세선화된 결과를 이용한 한글과 영어의 구별

전 일 수† · 원 남 식† · 이 두 한††

요 약

본 논문에서는 세선화된 결과로부터 다중 활자체에 적용 가능한 한글과 영어를 구별하는 알고리즘을 제안한다. 제안된 알고리즘에서는 각 글자의 연결요소의 개수를 가지고 한글과 영어를 구별하였으며, 연결요소의 개수만으로는 구별이 되지 않을 경우에는 한글에 존재하는 모음을 이용하여 이들을 구별하였다. 가장 널리 사용되는 세가지의 활자체에 대해 21,150 자를 실험한 결과 99.82%의 구별율을 얻었다.

Distinction of Korean and English Characters Using the Result of Thinning

Il Soo Jeon † · Nam Sik Won † · Doo Han Lee ††

ABSTRACT

This paper proposes a distinction algorithm of Korean and English characters which can be applied to multi-font from the results of thinning. The proposed algorithm distinguishes Korean and English characters as the number of connected components. If it can not distinguish those characters with the number of connected components, it distinguishes them as the vowel included in Korean characters. In experimental results, the distinction rate is about 99.82% for the 21,150 characters of three widely used fonts.

1. 서 론

최근 퍼스널 컴퓨터의 확산과 더불어 컴퓨터 하드웨어의 급속한 발전과 각종 유용한 소프트웨어의 개발은 고도 정보화 시대로의 변화를 가속시키고 있다. 고도 정보화를 달성하기 위해서는 문서 정보를 컴퓨터에 기록하는 것은 불가피하며, 이러한 방대한 양의 문서 정보를 처리하기 위해서는 문서인식 시스템을 필요로 한다.

현재 국내외에서는 문서인식을 위해 각 나라의 문자를 인식할 수 있는 단일 문자 인식 시스템[1, 2, 3, 4, 5, 6]이 많이 발표되고 있다. 그러나 우리가 접하는 문서의 대부분은 한글로만 쓰여지는 것이 아니라 한글과 영어 및 한자들이 혼용되어 쓰여진다. 그러므로 이들 혼용 문서의 처리에는 단일 문자 인식 시스템은 적합치 못하며 여러 나라 글자를 인식할 수 있는 문자 인식 시스템을 필요로 한다. 혼용 문서 인식을 위한 연구로서 한글과 한자로 쓰여진 혼용 문서를 인식하는 연구[7, 8]가 발표되었다. 혼용 문서 인식에 대한 연구는 단일 문자로 쓰여진 문서 인식과 비교하면 아직 연구가 많이 이루어지지 않고 있는 실정이다.

본 연구는 인쇄체 글자의 세선화된 결과로부터 한글과 영어를 구별하는 방법을 제안하였다. 제안된 방

† 정 회 원:경일대학교 전자계산학과
†† 정 회 원:경동전문대학 정보처리학과
논문접수:1996년 11월 22일, 심사완료:1997년 5월 22일

법에서는 각 글자의 연결요소와 한글의 모음 성분을 추출하여 한글과 영어를 구별하였다. 영어의 경우에는 각 글자의 연결 요소가 1개(소문자 i, j는 예외)인데 한글의 경우는 연결 요소가 1개 이상이 된다. 그래서 연결 요소가 1개 일 경우 한글에 존재하는 종모음과 횡모음의 존재 여부를 파악하여 한글과 영어를 구별하고, 연결 요소가 2개인 경우는 영어의 소문자 i, 혹은 j라고 판단되면 영어로, 그렇지 않으면 한글로 판정한다. 그리고 연결 요소가 3개 이상이면 무조건 한글로 판정한다.

문자인식에서 형태 분석을 쉽게 하고 또한 데이터 양을 최소화하기 위해 세선화를 한다. 세선화 과정은 문자인식을 위한 전처리 단계에서의 핵심적인 분야로 이에 대한 연구가 활발히 진행되어 왔으며 300 여 편 이상의 논문이 발표되었다[9]. 본 논문에서 세선화된 결과를 입력으로 하여 한글과 영어를 구별하는데 연결요소의 수를 가지고 판단을 하기 때문에 세선화된 결과에서는 연결성이 보장되어야 한다.

제안된 방법을 구현하여 실험해 본 결과, 입력 영상에 다소의 잡영이 존재하더라도 한글과 영어를 잘 구별하였으며, 한글과 영어가 50:50으로 작성된 문서에 대해 개개의 글자로 바르게 분리되었다고 가정했을 때 명조체의 경우는 99.96%, 신명조체의 경우는 99.87%, 고딕체의 경우는 99.53%가 구별되었다. 제안된 방법은 기존의 대부분 활자체에 적용될 수 있다.

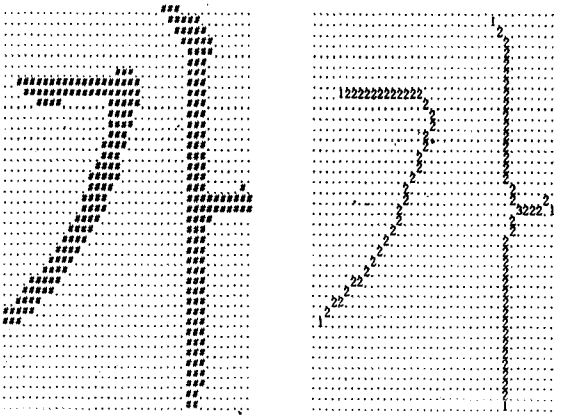
본 논문의 2장에서는 본 연구의 입력으로서의 세선화 알고리즘에 대해서, 3장에서는 한글과 영어의 구조적인 특징을 기술하였다. 4장에서는 본 연구에서 제안한 한글과 영어의 구별 방법을 기술하였고, 5장에서는 제안된 방법의 실험 결과 및 고찰을 기술하였다. 그리고 6장에서는 결론 및 향후 연구 방향을 제시하였다.

