고 있으며 패턴의 구성 요소인 성분이나 패턴 primitive에 의하여 패턴을 표현하고 언어의 syntax와 패턴 구조 사이의 유사성을 도입하여 주어진 syntax 법칙에 따라 패턴의 구조를 parsing하여 식별한다. 따라서 조직적인 면에서는 결정론적 방법의 결점을 보완했다고 볼 수 있다.

한글에 대한 인식 문제는 1969년 처음 시작되었고 최초로 조합문자에 대한 인식 연구를 시도하였다. 이 연구에서는 한글을 30종의 패턴으로 형식화하고 그것을 다시 6 내지 9종류로 변형하여 형태를 식별한 후 패턴 공간을 가변 분리하여 인식하는 방법에 관한 것이다. 인식율은 90%이고 1600여 자를 대상으로 기본 문자 식별로서 판독하는 방법이다. 또 구문론적 방법에 의하여 Tree 문법으로서 패턴그래프를 top-down적으로 순차 인식하는 방법도 발표되었다.

한글은 다른 언어의 문자와는 달리 문자수가 방대하고 처리 시간을 더욱 더 감소해야하는 문제점을 갖고 있다. 이러한 문제를 해결하는 하나의 방법으로서 본 연구에서는 한글 구조의 컴퓨터 입력시를 고려하고 인식율에 큰 영향을 미치고 있는 인식 전반부의 전처리과정과 한글의 구조상 자모가 서로 분리되어 있는 경우와 자모가 서로 연결되어 있는 경우를 서로 다른 알고리즘을 택하여 자소들을 추출해 내는 알고리즘을 제시한다.

II. 전처리 과정

전처리 과정은 문자 인식에 대한 인식율을 높이기위해 인식 알고리즘의 전반부에서 수행되는 과정이다. 문자 패턴은 ITV 카메라와 Image digitizer를 통해 32 × 32의 화소로 양자화되어 컴퓨터내부로 입력된다. 입력된 문자 패턴을 임으로 선정된 스레쉬홀드 레벨에 의하여 "1"과 "0"으로 2치화된다. "1"은 정보를 가지고 "0"은 정보를 가지지 않는다고 한다. 2치화된 패턴은 입력시 오차나 잡음을 흡수하기 위하여 평활화(Smoothing) 처리를 한다. 평활화 처리는 배경 부분의 불필요한 점을 제거하고 문자의 결함을 보충,또 문자의 요철을 제거하고 있다. 또 문자의 크기나 잡음의 종류를 조사하여 마스크 패턴의 크기를 조절함으로서 보다 좋은 인식 결과를 갖게한다. 이후에 stroke를 줄이기 위해 세선화 작업을 행한다.

III. 자소 추출

1] 한글의 표면 구조 식별

한글은 구조적으로 6 종류의 패턴으로 형식화 되었다. 이들 형식에서 횡모음과 종모음의 성분은 한글의 구조적 특징을 결정지워주므로 이의 정보는 6 종류의 형식화된 패턴의 포괄적인 구조로 판정할 수 있다. 우선 세선화된 패턴을 X,Y의 크기로 정규화하여 패턴 성분의 상대적 크기를 결정한다. 패턴 구조의 식별은 다음과 같이 행한다.

a) 횡모음 판정

최좌측에서 T를 시작점으로 하는 횡선분으로서 수색방향의 단점이 T이거나 B이면 그 길이가 X/2 이상일 때, 횡모음으로 판정하며 이 때의 횡선분을 Xm이라 한다.

b) 종모음 판정

우측 최상단에서 T를 시작점으로 하는 종선분으로서 수색방향의 단점이 T이거나 D또는 B이고 그 길이가 Y/2 이상일 때 종모음으로 판정하며, 이 때의 종선분을 Ym이라한다.

c) 종성 판정

Ym이 존재하고 Ym의 하단부에 B, D, L이 존재할 때 종성이 있음을 판정하고, Ym이 존재하지 않을 때 Xm의 하단부에 B, D, L이 있으면 종성이 있음을 판정한다.

