한국산학기술학회논문지
Vol. 7, No. 5, pp. 887-894, 2006

런 길이를 이용한 필기체 한글 자획의 교점 검출

정민철[1*]

Detection of Intersection Points of Handwritten Hangul Strokes using Run-length

Min-Chul Jung[1*]

요 약 본 논문은 런 길이를 이용해 필기체 한글 문자에서 자획의 교점을 검출하는 새로운 방법을 제안한다. 이를 위해 첫째로, 수평 런 길이와 수직 런 길이를 이용해 필기체 한글 문자의 자획 두께를 구하고, 둘째로, 자획 두께를 이용해 입력 문자의 자소를 수평 성분과 수직 성분으로 분리하며, 마지막으로, 자획의 수평 성분과 수직 성분을 이용해 자획의 교점을 구하는 기술을 제안한다. 수평 성분과 수직 성분 분석은 각도와 관계없이 자획 두께와 런 길이의 변화량만을 이용해 구한다. 자획의 교점은 오프라인 필기체 한글 인식을 위한 요소 기술 중 하나인 자소 분리를 위한 분리점 후보가 되며 분리된 자획은 필기체 한글 인식을 위한 특징을 나타낸다.

Abstract This paper proposes a new method that detects the intersection points of handwritten Hangul strokes using run-length. The method firstly finds the strokes' width of handwritten Hangul characters using both horizontal and vertical run-lengths, secondly extracts horizontal and vertical strokes of a character utilizing the strokes' width, and finally detects the intersection points of the strokes exploiting horizontal and vertical strokes. The analysis of both the horizontal and the vertical strokes doesn't use the strokes' angles but both the strokes' width and the changes of the run-lengths. The intersection points of the strokes become the candidated parts for phoneme segmentation, which is one of main techniques for off-line handwritten Hangul recognition. The segmented strokes represent the feature for handwritten Hangul recognition.

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

1. 서 론

한글은 하나의 문자가 1차원 배열로 나열되는 영어나 숫자와 달리 자음과 모음이 2차원 공간상에 복잡한 배열을 생성한다. 유니코드의 경우 한글의 한 문자는 초성 19개, 중성 21개, 종성 28개를 조합하여 만들어진다. 따라서 표현 가능한 글자의 수는 모두 19X21X28 = 11,172개이다. 이러한 많은 문자 모양의 패턴을 컴퓨터에 기억시켜 놓고 문자 패턴을 비교하여 문자를 인식할 경우, 그 패턴 데이터의 양은 엄청나게 많고 처리 시간도 많이 걸린다.

또한 필기체의 경우 하나의 문자도 필자에 따라 다양한 형태를 나타낸다. 우리가 한글을 쓸 때 자음, 모음의 조합을 통한 초성, 중성, 종성의 순서로 필기하듯이, 컴퓨터가 문자를 인식하는 과정도 조합된 문자를 개개의 패턴으로 인식하는 것이 아니라 초성, 중성, 종성으로 분리하여 수행함으로써 판별해야할 패턴의 종류를 줄이고 그 인식속도를 높일 수 있다. 따라서 하나의 문자를 인식하기 위해 자소를 분리하여야 하는데, 한글의 경우 많은 경우에 있어 초성과 중성, 중성과 종성이 서로 접촉, 접합되어 새로운 패턴을 형성한다. 그 결과, 중성 모음의 일부가 초성 자음과 접합되어 다른 초성으로 인식되거나 중성 모음의 일부가 종성 자음과 접합되어 다른 종성으로 오인식되는 경우가 발생한다. 인쇄체 한글 인식의 경우 세선화 기법을 이용하거나[1], 한글 문자를 6가지 형태로 구분하여 자소 분리를 시도해 왔다[2]. 그러나 세선화의 경우 한 문자가 입력되어 인식되기까지 소요되는 시간의 80%이상을 소모하며[3], 또한 문자 영상을 크게 변형시키고 잡영이 자획으로 오인식된다[1]. 또한 6가지 한글 문자 형태 분석에 의한 고정된 위치에서 자소분리는 필

[1]상명대학교 컴퓨터시스템공학과
*교신저자: 정민철(mjung@smu.ac.kr)

기체 한글의 경우 분리 최적 위치가 변동됨으로 인접 자소의 단편이 포함될 수 있고 자소 일부분이 잘려 나갈 수도 있다[1,4,5,6,7].

