

오프라인 필기체 한글 인식을 위한 자소 내 자획의 분리

정민철^{1*}

Stroke Extraction in Phoneme for Off-Line Handwritten Hangul Recognition

Minchul Jung^{1*}

요약 본 논문은 오프라인 필기체 한글 인식을 위한 요소 기술의 하나인 자소 분할을 위한 새로운 자획 추출법을 제안한다. 수평 런 길이를 이용하여 자소의 자획을 수직, 경사, 수평으로 구분 분리한다. 수직 자획이나 경사 자획의 수평 런 길이는 자획 두께가 되며, 수평 자획의 수평 런의 개수가 자획 두께가 된다. 수평 자획을 분리 추출한 후, 끊어진 수직, 경사 자획을 자획 두께의 수평 런으로 연결하여 분리한 자획들이 문자의 특징을 나타내게 한다. 추출된 자획들은 온라인 필기체 한글 인식 시스템에서 개발 사용되고 있는 자획 사전 정합을 통해 문자 인식을 할 수 있다.

Abstract This paper proposes a new stroke extraction algorithm for phoneme segmentation, which is one of main techniques for off-line handwritten Hangul recognition. The proposed algorithm extracts vertical, slant, and horizontal strokes from phonemes using run-length. The run-length of vertical or slant strokes becomes the width, and also the number of horizontal run-lengths the width. After extracting horizontal strokes from phonemes, the algorithm links two continuous vertical or slant strokes with run-lengths of the strokes' width to represent the features of a character. The extracted strokes can be utilized to recognize a character, using template matching of strokes, which is being adopted in on-line handwritten Hangul recognition.

Key Words : Stroke Extraction, Off-line Handwritten, Hangul Recognition. OCR.

1. 서론

디지털 카메라, 스캐너, 태블릿 등을 통해 입력 받은 문자 영상을 컴퓨터로 처리하여 문자를 자동으로 컴퓨터가 인식하는 광학 문자 인식(OCR: Optical Character Recognition) 기술은 키보드에 의한 문자 입력 방식을 보완, 대체할 새로운 기술 중 하나로써 그 사용과 응용 범위가 점차 확대되고 있다. 문자 인식 기술은 문자 영상정보를 획득하는 방법에 따라 온라인 문자 인식과 오프라인 문자 인식으로 나누어진다. 키보드를 장착할 수 없는 PDA 등에서 널리 사용되는 온라인 문자 인식은 전자펜으로 손쉽게 글이나 도형 기타 제스처 등을 입력할 수 있게 해주는 기술로써 더 이상 키보드의 어려운 자판을 암기할 필요가 없다. 온라인 문자인식은 필기체를 다루며 키보드보다 입력 속도가 느린 단점이 있다. 대용량의 문

서를 스캐너나 디지털 카메라를 통해 영상으로 입력 받아 처리하는 데 사용되는 오프라인 문자 인식은 많은 문자들을 빠르게 컴퓨터에 입력 처리할 수 있게 해주는 기술로써 더 이상 지루하고 느린 키보드 수작업을 할 필요가 없다. 이러한 것을 오프라인 문자인식이라 하며, 이것은 인쇄체와 필기체를 다룬다. 오프라인 인쇄체 문자인식은 워드 작업한 파일을 프린트하는 작업의 역이라 할 수 있다. 인쇄체 문자 인식은 기계에 의한 한정된 활자크기, 폰트 특성 등에 의해 높은 인식률은 보이고 있는 반면, 필기체 문자 인식은 복잡한 변형과 필체 등에 의해 미미한 인식률을 나타낸다. 더구나 필기방향이나 필기속도, 획순, 획수 등의 동적 정보를 이용할 수 있는 온라인 필기체 문자 인식에 비해, 공간적 바깥인 정적 정보만을 이용할 수 있는 오프라인 필기체 문자인식은 인식률을 높이기 위해서 아직 풀어야 할 많은 문제점이 존재한다. PDA 등에 사용되는 온라인 필기체 한글 인식 제품은 이미 상품화된 것도 있으나, 오프라인 필기체 한글 인식은

¹상명대학교 컴퓨터시스템공학과

*교신저자: 정민철(mjung@smu.ac.kr)

이 된다. 각 자획은 이미 접합되어 있고, 자소 또한 종종 접합되어 있으며, 심지어 입력 문자가 이웃 문자와 서로 접합되어진 경우도 많다. 본 논문에서는 이를 해결하기 위해 한글의 자소 분리에서 한 단계 더 내려가 자소를 자획으로 분리한다.