2. 본 연구 입력으로서의 세선화 결과

세선화는 인식 대상 영상을 양자화하여 영상을 구성하는 가장 기본 요소인 화소(Pixel)로 변환한 후, 특징 추출에 무관한 화소를 제거하여 골격선의 두께가 1인 화소로써 그 특징 골격을 형성하는 작업을 말한다. 세선화 알고리즘에 의해 추출된 골격선의 품질이 문자 인식율에 크게 영향을 미치게 되므로, 인식 대상 문자에 가장 적합한 세선화 알고리즘으로 수행되어야 한다.

세선화 알고리즘으로서 그 성능이 입증된 Zhang and Suen[10]이 제안한 알고리즘과, 이 알고리즘에서의 문제점을 부분적으로 개선한 알고리즘[11, 12, 13] 등 세선화에 관한 기존의 연구는 많이 있다. 그리고 원과 손[14, 15]은 각 화소들 간의 4 혹은 8방향의 연결 상태를 나타내는 연결값을 이용한 세선화 알고리즘인 WPTA를 제안하였고, WPTA의 결과 또한 연결값에 기반한 수치 정보로 남는다. 그림 1은 '가'자의 입력 영상을 WPTA를 이용한 세선화 결과를 나타내고 있다.

본 연구에서 한글의 모음 성분을 쉽게 찾아내기 위해 세선화된 결과로 남는 각 화소에 대해 8방향으로 조사하여 이웃한 화소의 수를 세어 그 화소의 연결값으로 부여한다. 그런데 WPTA에서는 세선화의 최종 결과가 연결값 형태로 남기 때문에 그것의 세선화 결과를 바로 본 연구의 입력으로 사용 가능하다. 그래서 본 논문의 입력으로서는 WPTA를 이용한 세선화 결과를 사용한다.



(a) 입력영상
(a) Input image

(b) 세선화된 결과
(d) Thinning result

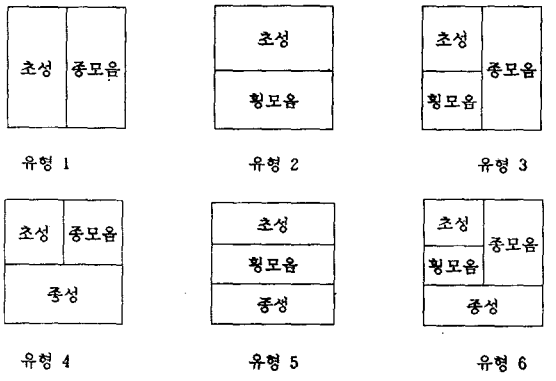
(그림 1) '가'자의 입력 영상과 WPTA를 이용한 세선화 결과
(Fig. 1) Input image of '나' and result of thinning using WPTA

3. 한글과 영어의 구조적인 특성

한글은 자소들의 조합으로 이루어진 문자로서 단자음 14자와 단모음 10자가 혼합되어 글자를 구성하며, 한글의 구성요소는 표 1[8]과 같다. 그리고 한글은 모음의 배치와 종성의 존재 유무에 따라 6가지 형식으로 분류될 수 있으며, 그림 2는 한글의 6 형식을 나타낸다.

〈표 1〉 한글의 구성요소
〈Table 1〉 Constructural element of Korean characters

종류	구성요소
한글문자	초성, 중성, 종성
초성	단자음, 쌍자음
중성	단자음, 쌍자음, 복합자음
종성	모음
모음	단모음, 복합모음
단자음	ㄱ, ㄴ, ㄷ, ㄹ, ㅁ, ㅂ, ㅅ, ㅇ, ㅈ, ㅊ, ㅋ, ㅌ, ㅍ, ㅎ
쌍자음	ㄲ, ㄸ, ㅃ, ㅆ, ㅉ
복합자음	ㄳ, ㄵ, ㄶ, ㄺ, ㄻ, ㄼ, ㄽ, ㄾ, ㄿ, ㅀ, ㅄ
단모음	ㅏ, ㅑ, ㅓ, ㅕ, ㅗ, ㅛ, ㅜ, ㅠ, ㅡ, ㅣ
복합모음	ㅐ, ㅒ, ㅔ, ㅖ, ㅘ, ㅙ, ㅚ, ㅝ, ㅞ, ㅟ, ㅢ
종모음	ㅏ, ㅑ, ㅓ, ㅕ, ㅣ, ㅐ, ㅔ, ㅒ, ㅖ, ㅖ
횡모음	ㅗ, ㅛ, ㅜ, ㅠ, ㅡ



(그림 2) 한글의 6 형식
(Fig. 2) 6 types of Korean characters

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

a b c d e f g h i j k l m n o p q r s t u v w x y z

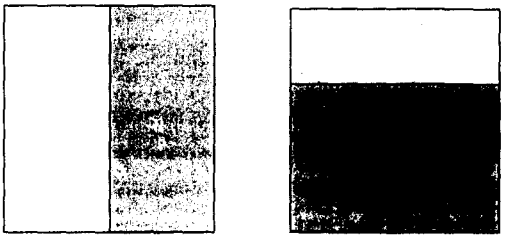
(그림 3) 영어의 알파벳
(Fig. 3) The alphabet

그림 2에서 알 수 있듯이 한글의 모든 글자에는 종모음과 횡모음 중 적어도 하나를 포함하고 있는 특징이 있다. 영어는 알파벳 26자로 구성되며, 그림 3은 영어의 알파벳을 표시하고 있다. 영어의 알파벳은 소문자 i와 j를 제외하고는 글자의 모든 획이 연결되어 하나의 연결요소를 이루는 특징이 있다.

4. 제안된 한글과 영어의 구별 방법

4.1 연결 요소가 1개 일 때 구별

i, j를 제외한 모든 알파벳은 연결요소가 1 개이고, 한글에서는 정상적인 경우 연결요소가 2개 이상이 된다. 그러나 자음과 모음의 접촉에 의해 1개의 연결요소로 구성될 수도 있다. 그러므로 연결 요소가 1개일 때 영어와 한글을 구별하기 위하여 한글에 존재하는 종모음이나 횡모음의 존재 유무를 조사하여 한글과 영어를 구별한다. 이 종모음과 횡모음은 그림 2의 한글의 6 형식에서 알 수 있듯이 그림 4에서 약간 검게 표시된 영역에 존재한다.