이에 본 논문에서는 오프라인 필기체 한글 문자의 자소 분리를 위한 요소 기술 중 하나인 자획의 분리와 자획의 교점을 검출하는 새로운 방법을 제안한다. 자획이란 자소를 구성하는 기본 단위인 선분을 말한다. 한글 자소는 방향성이 강한 자획들로 구성되어 있다. 본 논문에서 제안하는 자획 분리 기술과 자획의 교점을 검출하는 기술은 수평 런 길이와 수직 런 길이를 활용한다. 수평 자획은 '수평 런의 개수' 또는 '수직 런의 길이'가 그 자획의 두께가 되며, 수직 자획은 '수직 런의 개수' 또는 '수평 런의 길이'가 그 자획의 두께가 된다. 경사 자획은 본 논문에서는 그 경사 정도에 따라 수평 자획 또는 수직 자획에 포함되어 분리된다. 수평 자획과 수직 자획의 교점이 자획의 교점이 된다. 이러한 교점은 한 개의 자소 내에도 존재하지만 두 개 이상의 자소가 서로 접합할 때는 반드시 교점을 통해 접합됨으로 본 논문에서 검출한 교점의 위치는 자소 분리를 위한 일차 후보가 된다.

2. 연결 성분 분석

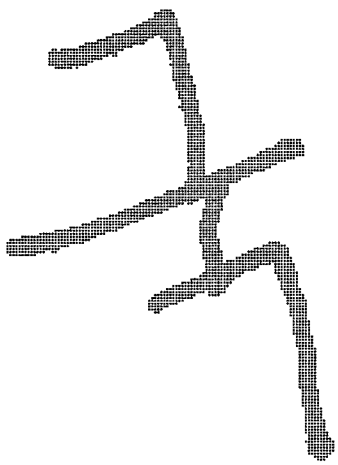


그림 1. 초성, 중성과 종성이 모두 접합되어 하나의 연결 성분으로 구성된 문자 '국'

입력 문자 영상에서 수평 자획과 수직 자획을 분리하기 위해 먼저 입력 문자 영상의 연결 성분 분석(connected component analysis)을 수행한다. 연결 성분 분석은 입력 영상을 위에서 아래로, 좌에서 우로 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 픽셀들의 런 길이(run length), 즉 런의 시작점과 끝점의 좌표에 의해 나타낼 수 있다. 흑색의 픽셀이 수평으로 연속되는 길이를 수평 런 길이, 수직으로 연속되는 길이를 수직 런 길이라 한다. 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 하나의 연결 성분에 '초성과 중성', '중성과 종성', 또는 '초성, 중성과 종성' 이 종종 접합되어 있으며 이는 자소 분리를 과정을 불가피하게 만든다. 그림 1은 초성, 중성과 종성이 모두 접합되어 하나의 연결 성분으로 구성된 문자 '국'을 보인다.

3. 수평 런 길이를 이용한 방향 성분

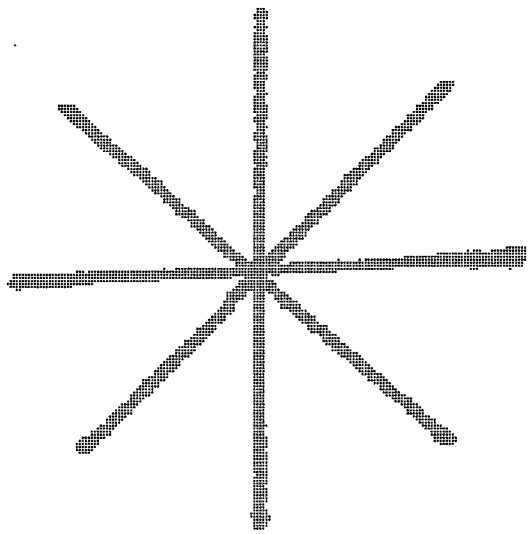


그림 2. 수평 런 길이에 의한 방향 성분 분류

그림 2에서 보듯이 각도 90도와 270도의 수직 성분은 수평 런 길이가 최소 3픽셀에서 최대 5픽셀의 값을 가지며 그 중 4픽셀의 값을 가장 많이 가진다. 또한 각도 45도, 135도, 225도와 315도의 경사 성분의 수평 런 길이는 대부분이 5에서 6픽셀의 값을 가진다. 따라서 그림 2에 있는 성분들의 두께는 오차범위 1픽셀 내외의 5픽셀이라 할 수 있다. 각도 0도와 180도의 수평 성분의 수평 런 길이는 앞에서 구한 두께 보다 훨씬 큰 값을 가진다. 따라서 입력으로 주어진 연결 성분을 왼쪽에서 오른쪽으로