한글 자소는 방향성이 강한 자획들로 구성되어 있다. 본 논문에서는 한글의 모든 자획을 0도(수평성분), 90도(수직성분), 45도와 135도(경사 성분), 360도(폐곡선 성분)의 방향을 가지는 표준 자획으로 분류한다. 수직성분과 경사성분은 수평 런 길이가 자획의 두께가 됨으로 자획 분리에서는 하나의 성분으로 간주 할 수 있다. 이러한 방향성을 이용해 분리한 자획의 조합으로 자소를 형성하고, 형성된 자소로 하나의 문자를 구성하여 인식하는 계층적 패턴 인식을 한다. 자획의 조합으로 형성된 초성(FS), 중성(MS), 종성(LS)으로부터 유니코드 문자를 구성하는 식은 아래와 같다.

$$Unicode = 44032 + (FS \times 21 \times 28) + (MS \times 28) + (LS) \quad (1)$$

유니코드는 컴퓨터에서 한글을 표현하는 방법으로, 조합될 수 있는 한글 11,172개의 글자 모두를 가나다순으로 정렬, 배치한 것이다.

3. 문자 영상의 전처리

그레이 레벨로 스캔된 영상은 첫 번째로, 히스토그램 스트레칭(Histogram Stretching) 기법을 사용하여 영상의 대비를 개선한다. 본 논문에서 사용하려는 명암 대비 스트레칭[4]은 영상 히스토그램이 모든 범위의 화소값을 포함하도록 식(2)에 의해 영상을 확장한다.

$$new\ pixel = \frac{old\ pixel - low}{high - low} \times 255 \quad (2)$$

위 식(2)에서 'high'는 영상 히스토그램에서 가로축 최대치를 나타내고 'low'는 가로축 최소치를 나타낸다. 두 번째로 배경(background)인 종이 영상으로부터 관심 대상이 되는 전경(foreground)인 문자 영상을 분리하기 위해 이진대비개선(binanzation)을 한다. 이 과정은 배경인 종이에 색이 있거나 얇은 무늬가 있을 때, 전경인 문자가 파란색 필기구나 연필 등으로 필기되었을 때 특히 중요하다. 만약 이진화 임계치가 적절한 값보다 크면 배경이 잡음 형태로 이진 영상에 출력되며, 임계치가 적절한 값보다 작으면 검은색 전경인 문자의 자획이 끊어지게 된다. 이 단계에서 끊어진 자획은 문자의 특징을 모호하게

하여 인식을 대단히 어렵게 한다. 끊어진 자획보다는 접합된 자획을 처리하는 것이 본 논문의 목적이므로, 임계치를 크게 한다. 세 번째로 높은 임계치로 인한 이진 영상 배경의 임펄스 잡음을 제거하기 위해 중간값 필터(median filter)[4]를 사용한다. 임펄스 잡음은 그 주변의 화소에 비해 월등히 밝거나 어둡기 때문에 일반적으로 3 X 3 그룹의 순위에서 최하위나 최상위에 위치한다. 3 X 3 그룹에서 중간값을 선택하여 임펄스 잡음을 영상에서 제거한다.

