論文96-33B-11-13

붙은 글자들이 포함된 인쇄체 한·영 혼용 문서에서의 효과적인 문자 인식 알고리즘

(An Efficient Character Recognition Algorithm in Printed Korean/English Documents Including Touching Characters)

金 桂 慶 *, 金 鎭 浩 **, 秦 成 一 *, 崔 興 文 *

(Kye-Kyung Kim, Jin-Ho Kim, Sung-Il Chien, and Heung-Moon Choi)

요 약

본 논문에서는 인쇄체 한·영 혼용 문서에서 붙은 문자들까지 정확히 세그멘테이션하여 효과적으로 인식해 낼 수 있는 문자 인식 알고리즘을 제안하였다. 한·영 혼용 문서에서 문자 영역 세그멘테이션을 위해 한글과 영문의 수직·수평 방향의 공간적 위치 정보를 기록 블록(writing block)으로 정의하고, 유형 분류기와 문자 인식기의 결과로부터 추출되는 신뢰도 지수값을 이용하여 오인식을 최소화시킬 수 있도록 하였다. 제안된 알고리즘을 대한전자공학회 논문지의 인식에 적용해 본 결과 영어 요약문의 경우 96.8%정도의 인식률을 얻을 수 있었고, 한·영 혼용의 경우 97.8% 정도의 인식률을 얻을 수 있었으며 특히 본 알고리즘은 세그멘테이션 에러를 줄이는데 상당히 효과적으로 이용될 수 있음을 확인하였다.

Abstract

In this paper, we present a character recognition algorithm in printed Korean and English documents including touching characters. We derived two rules to segment and recognize touching characters in the bilingual documents, one from the shape characteristics of Korean and English characters, of the writing blocks defined in this paper, and the other from the RF(reliability factor) values generated from the classifiers. Overall classification accuracy for the KITE paper of the proposed algorithm was about 96.8% for the English abstract, and about 97.8% for the bilingual parts. Also we confirmed the proposed algorithm significantly improves the accuracy of character segmentation of the actual mixed Korean and English documents including touching characters.

Ⅰ. 서 론

문서 영상 인식은 1980년대 초부터 패턴 인식 영역의 주관심 분야 가운데 하나로 연구되어 왔다. 최근의 연구 결과들을 보면 필기체 숫자들이나 다양한 활자체 문자들에 대해서 상당히 높은 인식 성능을 나타내고 있다[1-14]. 일반적으로 문자들이 깨끗하게 인쇄되어 있거나 개별 문자들의 띄어쓰기가 잘 되어 있는 (well-formed and well-spaced) 인쇄 문서들을 대상으로 하는 문자 인식 시스템을 설계하는 것은 쉬우며, 특히 신경회로망을 이용한 다중 해상도 OCR 시스템의 경우에는 다양한 활자체의 분리된 인쇄체 문자들에 대해 99.6% 이상의 높은 인식률을 나타내고 있다[1]. 그

* 正會員, 慶北大學校 電子電氣工學部
(The School of Electronics and Electrical Engineering, Kyungpook National University)
** 正會員, 慶北産業大學校 電子工學部
(Department of Electronic Engineering, Kyungpook Sanup University)
接受日字:1996年8月22日, 수정완료일:1996年11月11日

러나, 입력 문서 영상의 화질(quality)이 낮거나 다중 활자체와 다중 언어 문서에서 붙은 문자들이 많은 경우에는 높은 인식률을 얻을 수 있는 OCR 시스템의 개발이 비교적 어렵다.

실제 대부분의 인식 에러는 문자 영역 세그멘테이션을 잘못함으로써 초래된다[2]. 특히, 다양한 활자체로 구성된 한·영 혼용 문서에서 문자열을 추출한 다음 개별 문자 영역을 정확히 세그멘테이션해 내는 것은 쉬운 과제가 아니다[3]. 입력 문서 영상에 있는 영어 알파벳들은 다양한 활자체 및 크기, 또는 스캐너의 한정된 해상도 때문에 이웃하는 알파벳과 겹쳐 있거나 붙어 있는 경우가 흔하므로, 한 개의 알파벳 영역을 세그멘테이션 해내기 위해서는 수평 방향으로 붙은 문자들을 분할해야 되는 경우가 생긴다. 따라서, 영어 알파벳 영역을 정확히 세그멘테이션해 내기 위해서는 수평 방향으로 서로 붙은 알파벳들을 분할하여 세그멘테이션하고 인식하는데 주안점을 두어야 한다. 한편, 한글은 두 개, 세 개 또는 네 개의 자소들이 수평 또는 수직 방향의 2차원 형태로 결합되어 구성된다. 한글 문서 영상에서 한 문자 영역을 정확히 세그멘테이션해 내기 위해서는 영어와는 달리 수평으로 분리된 자소의 결합과 문자간의 실제적 접촉 또는 수직 방향 투영으로 인해 접촉되어 보이는 부분의 분할 문제를 동시에 고려해야 한다. 따라서, 한·영 혼용 문서에서 개별 문자 영역을 정확히 세그멘테이션하고 인식해 내기 위해서는 한글 및 영문의 구조적 특성상 상호 배타적인 요소들이 존재하므로, 이들을 고려하여 논리적이고 체계적인 방안을 마련할 필요가 있다.

본 논문에서는 인쇄체 한·영 혼용 문서에서 붙은 문자들까지 정확히 세그멘테이션하여 효과적으로 인식해 낼 수 있는 문자 인식 알고리즘을 제안하였다. 한·영 혼용 문서에서 개별 문자 영역을 정확히 세그멘테이션하기 위해 문자열에서의 한글 및 영문의 수직·수평 위치 정보를 기록 블록으로 정의하였고, 패턴 분류기의 분류 결과로부터 추출되는 신뢰도 지수값을 이용하여 오인식을 최소화시킬 수 있는 결과 판단 기준을 마련하였다. 한글의 경우에는 대부분 크기가 비슷한 사각형의 기록 블록으로 이루어져 있으나, 영어의 경우에는 크기와 위치가 각각 다른 기록 블록들로 구성되어 있다. 이들을 이용하여 한글과 영문 영역을 구분해 내고, 특히 한글 영역에 대해 하나의 글자를 구성하는 자소들이 수평 방향으로 분리되어 있을 경우에는 이들을 한 문자 영역으로 결합시키고 이웃하는 다른 문자에 속한 자소들이 서로 붙어 있을 경우에는 이들을 수평 방향으로 분할시킬 수 있는 결합과 분할(merging and splitting) 알고리즘을 마련하였다. 또한, 유형 분류기를 도입하여 문자 인식에 앞서 여섯 개의 한글 유형과 한 개의 비 한글 유형 즉, 전체 일곱 개의 유형 가운데 한 유형으로 문자를 분류한 다음, 각 유형별로 문자 인식을 시도함으로써 인식기의 분류 부담을 줄일 수 있도록 하였다. 제안한 방법을 이용하여 한·영 혼용 문서 인식용 OCR 소프트웨어를 구현한 다음 실제 문서에 대해 인식 실험을 수행하고 결과를 검토하였다.