수평 주사하면서 수평 런 길이가 급격이 변하는 부분을 자획의 방향 성분이 변하는 곳으로 일차적으로 정하고 자획을 분리한다. 자획을 분리한 후 분리된 자획의 크기가 임계치보다 작으면 분리를 취소한다. 즉, 분리된 후보 자획의 가로와 세로 길이가 위에서 구한 두께 보다 작으면 이를 독립된 자획으로 볼 수 없으므로 분리를 취소한다. 자획의 수직 성분과 수직 성분에 가까운 경사 성분은 수평 런 길이가 모두 자획의 두께와 같으므로 자획 분리에서는 하나의 성분으로 간주된다. 수평 성분에 아주 가까운 경사 성분은 수평 성분으로 간주되는 데, 이는 이러한 경사 성분의 수평 런 길이가 자획의 두께 보다 훨씬 크기 때문이다. 즉, 수평 런 길이를 이용한 방향 성분 분석은 각도와 관계없이 자획의 두께와 수평 런 길이의 변화량만을 이용해 구한다. 입력으로 주어진 연결 성분을 위쪽에서 아래쪽으로 수직 주사하면서 위에서 언급한 과정을 반복하면 수직 런 길이를 이용해 방향 성분 분석을 할 수 있다.

4. 수평 런 길이와 수직 런 길이 빈도수 히스토그램

문자의 자획 두께를 이용해 문자의 연결 성분을 방향 성분으로 분리하기위해 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 (1)에 따라 구할 수 있다.

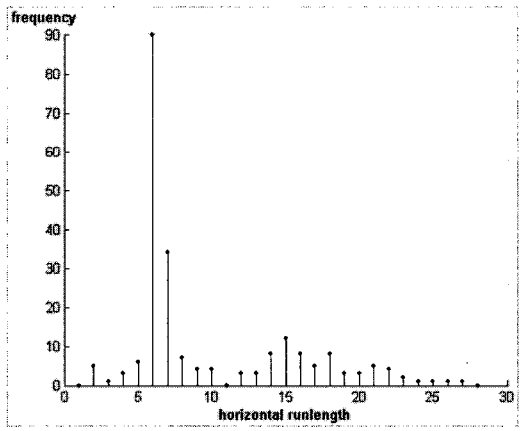
$$histogram[runlengh] \leftarrow histogram[runlengh] + 1 \cdots\cdots\cdots (1)$$

그 다음으로, 수평 런 길이 히스토그램의 빈도수에서 최대값을 가지는 수평 런 길이를 입력 문자 자획의 두께 w로 한다. 식 (2)는 이를 나타낸다.

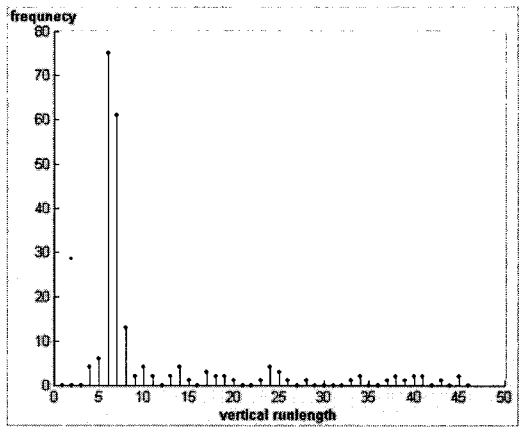
$$w = \max_{runlength=3}^{N} (hsitogram[runlength]) \quad (2)$$