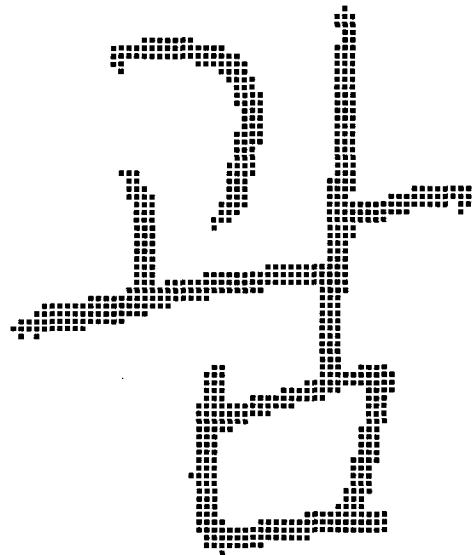


그림 2. 두 개의 연결 성분으로 구성된 문자

4. 문자의 자획 분할

동적 정보가 없는 단순한 정적 영상의 화소 정보만을 이용할 수 있는 오프라인 필기체 문자인식에서 문자의 필기 순서(획순)와 유사한 순서를 인위적으로 복원하기 위해서는 화소 정보를 보다 의미 있는 상태로 변환하여야 한다. 전처리된 영상은 아직 영상 정보의 최하위 계층인 화소 수준에 머물러 있다. 본 논문에서는 이를 연결 성분 분석(connected component analysis)을 통해 보다 상위 계층으로 나타낸다. 연결 성분 분석은 위에서 아래로, 좌에서 우로 영상을 스캔하면서 문자 전경(foreground)들 중 연결된 성분들을 찾아내는 것이다. 연결 성분을 분석하게 되면 화소들은 연결된 성분 덩어리(blob)들로 나타난다. 각 연결 성분 덩어리는 구성하고 있는 화소들의 수평 런 길이(run length), 즉 런의 시작점과 끝점의 좌표에 의해 나타낼 수 있다. 따라서 연결 성분 분석은 개개의

연결된 성분 덩어리마다의 크기와 위치를 나타낼 수 있으며, 이는 문자의 자획 결정에 중요한 정보로 사용된다. 이상적인 경우에는 초성, 중성, 종성은 각각 다른 한 개의 연결 성분이 되지만 실제로는 그림 2에서 보듯이 하나의 연결 성분에 초성과 중성 또는 중성과 종성 등이 접합되어 있다. 이를 분리하기 위해 문자의 자획 두께를 이용해 수직, 수평, 사선 등의 자획을 추출 분리한다. 즉, 본 논문에서 제안하는 새로운 자획 분리 방법은 그림 2의 문자를 예로 들면 첫 번째 연결 성분인 ‘ㄱ’은 ‘一’ 과 ‘|’ 의 2 개 자획으로, 두 번째 연결 성분인 ‘나’ 과 ‘口’의 접합은 8개의 자획인 ‘|’ 과 ‘一’ 으로 분리한다. 이를 위해 먼저, 수평 런 길이를 문자 영상 내에서 모두 구한다. 이를 가로축은 “수평 런 길이”로, 세로축은 “수평 런 길이의 빈도수”로 하는 히스토그램에 나타낼 수 있다. “수평 런 길이의 빈도수”는 아래 식 (3)에 따라 구할 수 있다.

$$histogram[runlength] \leftarrow histogram[runlength] + 1 \quad (3)$$

그 다음으로, 수평 런 길이 히스토그램의 빈도수에서 최대값을 가지는 수평 런 길이를 입력 문자 자획의 두께 w 로 한다. 등식 (4)는 이를 나타낸다.

$$w = \max_{runlength=3}^N (hsitogram[runlength]) \quad (4)$$

수평 런 길이가 2이하인 것은 이진화 과정에서 생긴 잡음이다. 따라서 수평 런 길이에서 제외하며, 위 식 (4)에서 $runlength = 3$ 은 이를 나타낸다. 일반적으로 수직이나 경사 자획의 가로 두께(=수평 런 길이)는 w 가 되며 수평 자획의 가로 두께(=자획의 길이)는 w 보다 훨씬 크고, 그 세로 두께가 w 가 된다.