(a) 종모음 존재 영역
(a) Vertical vowel existing area

(b) 횡모음 존재 영역
(b) Horizontal vowel existing area

(그림 4) 종모음과 횡모음의 존재영역
(Fig. 4) Existing area of vertical and horizontal vowel

4.1.1 한글에 존재하는 종모음과 횡모음

한글에서 연결 요소가 1개이기 위한 필요조건은 종

모음으로 'ㅏ'나 'ㅓ' 성분이 존재하거나, 횡모음으로 'ㅗ'나 'ㅜ' 성분이 존재해야 한다. 물론 종모음으로 'ㅑ', 'ㅕ', 횡모음으로 'ㅛ', 'ㅠ' 성분이 존재해도 마찬가지인데, 'ㅑ', 'ㅏ' 성분이 2개 있는 것으로, 그리고 'ㅕ', 'ㅛ', 'ㅠ'에 대해서도 'ㅓ', 'ㅗ', 'ㅜ' 성분이 각각 2개 있는 것으로 생각하면 된다. 'ㅟ' 및 'ㅢ'처럼 종모음의 'ㅣ'와 횡모음의 'ㅡ'가 들어간 한글에서 초성과 'ㅣ' 혹은 'ㅡ'가 붙어서 연결 요소가 1개로 나타날 수 있다. 그런 경우 초성과 'ㅣ', 'ㅡ'가 연결되어 'ㅣ'는 'ㅓ', 'ㅡ'는 'ㅗ' 성분으로 보인다. 그러므로 연결 요소가 1개 일 때는 한글에 존재하는 종모음의 'ㅏ', 'ㅓ', 횡모음의 'ㅗ', 'ㅜ' 성분의 유에 따라 한글과 영어를 구별할 수 있다.

4.1.2 종모음과 횡모음의 검출 방법

본 연구의 입력으로 사용되는 세선화 결과는 연결값으로 표현된다. 그림 5는 '오', '구', '거' 자의 입력영상을 나타낸 것이고, 그림 6은 그림 5의 입력영상을 WPTA를 이용하여 세선화한 결과이다.

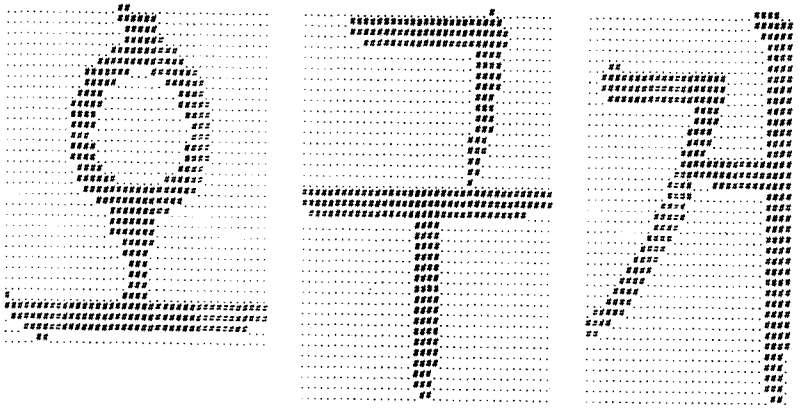
그림 1과 그림 6에서 알 수 있듯이 'ㅏ', 'ㅓ', 'ㅗ', 'ㅜ' 성분에서 두 획이 만나는 위치에는 연결값이 3인 화소가 1개 이상 존재한다. 그러므로 종모음 및 횡모음의 검출은 그림 4에서 나타낸 종모음과 횡모음의 존재 영역에서 화소의 연결값이 3인 것을 찾아, 그것을 기준으로 하여 한글의 'ㅏ', 'ㅓ', 'ㅗ', 'ㅜ' 성분에 해당하는지를 조사하면 된다.

연결값이 3 이상인 화소 P에 대해 그것이 한글의 'ㅏ', 'ㅓ', 'ㅗ', 'ㅜ' 성분을 가지고 있는지를 판단하기 위해 그 화소의 이웃화소의 배치형태를 조사하며, 그림 7은 화소 P에 대한 그것의 이웃한 8방향의 화소를 나타내고 있다.

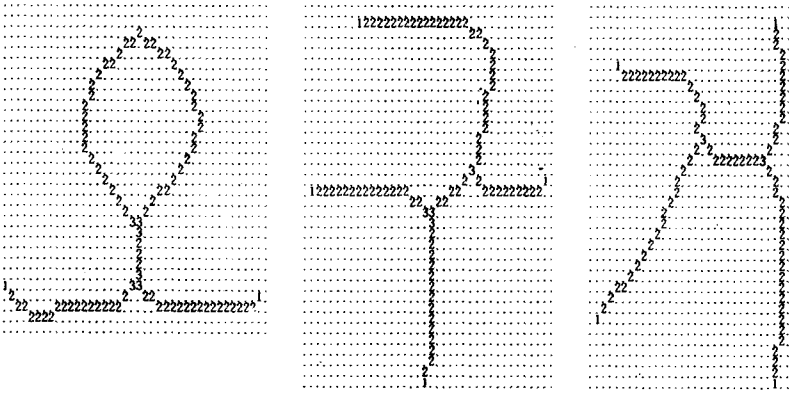
한글의 'ㅏ'를 세선화한 결과에서 두 획의 만나는 위치에 있는 연결값이 3인 화소 P와 연결값을 가진 그것의 이웃화소는 거의 대부분이 그림 8에서 나타낸 4가지 범주에 속한다. 화소 P의 연결값이 3일 때 그 화소와 이웃화소들의 배치 형태로 한글의 'ㅏ' 성분의 포함 여부를 알고자 한다면 일차적으로 그림 8의 4가지 경우 중의 한 형태로 되어 있는지를 조사해야 한다.