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

수직 런 길이에 대해서도 위의 과정을 되풀이 하면 수직 런 길이 히스토그램을 구할 수 있다. 그림 1에서 보인 문자 '국'은 그림 3에서 같이 수평 런 길이 빈도수 히스토그램과 수직 런 길이 빈도수 히스토그램으로 나타낼 수 있다. 그림 3(a)에서 보듯이 모두 220개의 수평 런이 있으며(수평 런 길이 2이하는 제외), 그 중 수평 런 길이가 6픽셀인 것은 90개, 7픽셀인 것은 34개이다. 즉, 수평 런 길이 히스토그램에서 최대 빈도수를 보이는 수평 런 길이는 6픽셀이다. 또한 그림 3(b)에서 보듯이 모두 206개의 수직 런이 있으며(수직 런 길이 2픽셀이하는 제외), 그 중 수직 런 길이가 6픽셀인 것은 75개, 7픽셀인 것은 61개이다. 즉, 수직 런 길이 히스토그램에서 최대 빈도수를 보이는 수직 런 길이는 6픽셀이다. 따라서 주어진 문자의 자획 두께 w는 6픽셀이고 오차 범위 1픽셀을 고려하면 5픽셀에서 7픽셀 사이이다.



(a)



(b)

그림 3. '국' 문자의 (a) 수평 런 길이 빈도수 히스토그램, (b) 수직 런 길이 빈도수 히스토그램

5. 문자의 수평 성분과 수직 성분 분리

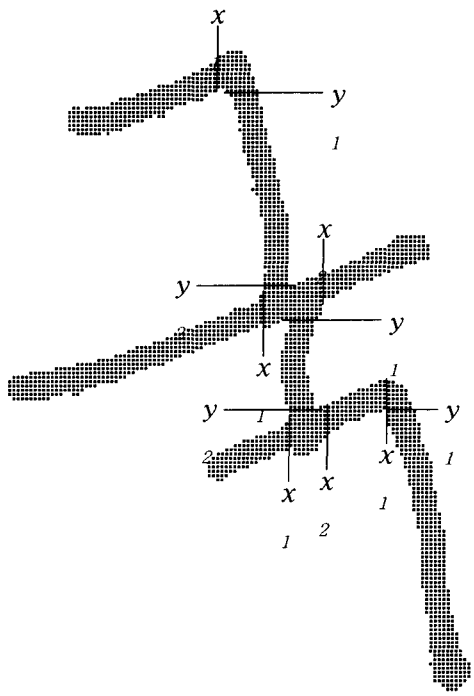


그림 4. 문자 '국'의 수직 자획의 시작 교점(x_1, y_1)과 수평 자획의 시작 교점(x_2, y_2)

입력 문자의 자획 두께를 구한 후, 입력 문자를 다시 수평으로 스캔하면서 수평 런 길이가 문자의 자획두께와 같아지는 부분을 y_1 으로, 문자의 자획 보다 2배 이상 커지는 부분을 y_2 으로 나타낸다. 즉, y_1 은 수직 자획의 시작 교점이고, y_2 는 수평 자획의 시작 교점이다. 마찬가지로 입력 문자를 수직으로 스캔하면서 수직 런 길이가 문자의 자획두께와 같아지는 부분을 x_2 로, 문자의 자획 보다 2배 이상 커지는 부분을 x_1 으로 나타낸다. 즉, x_1 은 수직 자획의 시작 교점이고, x_2 는 수평 자획의 시작 교점이다. 그림 4는 문자 '국'의 수직 자획의 시작 교점(x_1, y_1)과 수평 자획의 시작 교점(x_2, y_2)을 나타낸다. 이러한 교점들을 경계로 문자를 수평 성분과 수직 성분으로 분리한다. 단, 앞에서도 언급했듯이, 분리된 자획의 가로와 세로 길이가 자획의 두께 보다 작으면 이를 독립된 자획으로 볼 수 없으므로 분리를 취소한다.

그림 5는 문자 '국'이 수평 성분과 수직 성분의 자획으로 분리 된 것을 보인다. 문자의 방향 패턴을 추출하는 다른 알고리즘[1]과는 다르게, 본 논문에서 제안한 방법은 문자 자획이 꺾이는 부분과 문자 자획이 서로 만나는 부분이 수평 성분과 수직성분 모두에 포함된다. 따라서 수평 성분 영상(horizontal image)과 수직 성분 영상(vertical image)을 AND 연산을 하면 문자 자획이 꺾이는 부분과 문자 자획이 서로 만나는 교점 부분(intersection image)을 추출할 수 있다.

$$intersection\ image = horizontal\ image \cap vertical\ image \tag{3}$$

그림 6(a)는 그림 5(b)와 그림 5(c)를 AND 연산한 결과 영상이고, 그림 6(b)는 그림 5(a)에서 그림 6(a)를 뺀 영상이며, 그림 6(c)는 자획의 교점 부분을 독자의 이해를 돕기 위해 회색으로 표현한 영상이다.