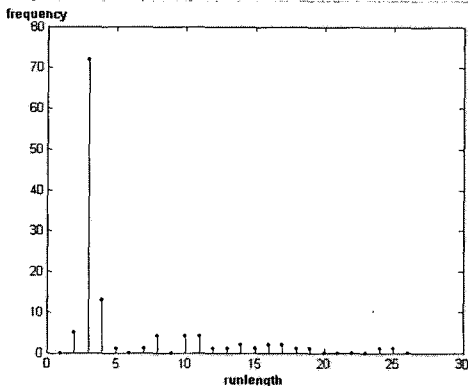


그림 3. ‘ㅁ’ 문자의 수평 런 길이 빈도수 히스토그램

그림 2에서 보인 ‘ㅁ’ 문자는 그림 3에서 같이 수평 런 길이 빈도수 히스토그램으로 나타낼 수 있다. 그림의 예에서는 모두 117개의 수평 런이 있으며(수평 런 길이 2이하의 제외), 그 중 수평 런 길이가 3인 것은 72개, 4인 것은 13개이다. 따라서 주어진 문자의 자획 두께 w 는 3에서 4가 된다. 수평 런 길이가 자획 두께 w 와 같으면 그 수평 런은 수직 자획 또는 경사 자획에 속하며 수평 런 길이가 자획 두께 w 보다 크면 그 수평 런은 수평 자획에 속한다. 수평 런 길이에 의하면 수직 자획과 경사 자획은 동일하다. 그 다음 단계로 문자의 자획 두께 w 를 가지고, 입력 문자의 연결 성분을 수평으로 주사한다. 그림 4는 수평 자획과 수직 자획이 만나는 시작점(y_1)과 수직 자획과 수평 자획이 만나는 시작점(y_2)을 나타낸다. 수평 런 길이가 자획 두께 w 보다 훨씬 크다가-이러한 수평 런은 수평 자획의 두께 만큼 반복한다-자획 두께 w 와 같아지는 부분이 수평 자획과 수직 자획이 만나는 시작점이다 (y_1). 수평 런 길이가 자획 두께 w 와 같다가 자획 두께 w 보다 급격이 커지는-커진 수평 런은 수평 자획의 두께 만큼 반복한다-부분은 수직 자획과 수평 자획이 만나는 시작점이다(y_2). y_1 과 y_2 에서 자획이 분리, 추출된다. 수평 자획의 런 길이는 y_2 에서 시작해 y_1 에서 끝나거나 y_1 (무교점)에서 끝나며, $|y_2 - y_1|$ 가 수평 자획의 두께 w 가 된다. 수평 자획의 런 길이가 y_2 에서 시작해 y_1 에서 끝나는 경우는 수평 자획과 수직 자획의 교점 영역을 형성하는데, 이때 수평 자획을 분리 추출하고 나면 수직 자획은 연결이 끊어지게 된다. 따라서 $|y_2 - y_1|$ 의 공간을 자획의 두께 w 의 수평 런으로 메워 끊어진 수직 자획을 이어준다. 그 결과, 그림 4에서 ①, ②, ③은 하나의 수직 자획이 된다. 또한 ④와 ⑤도 하나의 수직 자획이 된다. 즉, 방향이 다른 두 개의 자획이 겹치는 영역을 한 자획에만 분리하여 포함시키는 것이 아니라 두 개의 자획 모두에 포함시킨다. 이 영역은 문자의 고유 특징으로 문자 인식 시스템의 다음 단계에서 이용될 수 있다.

5. 실험 결과

본 논문에서 제안한 방법을 실험하기 위해 5명의 대학생이 쓴 필기체 한글을 각각 200dpi와 300dpi의 그레이 스케일로 스캔하여 영상을 획득하였다. 스캔하는 샘플링 수에 따라 영상 내 문자 자획의 두께가 달라지는 데, 제안한 방법은 이와 관계없이 문자 자획의 두께를 모두 구하였다. 자획의 분리 결과는 사람이 직접 자획 분리된 영상을 보고 분리 여부의 가부를 분석하였다. 그림 5는 그