한글에서 초성과 'ㅏ'가 붙어서 세선화한 결과에서 두 획의 만나는 위치에는 연결값이 4인 화소 P가 생길 수 있다. 그 화소 P와 그것의 이웃화소는 거의 대부분이 그림 9에서 나타낸 3가지 범주에 속한다. 그런데 연결값이 4인 화소 P와 그것의 이웃화소로부터 'ㅏ' 성분의 포함 여부만을 판단한다면 그림 9의 (a), (b), (c)의 경우는 모두 그림 8의 (d)에 포함시킬 수 있다.

그림 8과 그림 9를 식으로 표현하면 아래의 식 (1)이 된다. 같은 방법으로 화소 P의 연결값이 3 혹은 4인 화소와 그 이웃화소로부터 'ㅓ', 'ㅗ', 'ㅜ'의 포함 여부를 판단하는 식은 각각 식 (2), (3), (4)로 표현된다. 아래의 식 (1)-(4)에서 ∧는 논리곱, ∨는 논리합을 의미하고 연결값이 수치 형태로 남아 있는 화소는 논리적 1로, 연결값이 없는 화소는 논리적 0으로 간주



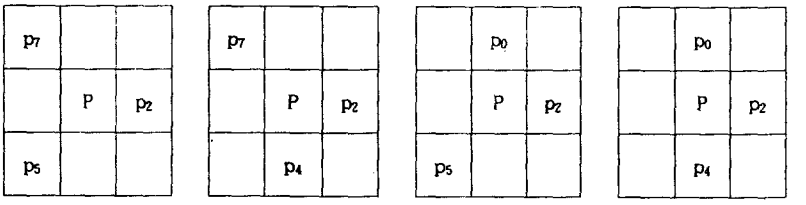
(그림 5) '오', '구', '거'자의 입력 영상
(Fig. 5) Input image of '오', '구', and '거'



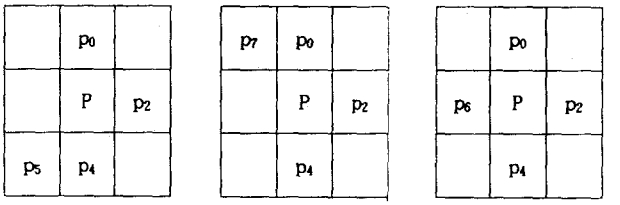
(그림 6) 그림 5의 세선화 결과
(Fig. 6) Thinning result of Fig. 5

p_7	p_0	p_1
p_6	P	p_2
p_5	p_4	p_3

(그림 7) 화소 P의 8 방향 이웃화소
(Fig. 7) 8-directional neighboring pixels of the pixel P



(그림 8) 연결값이 3인 화소 P 주변에서 나타날 수 있는 'ㅏ'의 4가지 모양
(Fig. 8) 4 types of 'ㅏ' surrounding about pixel P whose connection value is 3



(그림 9) 연결값 4주변에서 나타날 수 있는 'ㅏ'의 3가지 모양
(Fig. 9) 3 types of 'ㅏ' surrounding about pixel P whose connection value is 4

한다.

$$P_2 \wedge (P_0 \vee P_7) \wedge (P_4 \vee P_5) \quad (1)$$
$$P_6 \wedge (P_0 \vee P_1) \wedge (P_3 \vee P_4) \quad (2)$$
$$P_0 \wedge (P_2 \vee P_3) \wedge (P_5 \vee P_6) \quad (3)$$
$$P_4 \wedge (P_1 \vee P_2) \wedge (P_6 \vee P_7) \quad (4)$$

그러므로 'ㅏ', 'ㅓ' 성분의 존재 유무를 판단하기 위해서는 종모음 존재 영역 내에 연결값이 3 혹은 4인 화소 P에 대해 다음의 4 가지 조건(조건 1-조건 4)을 조사하는데, 모두 다 만족하는 화소가 있으면 'ㅏ' 나 'ㅓ' 성분이 존재한다. 아래의 조건들에서 화소 P의 행의 값을 i, 열의 값을 j라 가정하고 화소 P의 좌표를 (i, j)로 표기한다.

조건 1: 식 (1) ∨ 식 (2) = 1이어야 한다.

조건 2: 화소 P를 기준점으로 해서 수평 방향으로 향하는 가지에서 거리 3에 있는 화소의 좌표를 (k, l)이라 할 때 두 화소가 형성하는 선분의 기울기의 절대치 $\frac{|k-i|}{|l-j|}$가 1/3 이하이어야 한다.

조건 3: 화소 P를 기준점으로 해서 윗쪽으로 향하는 가지의 끝화소 좌표를 (m, n)이라 할 때 두 화소가 형성하는 선분의 기울기의 절대치 $\frac{|m-i|}{|n-j|}$가 2 이상이어야 한다.

조건 4: 화소 P를 기준점으로 해서 아래쪽으로 향하는 가지의 길이는 자소획의 폭 이상이어야 한다.

한글의 횡모음에서 'ㅗ', 'ㅜ' 성분의 유무를 판단하기 위해서는 횡모음 존재 영역 내에 연결값이 3 혹은 4인 화소 P에 대해 다음의 3 가지 조건(조건 5-조건 7)을 조사하는데, 모두다 만족하는 화소가 있으면 'ㅗ' 나 'ㅜ' 성분이 존재한다. 아래의 조건들에서 'ㅏ', 'ㅓ' 성분 조사시와 마찬가지로 화소 P의 좌표를 (i, j)로 표기한다.

조건 5: 식 (3) ∨ 식 (4) = 1이어야 한다.

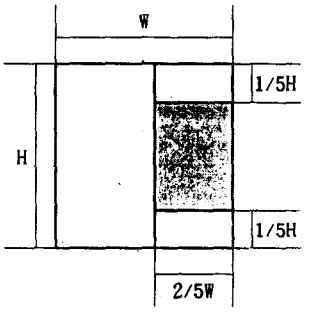
조건 6: 화소 P를 기준점으로 해서 왼쪽으로 향하는 가지의 화소수는 자소획 폭의 2배 이상이어야 하고, 그 방향의 끝화소 좌표를 (m, n)이라 할 때 두 화소가 형성하는 선분의 기울기의 절대치 $\frac{|m-i|}{|n-j|}$가 1/2 이하이어야 한다.