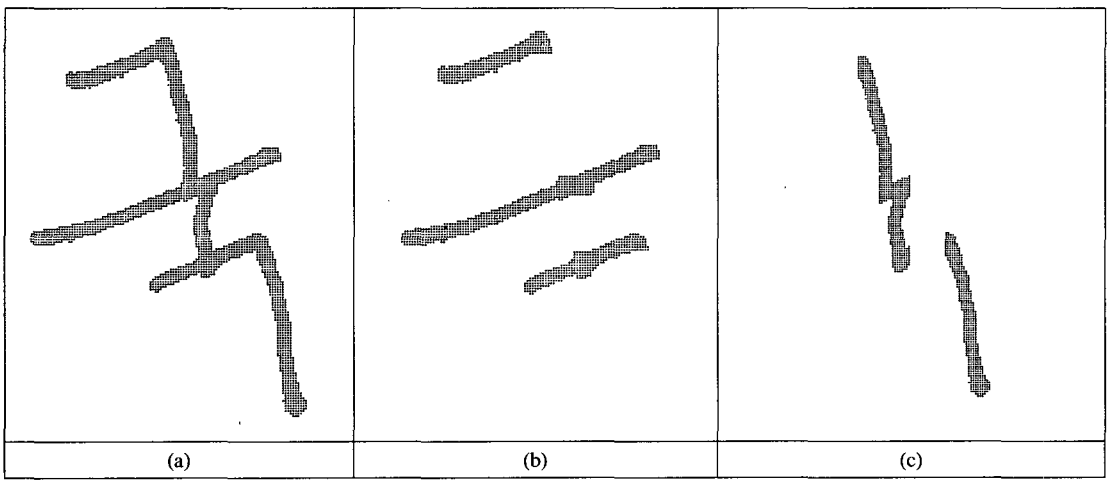


그림 5. 문자 '국'(a)의 수평 성분(b)와 수직성분(c)