Ⅱ. 한글과 영어의 구조적 특징에 대한 기록 블록의 정의

일반적으로 문자 영역을 세그멘테이션하기 위해서는 먼저 수평 방향 투영 기법으로 문자열을 추출해야 한다. 일단 문자열이 추출되고 나면 다시 수직 방향의 투영 정보를 이용하여 개별 문자 영역을 세그멘테이션하게 된다. 한글은 구조적 특징상 그림 1에서와 같이 여섯 개의 문자 유형으로 분류될 수 있다.

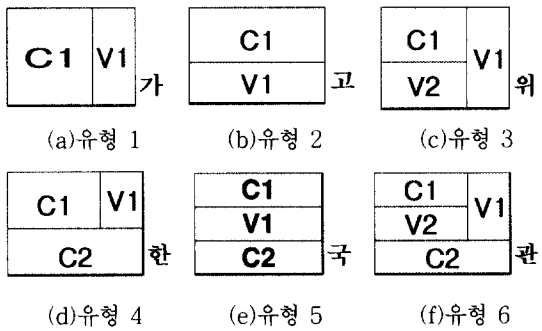


그림 1. 한글의 여섯 가지 문자 유형(C:자음, V : 모음)

Fig. 1. Six types of Hangul(Korean character) (C: consonant, V: vowel).

한글은 그림에서와 같이 두 개 또는 그 이상의 자음과 모음들이 수직 또는 수평 방향으로 결합되어 있다. 그림 1에서 유형 2, 4, 5 및 6에 속한 문자들은 수직 및 수평 방향으로 자소들이 배열되어 있기 때문에 이들을 수직 투영한 결과로부터 한 개의 문자 영역을 쉽게 세그멘테이션 할 수 있다. 그러나 유형 1과 3에 속한 문자들은 오른편에 분리된 수직 방향의 수직 모음

들이 존재하므로 수직 투영을 할 경우 한 개의 글자가 두 개의 영역으로 분리된 것처럼 나타나므로, 이 경우에는 분리된 자소들을 결합시켜 줄 수 있는 알고리즘이 마련되어야 한다.

한편, 영어의 경우 하나의 글자는 하나의 알파벳만으로 구성되므로 한글과 같이 분리된 자소를 결합시킬 필요는 없다. 따라서 연속적으로 나타나는 수직 방향의 국소 투영 블록들이 독립된 영문 알파벳인지 결합될 한글인지를 구분해 줄 수 있는 규칙이 마련되어야 한다. 특히 이들 블록들이 붙어 있을 경우 각 문자 영역을 정확히 세그멘테이션해 내는 것이 더욱 어려워지게 되며 또한 인접한 글자 블록들이 서로 붙지 않았더라도 겹친 영역에 존재하면 수직 투영 결과만으로 정확히 세그멘테이션해 낼 수 없게 되므로 블롭 컬러링(blob coloring) 기법[9]이 이용되어야 한다.

그림 2. 입력된 한 줄의 문자 영상에 대한 (a) 블롭 컬러링 결과와 (b) 수직 투영 결과
Fig. 2. (a) Blob coloring and (b) vertical projections for a scanned character line image.

그림 2는 입력된 문자 영상에 대해 블롭 컬러링과 수직 투영한 결과를 나타낸 것이다. 그림에서 수직 투영 결과만 보면 글자 영역이 서로 붙은 것처럼 보이나 블롭 컬러링 결과에서는 문자들이 서로 분리되어 있는 경우가 많이 나타나고 있다. 한글의 경우 유형 1 및 3에 속한 문자들 같이 수평으로 분리된 자소가 있을 때는 이들 자소들을 서로 결합시켜야 하고, 이웃하는 문자들 사이에 일부 자소가 서로 붙어 있을 때에는 이들을 분할시켜야 한다. 한편, 영어의 경우 독립된 블롭 컬러링 결과는 항상 개별 알파벳 영역으로 추출되어야 하며 만일 인접된 알파벳들이 붙어 있을 경우에는 이들을 분할시켜야 한다.

본 논문에서는 한·영 혼용 문서의 개별 문자 세그멘테이션 규칙을 마련하기 위해 먼저 글자의 위치 정보를 이용하여 글자의 유형별 기록 블록들을 정의하였다. 그림 3은 한·영 혼용 문자열에서 글자들의 위치 정보에 따라 기록 선(writing line)들을 도시한 것이다. 영어의 경우 upper_line, upper_baseline, lower_baseline, lower_line을 가지게 되며[5], 한글의 경우 문자의 상한 경계선인 top_line과 하한 경계선인 next_to_lower_line을 가지게 된다. 영어의 기록 영역은

그림 3. 한·영 혼용 문자열에서의 여섯 개 기록 선
Fig. 3. The six writing lines for mixed languages.

기록 선에 따라 하위(lower), 중간(middle), 상위(upper) 영역으로 나눌 수 있다. 하위, 중간 및 상위 영역은 각각 lower_baseline과 lower_line사이, upper_baseline과 lower_baseline사이 및 upper_line과 upper_baseline사이의 영역을 나타낸다. 기록 선을 보면 한글의 top_line이 가장 위쪽에 위치하게 되고 영어의 lower_line이 한글의 next_to_lower_line보다 낮은 가장 아래쪽에 위치하고 있다. 그림 4는 여섯 개의 기록 선에서 각 문자가 차지하는 영역을 사각 형태의 기록 블록으로 정의하여 나타낸 것이다. 영어의 경우에는 Ⓐ, Ⓑ 및 Ⓒ 세 개의 기록 블록 중 하나에 속하게 되고 모든 한글 문자의 경우에는 Ⓓ와 같은 한 개의 기록 블록 내에 포함된다. 이와 같은 기록 블록들은 일반적으로 비례 폰트로 쓰여진 글자들에 대해서는 동일하게 적용되므로 폰트의 종류에 무관하게 세그멘테이션 규칙에 적용할 수 있다.