조건 7: 화소 P를 기준점으로 해서 오른쪽으로 향하는 가지의 길이는 자소획의 폭 이상이어야 한다.

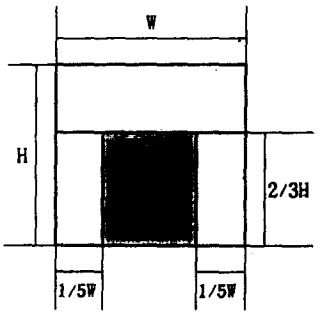
'ㅏ', 'ㅓ', 'ㅗ', 'ㅜ' 성분의 유무 판단시에 조건 2, 조건 3, 조건 6에서 기울기를 검사하는 것은 'ㅏ', 'ㅓ', 'ㅗ', 'ㅜ' 성분을 정확하게 가려내기 위한 것이고, 기울기 검사에서 사용된 한계 수치는 다양한 실험을 통해 얻은 수치이다. 그리고 'ㅏ', 'ㅓ' 성분 조사시에 사용한 조건 3에 대응되는 조건을 'ㅗ', 'ㅜ' 성분 조사시에는 사용하지 않은 것은 'ㅗ' 성분의 수직 방향의 획은 초성과 접촉시에 세선화된 결과에서 그 길이가 짧아 그러한 조건을 적용하기가 곤란했기 때문이다.

4.1.3 종모음과 횡모음 검출 영역의 축소

한글의 'ㅏ', 'ㅓ', 'ㅗ', 'ㅜ' 성분을 조사하기 위해서 종모음과 횡모음 존재 영역 내의 화소 중 연결값이 3이거나 4인 화소만 조사하면 되므로 조사대상 영역을 그림 10에서 표시한 영역으로 축소할 수 있다. 이 영역의 축소는 알고리즘 실행시간도 단축시켜 주지만 자소들의 가장자리 부분에 생기는 잡영들로 인해, 실제로는 'ㅏ', 'ㅓ', 'ㅗ', 'ㅜ' 성분이 아닌데 'ㅏ', 'ㅓ', 'ㅗ', 'ㅜ' 성분으로 판단할 가능성을 배제할 수 있으므로, 종모음과 횡모음의 존재 영역을 그림 10에서 약간 검게 표시한 영역으로 제한한다.



(a) 축소된 종모음 존재영역
(a) Reduced vertical voewel existing area



(b) 축소된 횡모음 존재영역
(b) Reduced horizontal voewel existing area

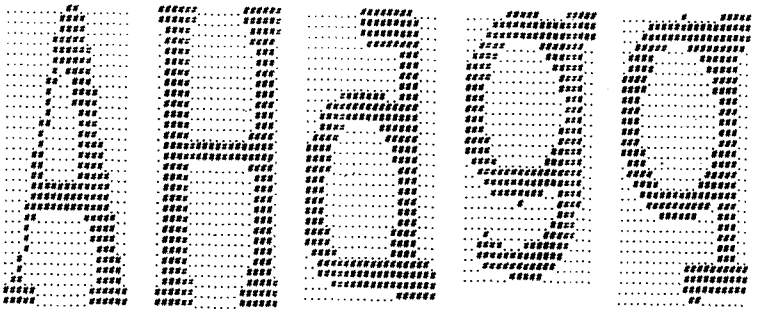
(그림 10) 축소된 종모음과 횡모음의 존재영역
(Fig. 10) Reduced existing area of vertical and horizontal vowel

4.1.4 종모음이 검출되나 영어인 글자들

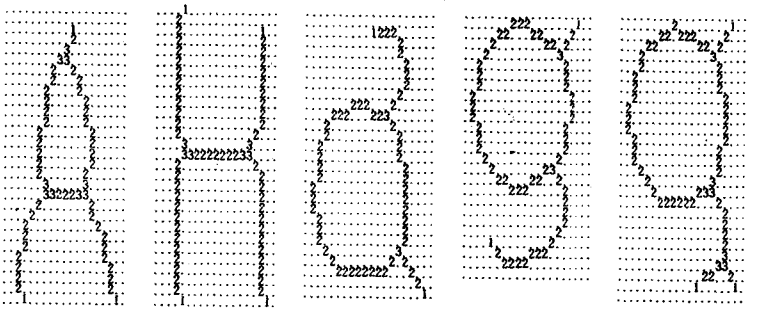
그림 11은 영어의 A, H, d, g, q 자의 입력 영상이고, 그림 12는 그림 11을 세선화한 결과이다. 그림 12에서 알 수 있듯이 영어의 A, H, d, g, q 자는 종모음 존재 영역에 세조건(조건1-조건3)을 모두 만족하는 'ㅓ' 성분이 존재할 수 있다. 종모음 존재 영역에 'ㅓ' 성분을 가진 A, H, d, g, q 자를 영어로 판정해 주기 위해서는 별도의 처리를 필요로 한다.

4.1.5 H, d, g, q 자를 영어로 판정하는 방법

세선화된 결과에서 종모음 존재 영역에서 'ㅓ' 성분이 2개 검출되면 그것은 'ㅕ'가 존재한다는 것을 의미하기 때문에 종모음 존재 영역에서 'ㅓ'성분이 2개 검출되면 그 글자는 영어의 A, H, d, g, q는 아니다.