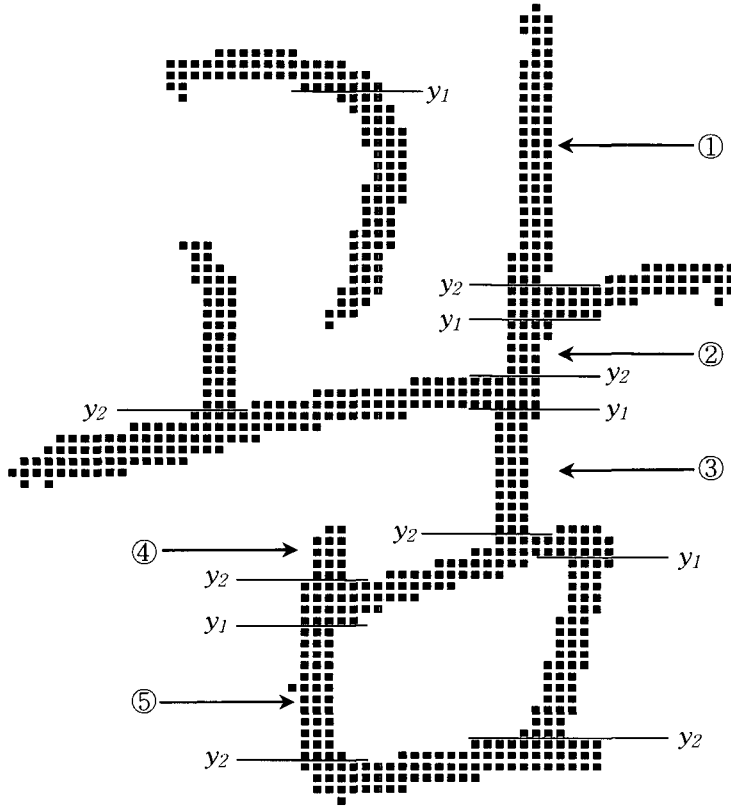


그림 4. 수평 자획과 수직 자획의 시작 교점(y_1)과 수직 자획과 수평 자획의 시작 교점(y_2)

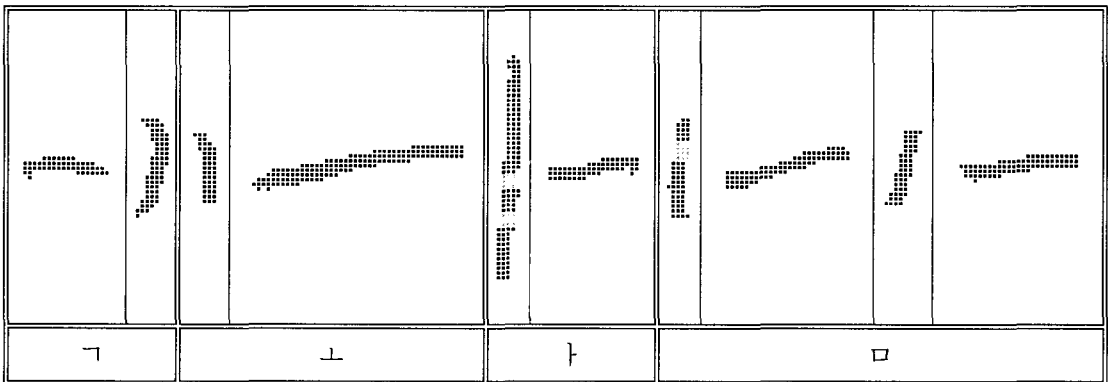


그림 5. 'ㅍ' 문자의 자획 추출

림 2에서 보인 'ㅍ' 문자를 자획으로 모두 분리한 것을 보인다. 회색 부분이 $|y_2 - y_1|$ 의 공간을 자획의 두께 w 의 수평 런으로 채워준 곳이다. 초성 자음 'ㄱ'은 'ㄱ'과 'ㄱ'으로, 중성 복모음 'ㅏ'는 'ㅏ', 'ㅑ', 'ㅓ'과 'ㅓ'으로, 종성 자음 'ㅍ'은 'ㅍ', 'ㅑ', 'ㅓ'과 'ㅓ'으로

자획 분할되었다. 분할된 결과만 보면 중성 복모음 'ㅏ'과 종성 자음 'ㅍ'은 동일하다. 그러나 종성 자음 'ㅍ'의 자획은 상단부에서 두 개의 접촉 지점 y_1 과 하단부에서 두 개의 접촉지점 y_2 를 가진다. 또한 두 개의 y_1 은 y_2 좌표에 대해 근사 값을 가지고, 두 개의 y_2 또한 근사 값을 가