2] 자소가 분리되어 있는 경우

한글은 구조상 절반 이상이 자소가 분리되어 있는 문자 패턴을 가지고 있다. 32 × 32 화소의 문자 패턴에서 자소가 분리된 경우 자소를 추출하기 위해서 pixel을 상하좌우로 이동시키고, 그 진행 방향을 고려하여 다음 단계로 행한다.

a) 문자 성분을 가지고 있는 기준 pixel에 3 × 3 window를 씌워 window내에 들어오는 pixel의 수를 계산한다.

b) 계산된 pixel의 수가 1개이면 이 pixel은 연결된 stroke상에 존재한다고 한다. 계산된 pixel의 수가 2개이면 이 pixel은 분기점을 이룬다고 한다. 계산된 pixel의 수가 0개이면 이 pixel은 stroke가 끝나는 경우라고 한다.

c) stroke가 시작되는 점을 시작점이라 하고 stroke가 갈리어 나가는 점을 분기점, stroke가 끝나는 점을 단점이라 한다.
진행하는 stroke의 방향과 다른 방향으로 진행하는 점을 굴곡점이라 한다.

d) 최좌측 단점을 점을 시작점으로 한다.

e) 세선화된 문자 패턴에서 시작점을 구한 후 3 × 3 window를 씌워가면서 이웃한 pixel로 이동하면서

중심에 위치하는 pixel의 값을 1 에서 2로 바꾸면서 진행한다.

f) e)에서 진행을 계속하다가 분기점을 만나면 window 주변에 걸치는 2개의 pixel중 하나의 pixel로 진행시키면서 그 위치를 기억한다. 계속해서 진행시키면서 pixel의 값을 1에서 2로 바꾼다.

g) 계속하여 진행하던 pixel이 단점에 도달하면 f)에서 기억된 분기점으로 이동한다.

h) 기억되어 있던 분기점에서도 앞에서와 같이 기준 pixel을 1에서 2로 바꾼다.

i) 더 이상 진행할 곳이 없으면, 그 점을 단점으로 한다.

j) pixel이 단점에 도달하면, 분기점에서 기억된 또 다른 방향으로 수색하여 기억된 분기점이 있으면 그 위치로 window를 옮겨서 위의 과정을 반복한다.

k) 더 이상 진행할 곳이 없으면 자소가 분리된 패턴의 pixel값이 모두 1에서 2로 바뀌게된다.

l) pixel값이 2인 부분을 선택적으로 추출하면 된다.

3] 자소가 연결 되어 있는 경우

자소가 서로 연결되어 있는 경우에는 문자의 모음을 중심으로 추출한다. 표1과 표2를 살펴보면 자모의 연결 상태를 알 수 있다. 초성과 중성, 중성과 종성의 연결 상태를 보면, 자소가 서로 연결되는 곳은 항상 분기점이 됨을 알 수 있다. 그러나 모든 분기점을 중심으로 자소를 추출할 수는 없다. 즉, 분기점에 연결되어 있는 자음의 상태를 먼저 관찰하여야 한다. 예를 들어 그림 1에 나타낸 "국"이라는 글자를 생각하자. 그림에 보이듯이 3개의 분기점을 중심으로 자소들을 추출하면 올바른 추출이라고 할 수 없다. 따라서 분기점을 중심으로 자소들을 추출하되 그 기준을 정해야 한다.

본 연구에서는 이러한 기준을 index mark(@)로 표시하고 그 정의와 알고리즘은 다음과 같다.

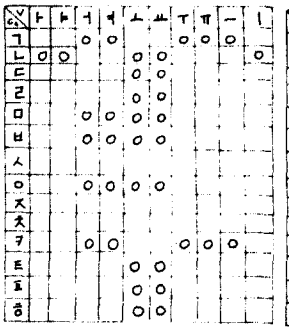

표 1

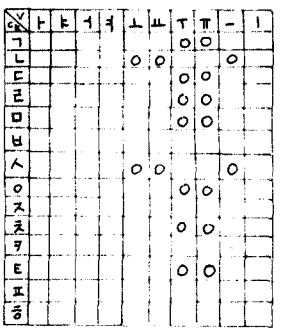

표 2

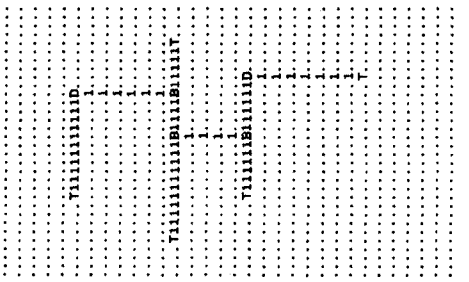

그림 1 세선화 패턴

<index mark의 정의>

pixel을 진행시키다가 분기점을 만나면 그 분기점을 중심으로 연결되는 자음의 상태를 고려하여 자소들을 떼어내야할 곳에 index mark(@)를 붙인다.