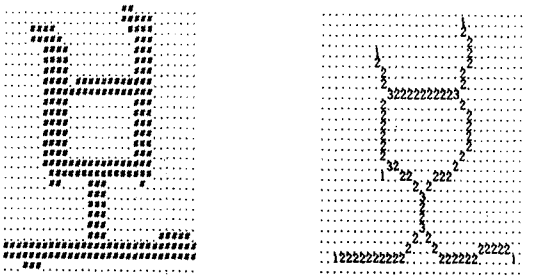


(그림 11) A, H, d, g, q 자의 입력 영상
(Fig. 11) Input image of A, H, d, g, and q



(그림 12) 그림 11의 세선화 결과
(Fig. 12) Thinning result of Fig. 11

종모음 존재 영역에서 'ㅓ' 성분이 1개 검출되고 'ㅏ'성분은 없으며, 또한 좌반면에 'ㅓ' 성분은 없고 'ㅏ'성분이 1개 검출되면 일단 영어의 A나 H로 판단할 수 있으나, 그럴 경우에 그림 13에서와 같이 한글의 '보'자 경우도 영어의 A나 H로 간주하여 영어로 오판할 수 있다. 그러므로 종모음 존재 영역에서 'ㅓ' 성분이 1개 존재하고 'ㅏ'성분은 없으며, 횡모음 존재 영역에서 'ㅗ'나 'ㅜ' 성분이 없고, 좌반면에 'ㅓ'의 대칭적인 위치에 'ㅏ' 성분이 1개 존재하면 영어의 A 혹은 H로 간주하여 영어로 판정한다.



(a)입력영상 (a) Input image

(b)세선화된 결과 (b) Thinning result

(그림 13) 한글 '보'자의 입력영상과 세선화된 결과
(Fig. 13) Input image and thinning result of Korean character '보'

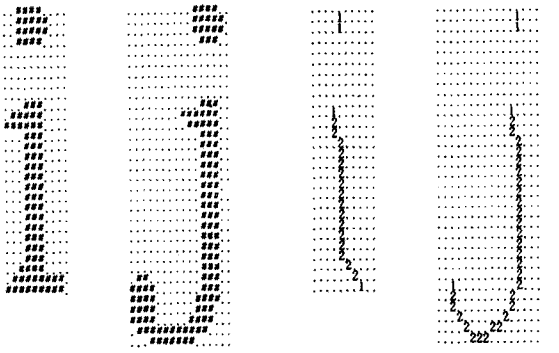
종모음 존재 영역에 'ㅓ' 성분이 1개 존재하고 'ㅏ' 성분은 없으며, 횡모음 존재 영역에서 'ㅗ'나 'ㅜ' 성분이 없고, 또한 연결값이 3인 화소를 포함하여 폐곡선을 이루고 폐곡선의 가로폭이 글자의 가로폭과 비슷하면 영어의 d, g, q 중 하나로 간주하고 영어로 판정한다.

영어의 A, H, d, g, q 자의 경우가 아니면서 횡모음과 종모음의 존재 영역에 'ㅏ', 'ㅓ', 'ㅗ', 'ㅜ' 성분이 1개라도 존재하면 한글로 판정한다.

4.2 연결 요소가 2개 일 때 구별

연결 요소가 2개 일 때 영어의 경우는 소문자 i와 j 밖에 없다. 그러므로 연결 요소가 2개 일 때 영어의 i나 j가 아니면 한글로 판정한다. 연결 요소가 2개 일 때 영어의 i와 j인지를 검사하기 위해 i와 j의 구조적인 특성을 알아본다.

그림 14에서 알 수 있듯이 i와 j는 글자 영역의 아래쪽 2/3 영역 내의 연결 요소가 1개이고, 그 영역 내의 화소수가 글자의 가로폭과 세로폭을 합한 것보다 작으며, 윗쪽으로부터 아래쪽으로 자소획 폭 크기보다 조금 큰 위치에서 좌에서 우로 주사하였을 때 골격화소를 하나도 만나지 않는다면 영어의 i나 j로 간주하여 영어로 판정한다.



(a) i, j의 입력영상 (a) Input image of i, j

(b) (a)의 세선화 결과 (b) Thinning result of (a)

(그림 14) i와 j의 구조
(Fig. 14) Structure of i and j

4.3 제안된 한글과 영어의 구별 알고리즘

각 글자의 세선화된 결과를 입력으로 하여 그것을 한글과 영어로 구별하는 알고리즘을 아래에 기술하였다.

단계 1. 세선화된 결과를 입력으로 하여 그것의 연결 요소를 구한다.

단계 2. 연결요소가 3개 이상이면 한글로 판정하고 종료한다.

단계 3. 연결요소가 2개이면 영어의 i나 j에 해당하는지를 조사하여 그것으로 간주되면 영어로 판정하고 종료하며, 그렇지 않으면 한글로 판정하고 종료한다.

단계 4. 종모음 존재 영역에서 'ㅓ' 성분이 1개 존재하고 'ㅏ' 성분은 없으며, 횡모음 존재 영역에서 'ㅗ'나 'ㅜ' 성분이 없고, 좌반면에 'ㅏ' 성분이 1개 존재하면 영어(A나 H로 간주)로 판정하

고 종료한다.

단계 5. 종모음 존재 영역에 'ㅓ' 성분이 1개 존재하고 'ㅏ' 성분은 없으며, 횡모음 존재 영역에서 'ㅗ'나 'ㅜ' 성분이 없고, 또한 연결값이 3인 화소를 포함하여 폐곡선의 가로폭이 글자폭과 비슷하면 영어(d, g, q 중 하나로 간주)로 판정하고 종료한다.

단계 6. 종모음의 존재 영역에 'ㅏ'나 'ㅓ'가 1개 이상 존재하거나, 횡모음의 존재 영역에 'ㅗ'나 'ㅜ'가 1개 이상 존재하면 한글로 판정하고 종료하며, 그렇지 않으면 영어로 판정하고 종료한다.

5. 실험 및 고찰

5.1 실험환경

본 연구의 실험 환경은 C 언어를 사용하여 알고리즘을 구현하여 Pentium PC에서 실행하였으며, 영상 입력을 위해 EPSON GT-9000 스캐너를 해상도 300 DPI로 해서 사용하였다. 스캐너 입력으로는 문서작성기의 일종인 로 문서를 작성하여 큐닉스 레이저 프린터(큐레이저 SFIIIs)에서 해상도 300 DPI로하여 출력한 것을 사용했다. 그리고 한글은 완성형 2350자를, 영어는 'The Korea Herald' 신문 사설의 내용 중 2350자를 실험 대상으로 하였으며 문서내의 글자들은 개개의 글자로 바르게 분리되었다고 가정하였다.