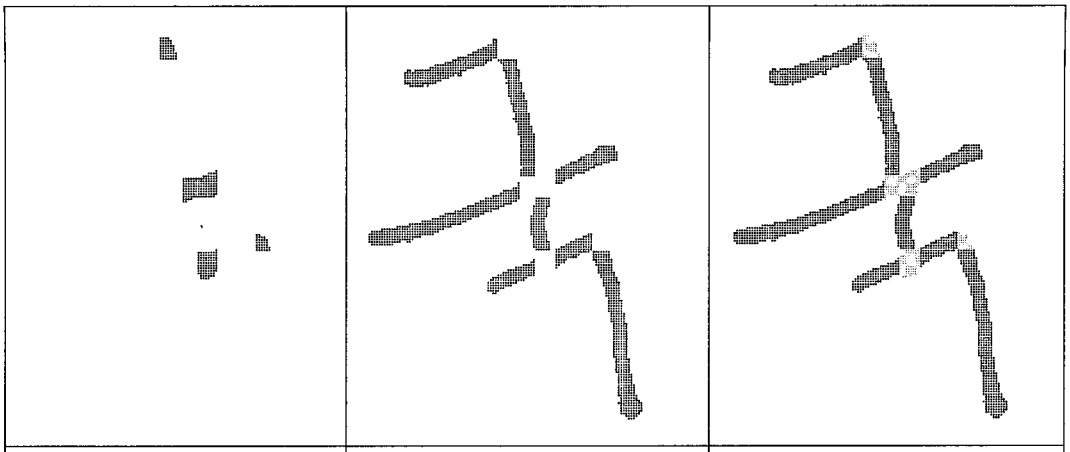


그림 6. 문자 '국'의 수평 성분과 수직성분의 교점

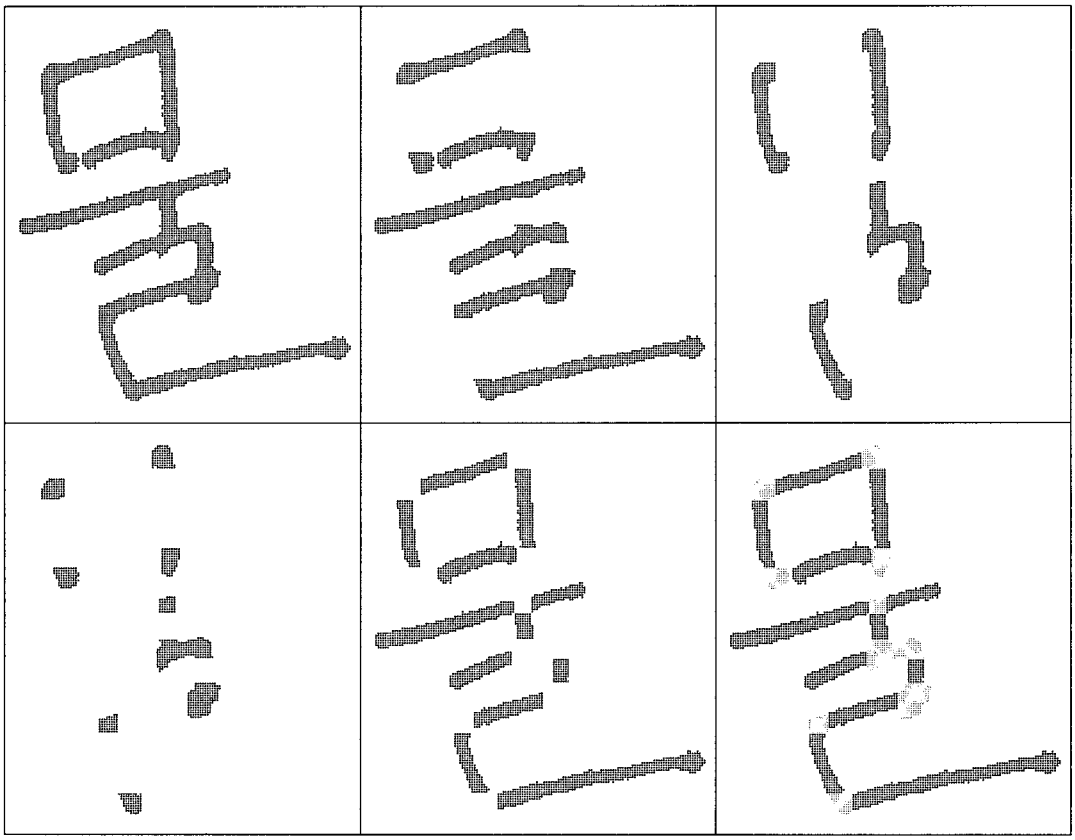


그림 7. 문자 '물'의 수평 성분, 수직 성분과 교점

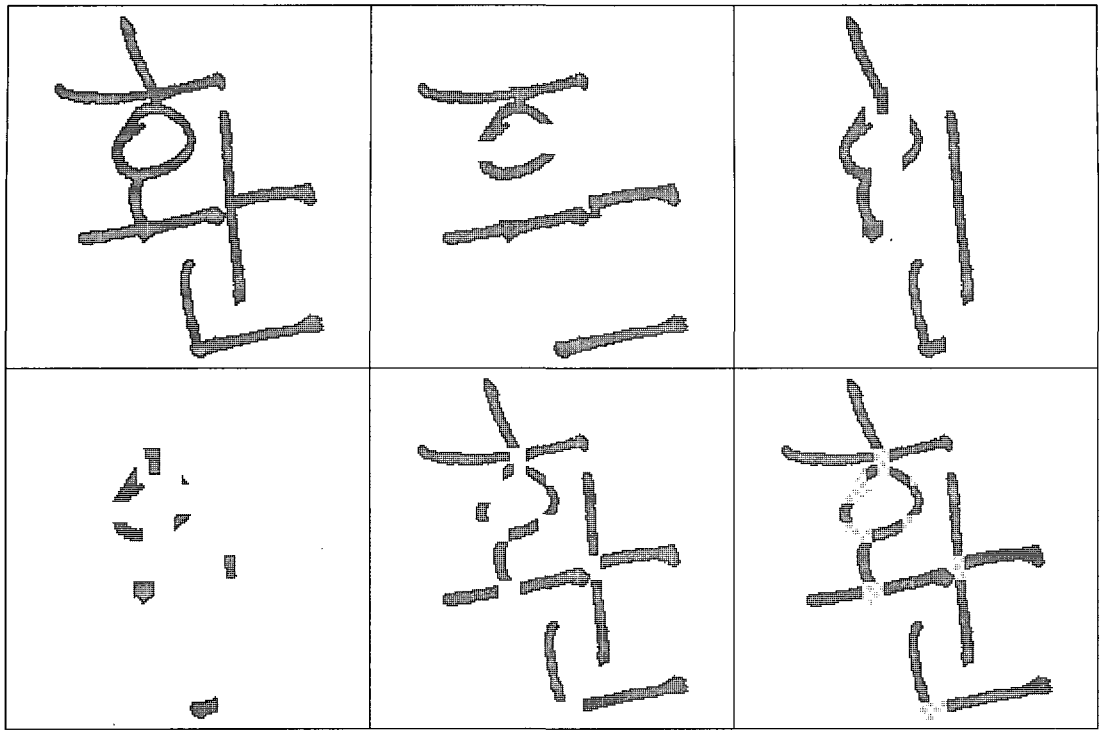


그림 8. 문자 '환'의 수평 성분, 수직 성분과 교점

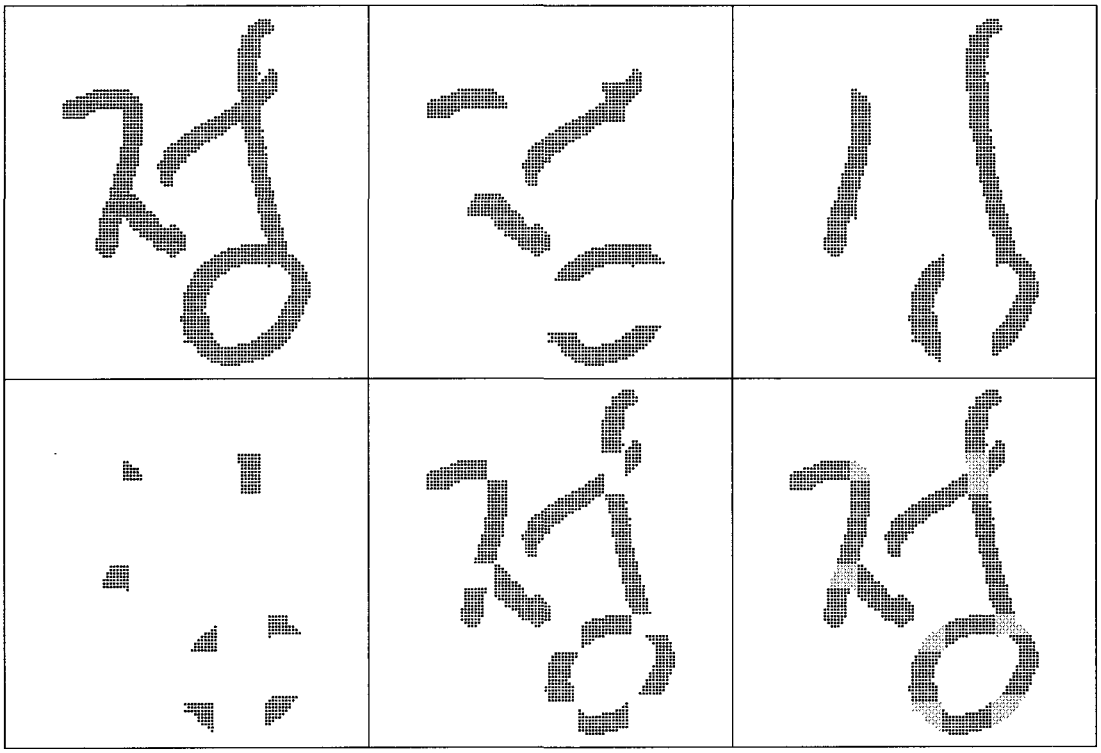


그림 9. 문자 '정'의 수평 성분, 수직 성분과 교점

6. 실험 결과

그림 7, 8, 9는 실험에 사용된 문자들 중 수평 성분, 수직 성분으로 분리된 영상과 그 교점 영상의 예를 보인다. 그림 7에서 문자 '물' 의 초성 'ㅁ'은 수평 성분과 수직 성분으로 성공리에 분리된 것을 보인다. 폐곡선 자음 'ㅁ'은 네 개의 교점을 가진다. 그러나 중성 'ㅜ' 의 수직 성분 'ㅣ'과 종성 'ㄹ'의 상단 수직 성분 'ㅣ' 부분이 하나의 수직 성분으로 나타난다. 이는 분리되었던 수평 성분이 임계치보다 적어 분리가 취소되고 두 수직 성분에 연결되어 나타난 결과이다. 여기서 임계치는 분리된 성분의 수평 런의 개수나 수직 런의 개수이다. 수평 런의 개수나 수직 런의 개수가 자획을 구성하기에 적으면 분리는 취소되고 보다 큰 성분의 일부분이 된다. 따라서 이 부분의 교점이 실제로는 두 개이나 하나의 긴 교점으로 나타났다. 