<index 알고리즘>

한글의 표면구조 식별에서 횡모음이나 종모음을 포함하고있는 문자의 횡모음과 종모음을 Xm과 Ym으로하였다. Xm, Ym의 시작점 T에서 진행하여 분기점을 만나면 그 점을 Bx_1, By_1 라하고 그 위치를 기억하고 진행 방향과 일치하는 방향으로 계속 진행하여 다른 분기점을 만나면 그 점을 Bx_2, By_2 이라하고 그 위치를 기억한다. 같은 방향으로 계속 진행하여 수색이 끝나면 그 점을 단점으로 한다. pixel은 다시 Bx_1, By_1 점으로 이동하여 분기점의 다른 한 방향으로 수색한다. 여기서 분기점, 굴곡점, 단점을 만나면 그 점에 @를 붙인다. 다시 pixel은 Bx_2, By_2 의 위치로 이동하여 그 분기점의 다른 한 방향으로 수색하여 분기점, 굴곡점, 단점을 만나면 그 점에 @를 붙인다.

종합하여 나타내면 다음과 같다.

a) Bx ⟶ T, T=ⓐ

b) Bx ⟶ D, D=ⓐ

c) Bx ⟶ B, B=ⓐ

d) Bx ⟶ T, T=ⓐ

e) By ⟶ D, D=ⓐ

f) By ⟶ B, B=ⓐ

g) Xm ⟶ By , By =ⓐ

h) Xm ⟶ T , T =ⓐ

i) Ym ⟶ T , T =ⓐ

index mark를 기준으로 하여 자소들을 추출하는데 초성, 중성, 종성을 windowing하여 다음과 같이 처리한다.

<window 알고리즘>

a) 초성

Xm 및 Ym에 ⓐ가 존재하면 ⓐ을 경계로 Xm의 상측 영역을 Ym의 좌측 영역을 떼어내고 Xm, Ym상에 ⓐ가 존재하지 않으면 Xm의 상측을, Ym의 좌측을 떼어낸다.

b) 중성

1> Xm은 존재하고하고 Ym은 존재하지 않을때 Xm에 연결된 짧은 선분에 ⓐ가 존재하면 ⓐ과 Xm을 중성으로 추출한다.

2> Ym은 존재하고 Xm은 존재하지 않을 때 Ym에 연결된 짧은 선분에 ⓐ가존재할 때 ⓐ과 Ym을 중성으로 추출한다.

3> Xm과 Ym이 존재할 때, Xm과 Ym에 연결되어 있는 짧은 선분에 ⓐ가 존재할때 ⓐ 및 Xm을 포함하는 ⓐ 및 Ym을 포함하는 중성으로 추출한다.

c) 종성

1> Xm이 존재하고 Xm의 하단에 ⓐ가 존재하면 ⓐ의 하단을 종성으로 추출한다.

2> Xm이 존재하고 ⓐ이 없으면, 횡모음의 하단에 B, D, L이 있으면 Xm의 하단을 종성으로 추출한다.

3> Ym만 존재할때 Ym의 하단의 ⓐ을 종성으로 추출한다.

IV. 결론

본 연구에서는 자소들이 분리되어 있는 경우와 연결되어 있는 경우에 대한 자소추출 알고리즘에 관한 것이다. 자소들이 정확하게 추출됨으로서 인식과정에서의 오인식을 줄일 수 있다고 본다.

a) 형식화된 6개의 formal 패턴의 표면 구조를 식별함으로서 자소의 추출 과정을 용이하게 할 수 있었다.

b) 자소가 분리되어 있는 경우는 pixel의 수색 방향과 구조적 특징을 이용하여 순차적으로 추출하였다.

c) 자소가 연결되어 있는 경우에는 index mark 에 의하여 자소를 떼어 낼 수있는 기준을 정하고 window를 이동하여 정확하게 추출할 수 있었다.

d) 문자 인식에서 세선화처리의 고속화로 막대한 양의 정보를 처리해야 하는 문제점이 발견되었으나 많은 연구와 전용 LSI의 개발로 이러한 문제는 해결될 수 있으리라 생각된다.

Reference

[1] 이 주근 ,"한글 문자의 인식에 관한 연구," 대한 전자공학회지 Vol. 9, No. 4, Sep. 1972

[2] 이 주근, 김 영근, 남궁 재찬, "Character pattern의 부분 분리와 인식에 관한 연구," 정보과학학술발표논문집, Arp. 1980

[3] 한글 기계화 연구소, "한글 기계화 연구," 1975

[4] J.K. Lee, "A method for the recognition of printed Korean chracters," JKIEE Vol. 7, No. 4, Dec. 1969

[5] T.K. Kim, T. Asui, "A study od the pattern recognition of Korean chracters by syntactic method," JKIEE Vol. 14, No. 5, Dec. 1977

[6] R. Stefanelli, A. Rosenfeld, "Some parallel thinning algorithm for digital pictures," J. ACM, Vol. 18, 1971

[7] H. Niemann, Pattern analysis, Springer-Verlag, 1981