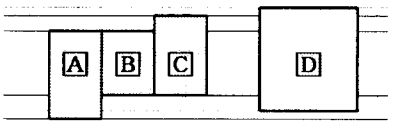


그림 4. 한글과 영문의 기록 블록들
Fig. 4. Character writing blocks of English(Ⓐ,Ⓑ, Ⓒ) and Hangul(Ⓓ).

표 1은 그림 4에 나타낸 기록 블록의 종류별 문자들의 예를 나타낸 것이다.

그림 5는 서로 인접하는 영어 알파벳 쌍이 가질 수 있는 기록 블록들의 종류를 나타낸 것이다. 연속된 두 영문자의 경우 이들 아홉 가지 중의 한 경우에 속하게 된다. 연속된 알파벳은 한 개의 한글과 비슷한 형태를 갖기 때문에 그림 5와 같은 연속 기록 블록 형태 정보

를 영어의 알파벳 영역 세그멘테이션 뿐만 아니라 한·영 혼용 문자열에서 한글 영역과 영문 영역을 구분하기 위한 중요한 정보로도 이용하였다.

표 1. 네 가지 기록 블록의 종류별 문자 예
Table 1. Character examples for each of the four writing block types.

Writing block types	Location	Character examples
A	Characters are located in the lower and middle zones	g, p, q, y
B	Characters are located in the middle zone only	a, c, e, m, n, o, r, s, u, v, w, x, z
C	Characters are located in the upper and middle zones	All capital characters, b, d, f, h, k, l, i, t
D	Hangul zones	All Hangul

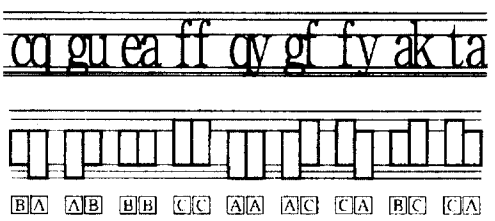


그림 5. 인접된 영어 알파벳 쌍에 대한 아홉 가지 기록 블록들의 형태
Fig. 5. Categorized writing blocks for contiguous pairs of alphabets in English.

그림 6은 전술한 여섯 개의 기록 선상에 여섯 가지 유형의 한글들을 도시한 것이다. 유형 2를 제외한 모든 한글은 영어 알파벳보다 문자의 상한 경계선이 더 높기 때문에 이들 영어 알파벳으로부터 한글을 분할하기 위한 특징으로 사용할 수 있다. 그림 6에 나타낸 여섯 가지 한글 유형에 따른 문자들의 구조는 그림 5에 도시된 영어 알파벳 쌍들의 구조와는 구분되는 특징들을 가지고 있다. 따라서 한글의 경우 한 문자의 넓이가 영어 소문자 두 개 또는 세 개를 합쳐 놓은 경우와 비슷하지만 그림 5의 정보로부터 이들을 구분할 수 있는 판단 기준을 얻을 수 있다.

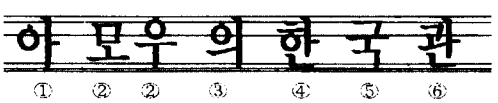


그림 6. 여섯 가지 기록 선상에 인쇄된 여섯 가지 한글 유형의 예
Fig. 6. Examples of six types of Hangul drawn on the six writing lines.

그림 7은 한글의 경우 한 글자가 수평 방향으로 분리되어 있어서 이들이 연속된 영어 알파벳 쌍으로 오분류될 수 있는 경우의 예를 나타낸 것이다. 유형 1과 3의 문자들은 왼쪽에 자음, 오른쪽에 모음의 형태로 결합되어 있다. 그림 7에서 한글의 유형 1과 3의 문자들은 문자 영역 내의 오른쪽에 수직 방향으로 긴 획이 존재하며 왼쪽에 위치한 자음이 오른쪽의 모음보다 길이가 짧은 특징이 있다. 따라서 이들의 기록 블록들은 연속된 영어 알파벳들의 기록 블록들과는 비교적 구분되는 특징을 가진다.

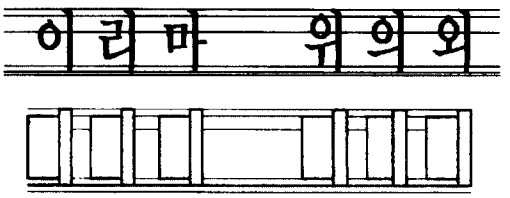


그림 7. 수평 방향으로 분리된 자소를 갖는 유형 1과 3 문자들의 기록 블록들의 예
Fig. 7. Examples of the horizontally partitioned Hangul of type 1 and type 3, and their writing blocks.

표 2. 한글의 각 유형별로 붙은 글자의 예
Table 2. Touching styles of contiguous pairs of the Hangul depending on the types.

C1 \ C2	Type1	Type2	Type3	Type4	Type5	Type6
Type1	가려	가수	아쉬	아침	아름	다웠
Type2	우이	소프	수위	구연	구름	구원
Type3	과배	과부	과위	화상	화통	화관
Type4	악어	악수	강쇠	망언	양군	양권
Type5	독야	독수	옥쇠	증언	훌륭	궁궐
Type6	황이	황우	황쇠	왕명	환웅	완월

그림 5, 6 및 7의 기본 특징들을 이용하여 한·영 혼용 문서에서 서로 분리된 문자들을 효과적으로 세그멘테이션해 낼 수 있다. 그러나, 실제 스캐너로 입력된 한·영 혼용 문서 영상에서는 이웃하는 문자들 사이에 붙은 경우가 많이 생기기 때문에 정확히 개별 문자를 세그멘테이션해 내기 위해서는 한글과 영문의 접촉 유형에 대한 분석이 요구된다. 표 2는 연속된 한글 문자들이 붙을 수 있는 유형별 예를 나타낸 것이다. 표에서 C1은 첫 번째 글자이고 C2는 두 번째 글자를 의미한

다. 한글은 여섯 개의 유형으로 구분되므로 세 개의 기록 블록으로 정의되는 영어에 비해 문자들 사이에 붙는 형태들이 더 다양하게 나타난다. 즉, 각 유형별로 두 개의 문자가 서로 인접해 있다면 표 2와 같이 36가지의 붙은 글자 형태들로 나타낼 수 있다. 붙은 글자 쌍의 첫 번째 문자는 'ㅏ', 'ㅑ'와 같이 수직 방향으로 긴 모음을 가지거나, 'ㅗ', 'ㅛ', 'ㅡ', 'ㅜ', 'ㅠ'와 같이 수평 방향으로 긴 모음을 가진다.