한글의 경우에 동일한 문서에 대해 에서 제공되는 명조체, 신명조체, 고딕체로 각각 출력한 것을, 영어의 경우도 동일한 문서를 대문자와 소문자로 각각 작성하여 그 각각에 대해에서 제공되는 영어의 명조체, 신명조체, 고딕체로 출력한 것을 스캐닝하여 실험하였다.

한글의 신명조체, 영어의 신명조체 대, 소문자 및 명조체 대문자의 경우에 획의 일부분이 가늘어서 스캐닝을 한 영상에서 자소획이 끊어질 수 있으므로 선의 끊어짐을 방지하기 위하여 스캐너의 환경설정시에 흑화소와 백화소를 구분할 때 사용되는 그레이 값(gray value)의 한계치(threshold)를 조정하여 선의 끊어짐이 없도록 하였다.

5.2 결과 및 고찰

한글의 경우 실험 결과를 표 2에, 그리고 영어의 경우 실험 결과를 표 3에 요약하였다. 한글과 영어의 소문자가 50:50으로 쓰여진 문서에 대해 명조체, 신명조체, 고딕체 각각에 대한 한글과 영어의 구별율을 표 4에 나타내었다. 표 4에서 알 수 있듯이 명조체, 신명조체의 경우는 100%에 가까운 구별율을, 고딕체의 경우도 99.5% 이상의 높은 구별율을 나타내었다.

〈표 2〉 한글의 실험 결과

〈Table 2〉 Experimental result of Korean characters

활자체	오판된 글자수	오판된 글자	구별율(%)
명조체	2	돌, 왓	99.91
신명조체	3	쭉, 찝, 크	99.87
고딕체	11	갈, 그, 뭴, 러, 려, 뵴 쒔, 올, 했, 출, 크	99.53

〈표 3〉 영어의 실험 결과(오판된 글자 뒤의 괄호 안은(오판된 글자수/2350자 내에 포함된 그 글자수)의 의미임)

〈Table 3〉 Experimental result of English characters

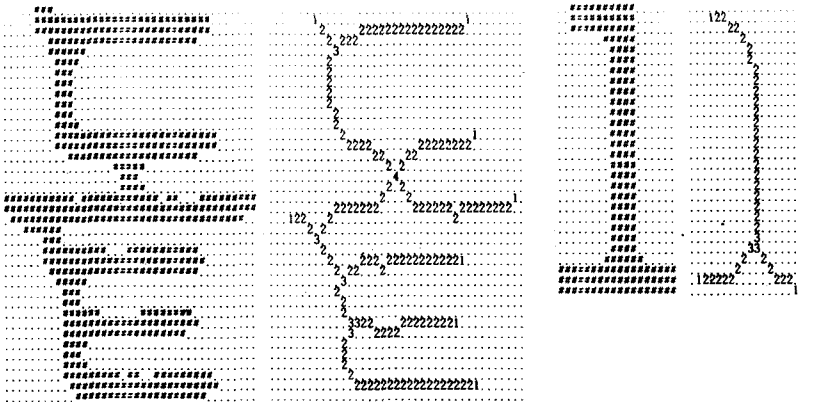
활자체		오판된 글자수	오판된 글자	구별율(%)
대문자	명조체	0		100.00
	신명조체	0		100.00
	고딕체	0		100.00
소문자	명조체	0		100.00
	신명조체	0		100.00
	고딕체	9	1(9/96)	99.62

〈표 4〉 한글과 영어의 소문자가 50 : 50일 때의 구별율

〈Table 4〉 Distinction rate for 50 : 50 document composed of Korean character and lower alphabet

활자체	구별율(%)
명조체	99.96
신명조체	99.94
고딕체	99.57
평균	99.82

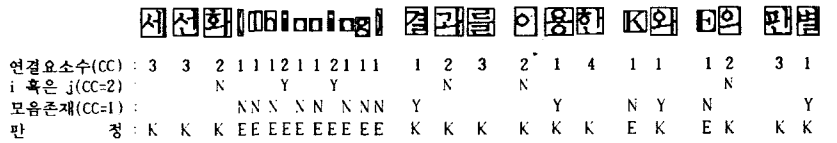
한글을 영어로 오판하는 경우는 그림 15(a)처럼 종모음과 횡모음 검출시에 연결값이 3 혹은 4인 화소를 중심으로 이웃화소들의 배치형태가 그림 8과 그림 9



(a) 한글 명조체의 '돌'
(a) '돌' of Korean Myungjo-font

(b) 영어 고딕체의 'l'
(b) 'l' of English Gothic-font

(그림 15) 오판된 글자의 예
(Fig. 15) An example of character ditinguished incorrect



(그림 16) 한글과 영어의 구별 예
(Fig. 16) A distinction example of the Korean and English characters

에서 보여준 형태가 아닌 경우가 대부분이었다. 표 3에서 오판된 영어의 고딕체 l은 그림 15(b)에서 알 수 있듯이 그 모양으로 봐서 종모음 존재영역에 'ㅗ' 성분이 존재하는 것으로 인식하여 한글로 오판했다. 영어의 고딕체 l의 경우는 구조가 특이하기 때문에 예외 처리를 하면 쉽게 해결할 수 있다. 그림 16은 글자들 간의 접촉이 없는 한 줄의 문장을 대상으로 한글과 영어를 구별하는 것을 예시하였다. 그림 16에서 'Y'는 예, 'N'은 아니오, 'K'는 한글, 'E'는 영어를 의미한다.

한글의 종모음과 횡모음의 유무 판단시 연결값이 3 혹은 4인 화소를 기준으로 3X3 윈도우내에서 식(1)-식(4)를 적용하였는데, 윈도우를 확장하여 좀 더 세밀하게 조사를 하면 구별율이 더 좋아질 것으로 생각된다.