이와 같은 현상은 그림 8에서도 나타났는데 문자 '환'의 'ㅇ' 의 왼쪽 수직 성분과 모음 'ㅗ' 의 수직 성분 'ㅣ' 이 하나의 수직 성분으로 표현된다. 이러한 현상은 필기체 한글의 특성 상 불가피한 것으로 이를 분리하려고 임계치를 적게 하면 이들 성분은 분리되나, 하나의 성분이 두개 이상으로 절편화되는 부작용이 발생하였다. 그림 9를 보면 문자 '정'의 초성 'ㅈ'에서 '/' 부분은 수직에 가까워 수직 성분으로 분리 되었고 '\' 부분은 수평 성분에 가까워 수평 성분으로 분리되었음을 알 수 있다. 이 결과는 만약 또 다른 필자가 만일 초성 'ㅈ'을 '/' 부분은 수평에 가깝게 '\' 부분은 수직에 가깝게 쓴다면 위의 결과와는 반대로 분리 될 수 있다. 즉, 수직 성분과 수평 성분을 가르는 판단은 수평 런 길이와 수직 런 길이의 최대 빈도치인 자획의 두께이다. 비록 경사 성분이 경우에 따라 수평 성분으로, 수직 성분으로 분리 될 지라도 그 성분들의 교점은 불변이다. 또한 그림 9는 폐곡선 자음 'ㅇ'은 네 개의 교점을 가지는 수직 성분과 수평 성분으로 성공리에 분리됨을 보인다.

7. 결 론

본 논문에서는 수평 런 길이와 수직 런 길이를 이용해 입력 문자의 자획 두께를 구하고, 이를 이용해 입력 문자의 자소를 수평 성분과 수직 성분으로 분리하며, 또한 자획의 교점을 구하는 기술을 제안하였다. 즉, 방향이 다른 두 개의 자획이 겹치는 영역을 한 자획에만 분리하여 포함시키는 것이 아니라 두 개의 자획 모두에 포함시키고 이를 이용해 자획의 교점을 구했다. 