표 3. 영어의 기록 블록에 따른 붙은 알파벳들의 예

Table 3. Touching styles of contiguous pairs of English alphabets depending on their writing blocks.

C1 \ C2	Block A	Block B	Block C
Block A	qy gy	gu pa	gf qf
Block B	cq wy	ns ea	ak ei
Block C	fy fg	ta is	ff ft

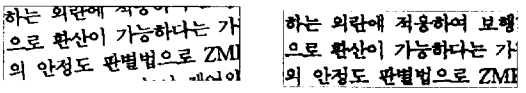
표 3은 영어 알파벳들이 가지는 기록 블록에 따라 붙어서 나타날 수 있는 알파벳 쌍의 예를 나타낸 것이다. 영어 알파벳들은 세 가지의 기록 블록 가운데 한 가지의 형태를 가지므로 표 3과 같이 기록 블록별로 붙은 알파벳 쌍은 모두 아홉 가지로 나타낼 수 있다.

한글이나 영문이 세 개 이상 연속해서 붙어 있는 경우에도 세그멘테이션 단계에서 좌측 첫 문자부터 연속된 두 개 문자 영역을 세그멘테이션 후보 영역으로 하여 문자를 추출하게 되므로 표 2 또는 표 3의 규칙이 동일하게 적용될 수 있다.

한글은 문자 세그멘테이션의 첫 단계에서 기록 블록의 형태 정보로부터 대부분 한글임을 알 수 있고 상호 붙어 있을 경우 표 2의 정보에 따라 이들을 해석할 수 있다. 그러나, 한글의 기록 블록 형태와 다른 기록 블록들이 있다면 표 3의 정보를 이용하여 붙거나 붙지 않은 연속된 영어 알파벳 영역인지를 구분해 낼 수 있다. 그리고 한글과 영문이 붙은 경우에도 좌측부터 전술한 기록 블록을 해석하여 차례대로 영문 영역인지 한글 영역인지를 구분해 낼 수 있다.

Ⅲ. 한글과 영어의 문자 영역 세그멘테이션 및 문자 인식

문자 분할의 첫 단계에서는 문서 영상에 대해 수평 및 수직 투영한 결과로부터 문자열을 찾아내고 이어서 단어 영역들을 찾아낸다. 문자열을 찾기 위해서는 수평 방향 화소값 투영 결과를 이용하여 배경 영역과 문자 영역을 구분하는 기법을 이용하면 된다. 문서 영상이 기울어져 있거나 휘어져 있을 경우에는 수평 방향 투영 결과만으로 문자열을 찾을 수 없고 왜곡에 대한 보정을 수행한 다음 문자열을 추적해야 된다[15-16]. 본 논문에서는 국소 영역을 레이블링해서 수평 방향의 후보 문자 영역을 추적한 다음 문서의 기울어진 각도를 계산해 내고 영상 변환을 통해 기울기 보정한 결과로부터 문자열을 추출하였다[16]. 그림 8은 기울어진 전자공학회 논문지의 문서 영상에 대해 기울기 보정한 결과를 도시한 예이다.



(a) (b)

그림 8. 기울어진 문서 영상을 기울기 보정한 결과
(a) 기울어진 문서 영상의 일부 (b) 기울기 보정이 된 문서 영상의 일부

Fig. 8. Rotation compensated document image for rotated document image.
(a) Rotated document image. (b) Rotation compensated document image.

개별 문자 세그멘테이션은 일단 문자열을 추출한 다음, 수직 투영 결과로 찾아낸 단어 영역에 대해 행해진다. 개별 문자 영역을 세그멘테이션하기 위해서 한글과 영어의 형태상 특징들을 이용하며, 블롭 컬러링은 단어 영역 안의 모든 블롭들을 구분해서 찾기 위해 사용된다.

본 논문에서는 블롭 컬러링의 결과를 해석하여 한 문자 영역을 정확히 세그멘테이션해 내기 위해 블롭 파라미터들을 정의하였다. 즉, 문자열에서 해당 블롭의 위치 및 수평, 수직 방향의 길이를 정량적으로 표시할 수 있도록 하였다. 그림 9는 "가"를 블롭 컬러링 하였을 때 블롭 "ㄱ"을 해석하기 위해 정의한 블롭 파라미터들의 예를 나타낸 것이다. 그림 9에서 Cs와 Ce는 블롭 컬러링된 한 문자의 시작점과 끝점을 나타낸다. Xs와 Xe 및 Ys와 Ye는 각각 한 문자에서 첫 번째 블롭의 x방향 시작점과 끝점, y방향 시작과 끝점을 나타내며, Wu와 Wn은 해당 블롭의 각각 상한과 하한의 여유 공백을 나타낸다. H와 W는 문자의 높이와 폭을 나타낸다.

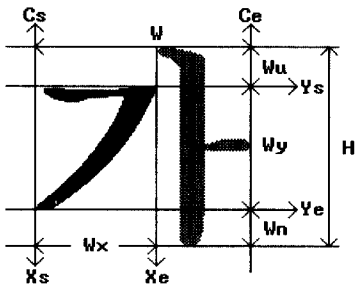


그림 9. 정의한 블롭 파라미터들의 실제 예
Fig. 9. Blob parameters defined to analyze the results of blob coloring.