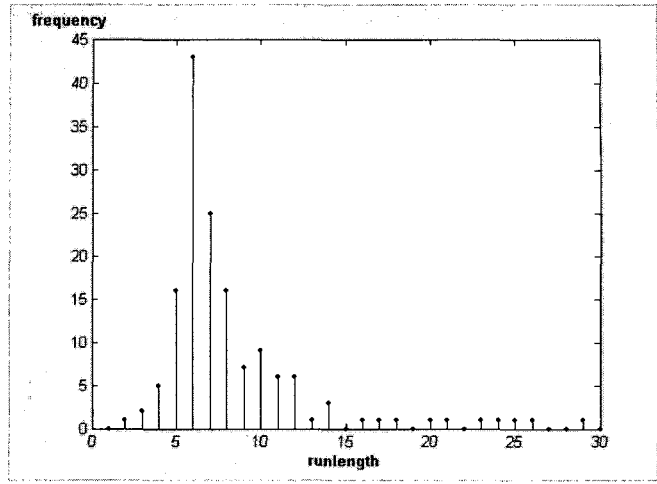
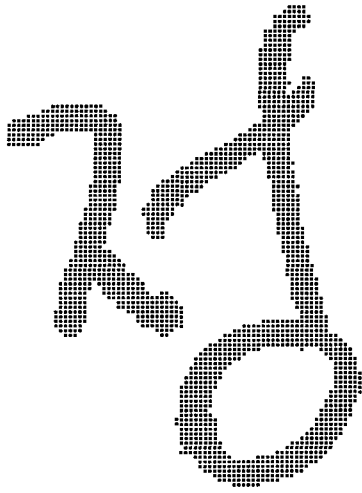


그림 6. '정' 문자와 수평 런 길이 빈도수 히스토그램

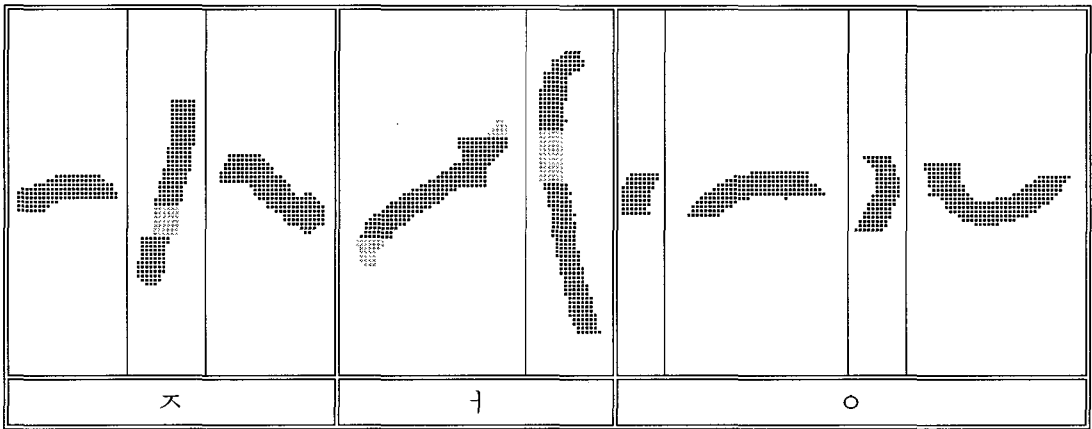


그림 7. '정' 문자의 자획 추출

진다. 네 개의 자획이 이러한 조건을 만족시키면 네 개의 자획은 폐곡선을 이룬다고 판단하고, 문자 인식 시스템의 다음 단계에서 네 개의 자획을 합성해 하나의 자소로 만든다.

그림 6은 '정' 문자와 수평 런 길이 히스토그램을 보인다. 그림에서 보듯이 모두 150개의 수평 런이 있으며(수평 런 길이 2이하의 제외), 그 중 수평 런 길이가 6인 것은 43개, 7인 것은 25개이다. 따라서 주어진 문자의 자획 두께 w 는 6에서 7이 된다. 그림 7은 '정' 문자를 자획 분할한 것을 보인다. 중성 모음 'ㅣ'의 수평 자획의 양끝 부분(회색 부분)은 수평 런 길이가 수직 자획의 두께와 동일하여 일단 수직 자획으로 분할되나, 수평 런의 개수가 수직 자획을 구성하기에 적어 분할이 취소되고 수평 자획의 일부만이 된다. 중성 자음 'ㅇ'은 네 개의 자획으로 분할되나 'ㅁ'과 마찬가지로의 조건을 만족시킴으로,

문자 인식 시스템의 다음 단계에서 하나의 폐곡선으로 합성되어 하나의 자소가 된다.