6. 결 론

본 논문에서는 다중 활자체 환경에서 세선화된 결과를 이용하여 한글과 영어를 구별하는 방법을 제안하였다. 제안된 방법에서는 한글과 영어를 구별하기 위하여 각 글자의 연결요소 개수와 한글의 모음 성분을 추출하여 한글과 영어를 구별하였다.

영어의 경우에는 연결 요소가 1개(소문자 i, j는 예외)인데 한글의 경우는 연결 요소가 1개 이상이 된다. 그래서 연결 요소가 1개 일 경우 한글에 존재하는 종모음과 횡모음을 검출하여 한글과 영어를 구별하였고, 연결 요소가 2개인 경우는 영어의 소문자 i, j에 해당하는지를 조사하고 난 뒤 한글과 영어를 구별하였다. 그리고 연결 요소가 3개 이상이면 한글로 판정하

였다.

본 연구에서 제안된 알고리즘을 구현하여 실험한 결과, 명조체, 신명조체의 경우는 100%에 가까운 구별율을, 고딕체의 경우도 99.5% 이상의 높은 구별율을 나타내었다.

향후 연구 과제로 본 연구에서는 WPTA 결과를 입력으로 사용하였는데 연결성이 보장되는 어떤 세선화 알고리즘으로도 한글과 영어를 구별하는 보다 일반성 있는 알고리즘를 제안하는 것이다.

참 고 문 헌

[1] S. Kahan, T. Pavlidis, "On the Recognition Printed Chacters of Any Font and Size," IEEE Trans. on Pattern Analysis and Machine intelligence, Vol. 2, no. 5, pp. 274-288, Sep. 1987.

[2] P. K. Kim, H.J. Kim, "On-Line Recognition of Run-On Korean Characters," International Conf. on Document Analysis and Recognition, no. 5, pp. 54-57, Montreal, Canada, Aug. 1995.

[3] 이진수, 권오준, 방승양, "개선된 자소 인식 방법을 통한 고인식율 인쇄체 한글 인식," 정보과학회논문지(B), 제 23권 제 8호, pp. 841-851, 1996년 8월.

[4] 박덕원, 박종원, "3×3 템플레이트를 이용한 여러 영문 활자체의 인식," 정보과학회논문지(B), 제 23권 제 6호, pp. 625-634, 1996년 6월.

[5] 이성환, "다양한 활자체 및 크기를 갖는 대용량 한글의 고속 인식을 위한 최적 트리 분류기," 정보과학회논문지, 제 20권 제 8호, pp. 1083-1092, 1993년 8월.

[6] 조성배, 김진형, "인쇄체 한글 문자의 인식을 위한 계층적 신경망," 정보과학회논문지, 제17권 제3호, pp. 306-316, 1990년 5월.

[7] 김우성, 방승양, "신경회로망을 이용한 한글 한자 혼용 문서 인식에 관한 연구," 대한전자공학회 논문지, 제29권 B편 제2호, pp. 50-59, 1992년 2월.

[8] 심상완, 이성범, 남궁재찬, "인쇄체 문서의 문자 영역에서 한글과 한자의 구별에 관한 연구," 한국통신학회 논문지, 제18권 제 6호, pp. 802-814, 1993년 6월.

[9] 이성환, '문자인식', pp. 218-219, 홍릉과학출판사, 1993.

[10] T.Y. Zang and C.Y. Suen, "A fast parallel algorithm for thinning digital patterns," Communication of the ACM, Vol. 27, no. 6, Mar. 1984.

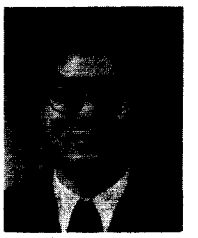
[11] A.D. Mandalia, A.S. Pandya, R. Sudhaker, "Modified fast parallel thinning algorithm for noisy handprinted characters," 92 Proceedings of the 2ND Singapore International Conf. on Image Processing, pp. 7-11, Singapore, Sep. 1992.

[12] Lu, H.E., Wang, P.S., "An Improved Fast Parallel Thinning Algorithm for Digital Patterns," In Proc. of the IEEE Conf. On Computer Vision and Pattern Recognition, pp. 364-367, 1985.

[13] Wang, P.S.P, Hui, L., Fleming Jr., T., "Further improved fast parallel thinning algorithm for digital patterns, In Computer Vision, Image Processing and Communication Systems and Appli. edt. by P.S.P Wang," pp. 37-40, 1986.

[14] 원남식, 손윤구, "4-인접 연결값을 이용한 병렬 세선화 알고리즘," 한국정보과학회논문집 제22권 제7호, pp. 1047-1056, 1995년 7월.

[15] 원남식, 손윤구, "8-이웃 연결값에 의한 병렬 세선화 알고리즘," 한국정보처리학회 논문집 제2권 제5호, 1995년 12월.



전 일 수

1984년 경북대학교 전자공학과 (학사)

1988년 경북대학교 전자공학과 (석사)

1995년 경북대학교 전자공학과 (박사)

1984년~1985년 삼성전자(주)

1989년~현재 경일대학교 전자계산학과 부교수

관심분야: 문자인식, 데이터베이스

원 남 식

1974년 인하대학교 전자과 졸업(학사)
1984년 영남대학교 대학원 전자과 졸업(석사)
1996년 영남대학교 대학원 전산공학과 졸업(박사)
1976년~1978년 한국과학기술연구소 연구원
1978년~1981년 한국전자기술연구소 연구원
1981년~현재 경일대학교 전자계산학과 교수
관심분야:문자인식, 세선화 알고리즘, 네트워크, 컴퓨터 그래픽스

이 두 한

1987년 경북대학교 전자공학과 공학사
1991년 경북대학교 전자공학과 공학석사
1993년 경북대학교 컴퓨터공학과 박사과정 수료
1994년~현재 경동전문대학 전산정보처리과 조교수
관심분야:객체지향 데이터베이스, 인공지능