한 문자를 세그멘테이션하기 위해서 "가"를 블롭 컬러링하였을 경우를 예시한 그림 9의 파라미터들을 이용하여 세그멘테이션 알고리즘을 구현하였다. 한 개의 한글 문자와 두 개의 인접한 영어 알파벳 쌍은 수평 폭의 길이가 비슷한 경우가 많으므로 분할 단계에서 상호간에 혼동될 수도 있으나, 이 경우에 각각의 블롭 파라미터 값들이 서로 다르게 설정되므로 이들을 구분할 수가 있다. 그림 5에서 설명한 것처럼 영어의 경우 첫 번째 블롭과 두 번째 블롭은 상한과 하한 여유 공백 중에 하나 또는 두 개가 서로 같거나 두 개 모두 같지 않은 경우로 나타난다. 그러나, 한글의 한 문자가 두 개의 수평 방향으로 인접한 자소 블롭들로 구성될 경우 연속된 두 개의 영어 알파벳과는 달리 첫 번째 블롭의 상한과 하한의 여유 공백이 모두 두 번째 블롭보다는 크다는 특징이 있다.

그림 10은 본 논문에서 정의한 블롭 파라미터들을 이용하여 개별 문자를 세그멘테이션하는 알고리즘을 나타낸 것이다. 세그멘테이션 알고리즘에서는 먼저 단어 영역을 추출하고 블롭 컬러링을 행한 다음 한 개의 한글 문자 영역을 구성하는 후보 블롭들을 다음과 같은 조건으로 선택하였다.

$$X^i s - Cs < H \qquad (1)$$

식 (1)에서 $X^i s$ 는 i번째 블롭이 시작하는 x점을 나타낸다. 식 (1)의 과정으로 선택된 후보 블롭들의 오른쪽 한계점 $X^i e$를 조사하여 해당 후보 블롭이 한 글자 영역 내에 있는지 또는 뒷 글자와 접촉되어 있는지의 여부를 판단하기 위해 다음과 같은 조건으로 선택하였다.

$$X^i e - Cs < H + \theta \qquad (2)$$

식 (2)에서 $H + \theta = W$이고, θ는 문서 영상에서의 실제 한글 한자의 폭 W와 한글의 수직 길이 H사이의 근사적인 차이 값이다. 식 (1)을 만족하는 블롭들이 식 (2)를 모두 만족할 경우에는 한 개의 한글 문자 영역으로서 뒷 글자와 접촉이 되지 않았거나 연속된 영문 알파벳으로 볼 수 있다. 한편, 식 (1)을 만족하는 블롭들 중에서 식 (2)를 만족하지 않는 블롭들이 존재할 경우 한글 문자 영역으로서 뒷 글자와 접촉되어 있거나 연속된 영문 알파벳으로 볼 수 있다. 따라서, 본 알고리즘에서는 한글에 대해 우선적으로 세그멘테이션을 시도하고 다음에 영문 알파벳에 대한 세그멘테이션 조건을 판단하였다. 블롭을 해석할 때 잡음 부분을 제거시키기 위하여 문자 높이의 5%이내의 화소 개수를 갖는 블롭들을 고려 대상에서 제외시켰다.

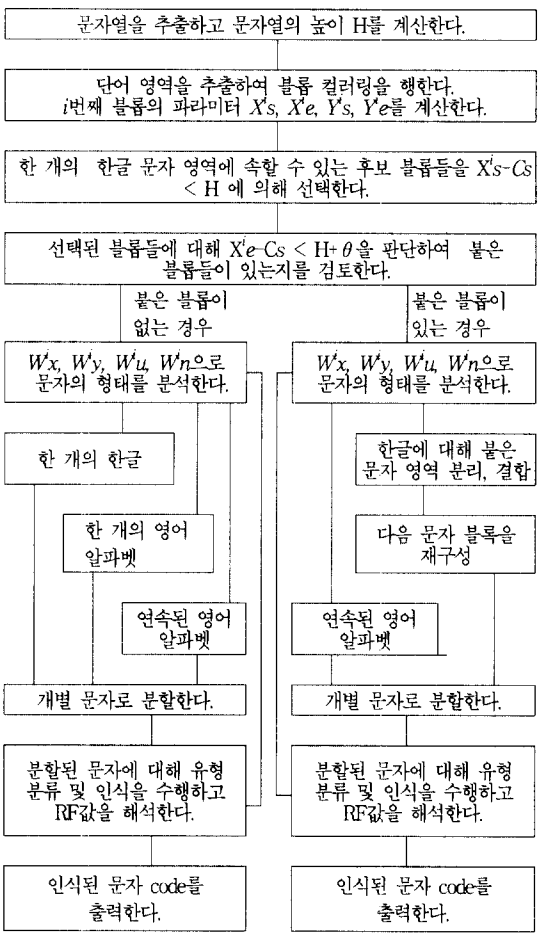


그림 10. 한·영 혼용 문서에서 개별 문자 분할 및 인식 알고리즘
Fig. 10. Character segmentation and recognition algorithm for the mixed Korean and English document.

그림 11은 서로 분리된 문자들에 대해 블롭 컬러링을 행한 결과를 나타낸 것이다. 그림에서는 15개의 블롭들이 서로 분리되어 있으므로 한 문자 영역 내에 있는 블롭들은 식 (1)과 (2)에 따라 선택할 수 있다.

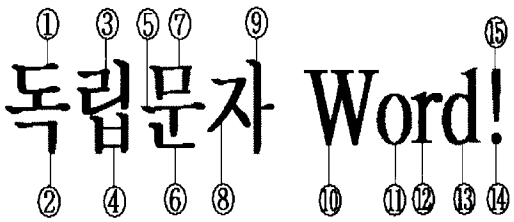


그림 11. 분리된 문자들에 대해 블롭 컬러링한 결과
Fig. 11. Examples of untouching characters after blob coloring.