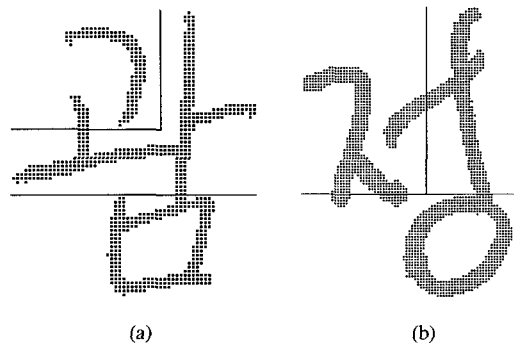


그림 8. 한글 자소 조합 형태에 따른 자소 분리

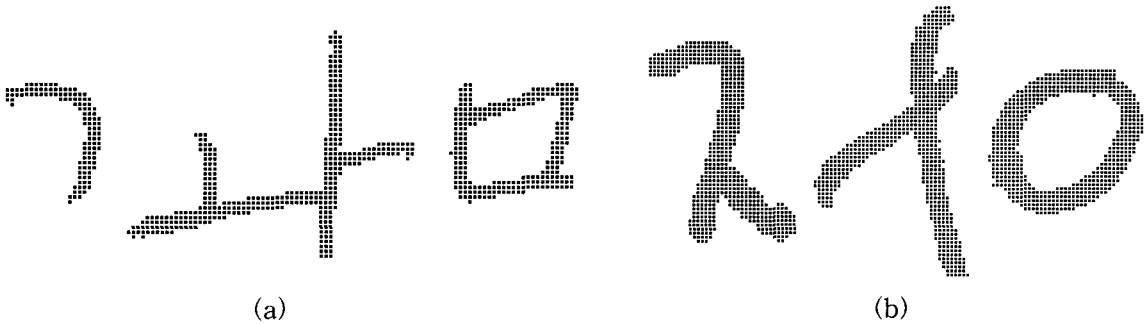


그림 9. 자획 분리를 통한 자소 분리

그림 8은 기존의 한글 자소 조합 형태에 따른 자소 분리선을 나타낸다. 그림 8(a)는 그림 1에서 제시된 유형 6에 의한 분리선을 보이고, 그림 8(b)는 유형 4에 의한 분리선을 보인다. 그림에서 보듯이 필기체 한글의 경우 분리 최적 위치가 변동됨으로 인접 자소의 단편이 포함되고 자소 일부분이 잘려 나간다. 그림 9는 본 논문에서 제시한 자획 분리를 통한 자소 분리 결과를 나타낸다. 한글 자소 조합 형태에 관계없이 자소의 원형대로 잘 분리됨을 알 수 있다.

6. 결론

본 논문에서는 오프라인 필기체 한글 인식을 위한 요소 기술 중 하나인 자소 분리를 위한 새로운 자획 분리 기술을 제안하였다. 우리가 필기체 한글을 쓸 때 자획을 바탕으로 초성 자음, 중성 모음, 종성 자음의 순서로 조합하여 한글 문자를 구성하여 쓰는 것처럼, 문자 인식은 필기 과정의 역으로 조합된 문자를 자소로 분리하고, (접합) 자소를 자획을 분리하여 구조적으로 수행한다. 이는 계층적 구조를 갖는 한글 문자의 특성을 최대한 활용한 것이다. 또한 본 논문이 제안한 자획 분리 기술은 오프라인 필기체 한글 문자를 세선화 과정 없이 수행한다. 골격선 추출을 위한 세선화 과정은 한글 문자 구조에 왜곡 현상을 발생시키고, 과거의 많은 연구가 이를 제거하기 위해 시간 소모적인 복잡한 알고리즘을 사용하였다. 본 논문에서는 세선화 과정 없이 문자의 자획을 수평 런 길이를 이용해 수평 자획, 경사 자획을 포함한 수직 자획, 폐곡선 자획으로 분리하였다. 분리된 자획은 온라인 필기체 한글 인식을 위해 개발되어 사용되고 있는 자획 사전 등과 같은 기술 등을 이용해 문자를 인식하는 데 이용될 수 있다 [9, 10, 11, 12, 13]. 부가적으로 본 논문에서 제안한 자획 분리 기술은 사용자 확인을 위한 서명 자동 인식(자획이

주요한 특징이 된다)에도 적용할 수 있으며 한자나 영문, 숫자 등의 인식에도 활용될 수 있다.

참고문헌

- [1] 최환수, 정동철, 공성필, “잡영과 왜곡이 심한 한글 문자의 자소분리 및 인식에 관한 연구”, 한국통신학회 논문지, Vol. 22, No. 6, pp. 1160-1169, 1997.
- [2] 백승복, 강순대, 손영선, “경계선 기울기 방법을 이용한 다양한 인쇄체 한글의 인식”, 한국퍼지 및 지능시스템학회논문지, 제13권, 1호, pp. 1-5, 2003.
- [3] 최재균, 강신욱, “자소의 논리적 분리를 이용한 고딕체 한글 문자 인식에 관한 연구”, 제1회 신호처리 합동 workshop 논문집, 제1권, 제1호, pp. 111-113, 1988.
- [4] Randy Crane, "A Simplified Approach to Image Processing", Prentice-Hall, 1997.
- [5] C. Sung Bae, K. Jin H, "Recognition of Large-set Printed Hangul(Korean Script) by two-stage Back propagation Neural Classifier", Pattern Recognition, Vol. 25, No. 11, pp. 1253-1360, 1992.
- [6] 권재욱, 조성배, 김진형, “계층적 신경망을 이용한 다중 크기의 다중 활자체 한글 문서 인식”, 한국 정보과학회 논문지, Vol. 19, pp. 69-79, 1992.
- [7] 김창운, “신경망 모델을 이용한 인쇄체 한글 문자의 인식에 관한 연구”, 광운대학교 대학원 석사학위 논문, 1992.
- [8] 김석중, 김대훈, 김기두, 김우성, 김영욱, “퍼지이론을 이용한 오프라인 필기체 한글 문자의 유형분석 및 자소 분리”, 대한전자공학회 하계종합학술대회, 제16권, 제1호, pp. 599-602, 1993.
- [9] 홍성민, 국일호, 김현우, 박수성, 이정환, 조원경, “구조해석적 방법과 획 사전을 이용한 문자 인식 시스템”, 한국통신학회 문자인식 워크샵, 제1권, pp.

195-198, 1993.

- [10] 정진국, “획 상대 위치 판별을 통한 온라인 필기체 한글 문자 인식에 관한 연구”, 한국데이터베이스학회 정보기술과 데이터베이스저널, 4권, 2호, pp. 65-78, 1998.
- [11] 김현우, “온라인 문자인식 시스템에서 획 인식을 위한 획 사전 구성에 관한 연구”, 경희대 대학원 석사학위논문, 1993.
- [12] 장승익, 임길택, 김호연, 정선화, 암윤석, “낱자 인식기와 자소 조합 인식기를 혼용한 인쇄체 한글 인식 방법”, 한국정보과학회 봄학술발표논문집, Vol. 30, No.1, pp. 244-246, 2003.
- [13] 김춘영, 석수영, 정호열, 정현열, “Substroke HMM 기반 온라인 필기체 문자인식”, 한국신호처리시스템학회 하계학술발표논문집, 제4권, 1호, pp. 74-77, 2003.

정민철(Minchul Jung)

[정회원]



- 1993년 2월 : 인하대학교 전자재료공학과 (공학사)
- 1995년 9월 : 미국뉴욕주립대 (버팔로) 전기, 컴퓨터공학과 (공학석사)
- 2001년 6월 : 미국뉴욕주립대 (버팔로) 전기, 컴퓨터공학과 (공학박사)

- 1993 ~ 1999년 : (주)포스데이타 OA사업부
- 1995 ~ 1999년 : 미국 CEDAR 연구소 연구원
- 1999 ~ 2001년 : 미국 Cedar Tech Inc. 연구원
- 2002년 ~ 현재 : 상명대학교 컴퓨터시스템공학과 조교수

<관심분야>

영상처리, 컴퓨터 비전, 인공신경망, 인공지능.