식 (1)에서 선택된 블롭들이 식 (2)를 모두 만족할 경우 한글 한 문자나 영어의 한 알파벳 또는 알파벳들이 서로 붙어 있는 경우로 볼 수 있다. 알파벳들이 서로 붙은 경우의 블롭들이라면 그림 5의 정보에 따라 이들을 서로 분리시키도록 한다. 블롭들에 대한 해석을 통해 한 문자 영역이 분할되고 나면, 유형 분류기에서 문자의 유형을 결정하고, 문자 분류기에서 문자의 인식 코드를 출력하게 된다. 이들 두 분류기에서 산출된 신뢰도 지수(RF:이하 RF로 약함)값으로부터 문자 분할 및 인식이 제대로 이루어졌는지를 최종 결정하게 된다. RF값은 다음과 같이 정의된다.

$$RF = \frac{O_{first} - O_{second}}{O_{first}} \tag{3}$$

식 (3)에서 O_{first}와 O_{second}는 분류기의 최고 출력값과 두 번째 높은 출력값을 나타낸다. 식에서 최고 RF값은 1.0이 되며 실험 결과 RF값이 0.4이하이면 문자 영역이 잘못 분할되어 분류기에서 인식할 수 없는 글자로 판정한다. 분류기에서 산출된 RF값으로부터 문자가 잘못 분할된 것으로 판정되었을 때 문자의 형태를 분석하는 과정에서 좌우 방향 쉬프트를 하여 다시 분할을 시도하게 된다.

식 (1)을 만족하는 블롭들이 식 (2)를 만족하지 못하는 경우가 있을 경우에는 선택된 블롭들을 한글의 붙은 형태 또는 영어의 알파벳들이 서로 붙어 있는 형태로 보고 해석을 시도한다. 그림 12는 붙은 글자들에 대해 블롭 컬러링한 결과를 도시한 것이다.

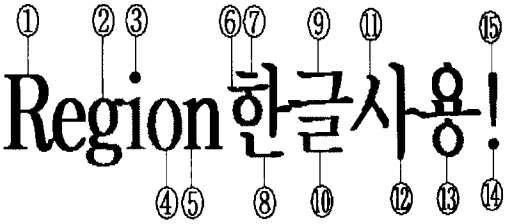


그림 12. 붙은 글자들에 대해 블롭 컬러링한 결과
Fig. 12. Examples of the touching characters after blob coloring.

그림에서 블롭 ①은 두 개의 알파벳으로 이루어져 있으며, "한글"은 ⑥에서 ⑩까지 다섯 개의 블롭으로 되어 있다. 따라서 이 블롭들의 블롭 파라미터값들과 기록 블록의 위치 정보에 따라 개별 문자 영역을 세그멘테이션해 내기 위한 결합 또는 분리를 시도한다. 한 개의 한글이 다음 글자와 접촉이 되었을 때는 표 2에 제시한 경우 중에 하나일 수가 있으므로, 이를 확인하여 붙은 영역을 분할할 수 있다. 표에서 한글이 접촉될 수 있는 경우는 모음이 "ㅏ" 인 경우나 모음이 "ㅡ" 인 경우이므로 블롭 파라미터 값을 이용하면 쉽게 판단할 수 있다. 한편, 연속된 영어 알파벳의 경우 그림 5에 제시한 기록 블록 정보를 이용하면 정확히 판단해 낼 수 있다. 붙은 블롭들에 대해 전술한 방법에 따라 한 개의 문자 영역을 분할한 다음 문자간에 서로 분리된 경우에서와 같은 방법으로 문자의 유형과 코드를 분류해 낸다. 마찬가지로 RF값을 이용하여 분할 결과가 옳은 지의 여부를 RF<0.4의 조건으로 판단하고 잘못 분할 되었을 경우 분할위치를 좌우방향으로 쉬프트시켜 다시 분할한다.

Ⅳ. 시뮬레이션 결과 및 고찰

제안된 방식에 따라 팬티엄 166MHz, 32MB의 개인용 컴퓨터에서 C언어를 이용하여 OCR 소프트웨어를 구현하고 문자 인식을 시도하였다. 5종류의 폰트(명조체, 신명조체, 고딕체, 중고딕체, 신문 명조체)에 대해 한글 1000자, 영어, 숫자 그리고 특수 기호 84자 등 전체 1084자를 각각 출력시켜 구한 전체 5420개의 글자를 유형 분류기와 문자 인식기에 학습시킨 다음 테스트 패턴을 이용하여 그 인식 성능을 분석하였다. 유형 분류를 위해 역전파 신경회로망을 이용하였고 메쉬피처[6]를 입력으로 사용하였다. 유형 분류기의 구성 및

학습 결과는 표 4와 같다.

표 4. 유형 분류기의 구조 및 학습 결과
Table 4. Structure of the type classifiers and its training results.

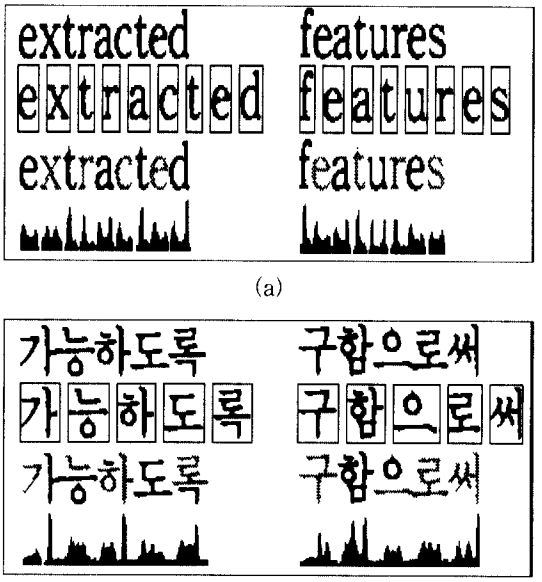
Classifier	Structure of the type classifiers			Training results	
	Input neuron	Hidden neuron	Output neuron	Average node errors	Total training rate(%)
Type classifier	256	35	7	0.0022	99.94

유형 분류기에서는 한글과 비 한글 1084개의 글자를 한글 여섯 종류 및 영문, 숫자, 특수 기호 한 종류 등 일곱 개의 유형 중에 하나로 구분해 낼 수 있도록 하였다. 한편, 유형별로 개별 문자 인식을 위한 문자 인식기로도 역전파 신경회로망을 사용하였으며 메쉬피쳐를 입력으로 하였다. 표 5는 유형별 문자 인식기의 구조 및 학습 결과를 도시한 것이다.

표 5. 유형별 문자 인식기 구조 및 학습 결과
Table 5. Structure of the character recognizer and its training results.

Recognizer	Structure of the character recognizer			Training results	
	Input neuron	Hidden neuron	Output neuron	Average node errors	Total training rate(%)
Type 1 recognizer	256	35	115	0.0006	100.00
Type 2 recognizer	256	30	70	0.0011	100.00
Type 3 recognizer	256	25	50	0.0013	100.00
Type 4 recognizer	256	40	450	0.0002	99.95
Type 5 recognizer	256	35	265	0.0003	99.98
Type 6 recognizer	256	25	50	0.0014	100.00
Type 7 recognizer	256	30	84	0.0012	99.60

표 4 및 5에서 평균 노드 에러는 전체 5420개의 글자에 대한 출력 노드 뉴런들의 에러 값의 합을 평균한 것이고 학습률은 학습 패턴에 대한 인식률을 나타낸 것이다. 성능 분석에 사용된 테스트 데이터는 대한전자공학회지 KITE(Korean Institute of Telematics and Electronics)에서 임의로 선택한 5쪽의 문서를 HP IIc 스캐너로 300dpi해상도에서 입력하여 만들었다. 그림 13은 테스트에 사용된 데이터의 세그멘테이션, 블롭 컬러링 및 수직 방향 투영 결과의 일부 예를 도시한 것이다. 그림13의 (a)는 영문 데이터의 세그멘테이션, 블롭 컬러링 및 수직 방향 투영 결과 예를 도시한 것으로서 실제로는 붙어 있지 않은 알파벳들이 수직 방향 투영 결과에서는 붙어 있으나, 이를 정확히 세그멘테이션해 낼 수 있음을 확인할 수 있다. 또한 그림 13의 (b)는 한글의 경우로서 수직 방향 투영 결과에서는 대부분 붙은 글자로 보이지만 본 알고리즘으로 정확히 세그멘테이션 해냄을 알 수 있다.



(a)

(b)

그림 13. 테스트 데이터의 일부 예, 세그멘테이션 결과 및 수직 방향 투영 결과
(a) 영문 데이터의 세그멘테이션, 블롭 컬러링 및 수직 방향 투영 결과 예 (b) 한글 데이터의 세그멘테이션, 블롭 컬러링 및 수직 방향 투영 결과 예
Fig. 13. Examples of testing data, segmentation results, and vertical projection results.
(a) Segmentation, blob coloring, and vertical projection results of English testing data. (b) Segmentation, blob coloring, and vertical projection results of Hangul testing data.

제안된 알고리즘을 전자공학회 논문지에서 임의로 선정한 특정 논문의 영문 요약 부분에 적용하여 인식 실험한 결과를 표 6에 도시하였다.

영문 요약 부분 인식 결과에서 붙은 글자로 보이는 부분은 블롭 컬러링 결과로 모두 정확하게 분리해 낼 수 있었으며, 세그멘테이션 에러 글자들은 실제 영상에

서 인접한 글자들과 붙은 글자들로서 한·영 혼용 세그멘테이션 규칙을 적용하였기 때문에 대부분 한글 영역으로 오분할되어서 나타난 것이다. 전체 인식 결과는 세그멘테이션 에러와 문자 인식 에러를 제외하고 올바르게 인식된 글자들의 개수로 계산된 값이다.

표 6. 전자공학회지의 영문 요약 부분에 대한 인식 결과

Table 6. Recognition results for English abstract of the journal of the KITE.

Total no. of characters	No. of touching characters after vertical projection	No. of touching characters for real image	No. of segmentation error characters	No. of recognizer error characters	Total no. of recognizer characters
819	105	37	14	12	793
rate(%)	12.8	4.5	1.2	1.5	96.8

한편 제안된 알고리즘을 전자공학회 논문지에서 임의로 선정한 논문의 한·영 혼용된 본문에 적용하여 인식 실험한 결과를 표 7에 도시하였다. 테스트용 문서에서 그림이 없고 비교적 글자들로만 이루어진 쪽의 글자 수는 평균 1700자 정도가 되며 그림이 있는 경우에는 글자 수가 상당히 줄어들게 된다. 하지만 투영 결과로 붙어 보이는 글자 수나 실제 붙어 있는 글자 수는 각 쪽마다 전체 글자 수에 관계없이 비슷하게 나타난다는 것을 알 수 있었다.

표 7. 전자공학회지의 본문 내용에 대한 인식 결과

Table 7. Recognition results for the bilingual parts of the journal of the KITE.

Total no. of characters	No. of touching characters after vertical projection	No. of touching characters for real image	No. of segmentation error characters	No. of recognizer error characters	Total no. of recognizer characters
1741	536	210	23	15	1703
rate(%)	30.7	12.1	1.3	0.9	97.8

본문 내용의 경우 한·영 혼용이지만 한글이 다수를 차지하고 있으며, 이때 투영 결과 붙은 것으로 보이는 글자나 실제 영상에서 붙은 글자가 영문 전용에 비해 상당히 많이 나타났다. 즉, 한·영 혼용 문서에서 글자들의 간격을 일정하게 할 경우 한글이 영문에 비해 붙은 글자들이 더 많음을 알 수 있다. 본 논문에서 테스트 패턴으로 사용한 전자공학회지 본문내용을 상용 OCR인 글눈에서 인식시켜 본 결과 약 94%의 인식률을 얻을 수 있었다. 또한 상용 OCR이 초당 50자 이상을 인식할 수 있는데 비하여 본 실험 결과에서는 초당 40자 정도를 인식할 수 있었다.

제안된 알고리즘은 영문에 비해 한글의 세그멘테이션 및 인식 능력 제고에 우선 순위를 부여하여 작성하였으므로 한글 문서에 적용한 결과가 붙은 글자 수가 영문에 비해 많음에도 불구하고 비교적 높은 인식률을 나타낸다는 것을 확인할 수 있다.

실험 결과에서 참고문헌 인용 표시 기호는 고려 대상에서 제외시켰고 "."와 "," 에 대한 인식 오차는 고려하지 않았다. 한글의 경우 대부분 정확히 세그멘테이션 되었으나, "가능"과 같이 심하게 붙은 글자에 대해서는 두 글자 사이의 절단 점이 왼쪽으로 약간 이동되어서 "기" 와 "능"으로 잘못 인식되는 경우도 있었다. 반면, 영어의 경우 세 개 또는 그 이상 붙은 글자들에 대해서 세그멘테이션해 내기가 어려웠으며, 이는 연속적으로 붙어 있는 알파벳들이 한 개의 알파벳으로 보여지기 때문이었다. 예로써, 인식 시스템은 "r"과 "n"이 서로 붙어 있을 때는 "m"으로, "w"를 "v"와 "v"로 잘못 분류하는 경우가 있었다. 따라서 한글뿐만 아니라 영어의 경우 후처리 단계에서 단어 사전이나 문맥 사전 검사 방법을 추가하면 좀 더 높은 인식률을 얻을 수 있을 것이다.

V. 결 론

본 논문에서는 인쇄체 한·영 혼용 문서에서 붙은 글자들까지도 효과적으로 세그멘테이션하여 정확히 인식해 낼 수 있는 문자 인식 알고리즘을 제안하였다. 한·영 혼용 문서에서 개별 문자 영역을 정확히 세그멘테이션하기 위해 문자열에서의 한글 및 영문의 수직·수평 위치 정보를 기록 블록으로 정의하였고, 패턴 분류기의 분류 결과의 신뢰도 지수값을 이용하여 오인식을 최소화시킬 수 있는 결과 판단 기준을 마련하였다.

제안된 알고리즘의 문자 영역 세그멘테이션 및 문자 인식 성능을 확인하기 위해 대한전자공학회지에서 임의로 선택한 5쪽의 문서를 HP IIc 스캐너로 300dpi해상도에서 입력한 다음 문자 인식을 시도하였다. 실험 결과 영문 요약 부분의 경우 96.8% 정도의 인식률을 얻을 수 있었으며 한·영 혼용된 본문에 대해서는 약

97.8%정도의 인식률을 얻을 수 있었다. 결과에서는 제안된 알고리즘을 이용하여 개별 문자 영역 세그멘테이션에서 초래되는 에러들을 상당 부분 줄일 수 있음을 확인하였다. 또한 문서 영상에서 한글이 영문에 비해 실제 붙은 글자들의 수가 더욱 많았지만, 제안된 알고리즘은 한글에 우선권을 부여하여 작성되었기 때문에 이와 같이 붙은 한글들을 매우 효과적으로 세그멘테이션해 내고 인식해 낼 수 있음을 확인하였다.

일부 글자들의 경우 단독 세그멘테이션 결과만으로는 정확히 인식해 낼 수 없었으며 이들을 인식하기 위해 후처리 단계에서 단어 사전이나 문맥 사전 검사 방법을 추가하기 위한 연구가 요구된다.

참 고 문 헌

[1] S. Kiang, M. Shridhar, and M. Ahmadi, "Segmentation of touching characters in printed document recognition," *Pattern Recognition*, vol. 27, no. 6, pp. 825-840. 1994.

[2] J. Wang and J. Jean, "Segmentation of merged characters by neural networks and shortest path," *Pattern Recognition*, vol. 27, no. 6, pp. 649-658. 1994.

[3] S. Wood, X. Yao, K. Krishnamurthi, and L. Dang, "Language identification for printed text independent of segmentation," *IEEE Int'l Conference on Image Processing*, Washington, D.C. vol. 3, pp. 428-431, 1995.

[4] R. M. Bozinovic and S. R. Srihari, "Off-line cursive script word recognition," *IEEE Trans. Pattern Analysis Mach. Intell.* vol. 11, pp. 68-83, 1989.

[5] C. Huang and H. Lee, "Increasing character recognition accuracy by detection and correction of erroneously identified characters," *Pattern Recognition,* vol. 27, no. 9, pp. 1259-1266. 1994.

[6] S. I. Chien, "Hangul(Korean) and English OCR system using multiple hypotheses driven neural nets," *Proc. of Korean-French Character Recognition Workshop,* pp. 37-52, 1994.

[7] S. Shlien, "Multifont character recognition for typest documents," *IEEE Trans. Pattern Analysis Mach. Intell.* vol. 2, no. 4, pp. 603-620, 1988.

[8] H. Takahashi, "A neural network OCR using geometric and zonal pattern features," *ICDAR '91*, Paris, vol. 2, pp. 821-828, 1991.

[9] K. S. Fu and A. Rosenfeld, "Pattern recognition and computer vision," *IEEE Computer,* vol. 17, no. 10, pp. 274-282, Oct. 1984.

[10] J. Wang and J. Jean, "Multiresolution neural networks for omnifont character recognition," *IEEE Int. Conf. on Neural Networks,* vol. 2, no. 5, pp. 1588-1593, 1993.

[11] P. Luca and A. Gisotti, "Printed character preclassification based on the word structure," *Pattern Recognition,* vol. 24, pp. 609-615, 1991.

[12] C. Wells, L. Evett, P. Whitby, and R. Whitrow, "Fast dictionary look-up for contextual word recognition," *Pattern Recognition,* vol. 23, pp. 501-508, 1990.

[13] H. Takahashi, N. Itoh, T. Amano, and A. Yamashita, "A spelling correction method and its application to an OCR system," *Pattern Recognition,* vol. 23, pp. 363-377, 1990.

[14] S. Kahan, T Pavilids, and H. Baird, "On the recognition of printed characters of any font any size," *IEEE Trans. Pattern Analysis Mach. Intell.,* vol. 9, no. 2, pp. 274-287, 1987.

[15] S. W. Lee, "Nonlinear shape restoration of distorted images with Coons transformation," *IEEE Int'l Conference on Document Analysis and Recognition,* Montreal, Canada, vol. 1, pp. 235-238, 1995.

[16] Y. M. Baek, "Layout understanding of documents using effective region labeling," 경북대학교 전자공학과 석사학위논문, 1994.

[17] Z. Kovacs and R. Guerrieri, "Massively parallel handwritten character recognition based on the distance transform," *Pattern Recognition,* vol. 28, no. 3, pp. 293-301. 1995.

저 자 소 개

金 桂 慶(正會員)

1966년 10월생. 1989년 2월 경북대학교 전자공학과 졸업(공학사). 1992년 2월 경북대학교 전자공학과 석사과정 졸업(공학 석사). 1993년 3월 ~ 1996년 11월 현재 경북대학교 전자공학과 박사과정 재학중. 주관심분야는 문자인식, 패턴인식 등임.

秦 成 一(正會員) 第 28卷 B編 第 12號 參照

金 鎭 浩(正會員) 第 33卷 B編 第 2號 參照

崔 興 文(正會員) 第 33卷 B編 第 2號 參照