이러한 자획의 교점은 오프라인 필기체 한글 인식을 위한 요소 기술 중 하나인 자소 분리를 위한 분리점 후보가 된다. 즉, 오프라인 필기체 한글은 '초성과 중성', '중성과 종성', 또는 '초성, 중성과 종성' 이 종종 접합되는 데 효율적 인식을 위해 이를 분리해야 한다. 그러나 정확한 접합부를 찾는 것은 쉽지 않은 문제 일 뿐 아니라 만약 잘못된 분리를 통해 중성의 일부가 초성으로, 또는 종성으로 분리되거나 초성이나 종성의 일부가 중성의 일부로 분리된다면 아무리 인식 모듈 성능이 좋다하더라도 인식 오류는 필할 수 없다. 본 논문에서 제안한 방법으로 구한 자획의 교점들은 자소 분리를 위한 교점, 즉 경계의 일차 후보가 된다. 또한 한글의 일반적인 6가지 형태[2]를 이용해, 즉 문자 내 교점들의 상대적 위치에 따라 자소 내의 교점인지 자소간의 교점인지 알 수 있다. 또한 자획의 교점은 자소의 구조적 특징을 잘 나타냄으로 특징 벡터의 좌표로 사용될 수도 있다[8]. 즉, 분리된 자획의 길이(r)과 그 각도(Θ), 자획의 교점(x, y)는 오프라인 필기체 한글 인식을 위한 특징 벡터의 좋은 구성 요소가 된다.

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, 강신옥, "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] 권재옥, 조성배, 김진형, "계층적 신경망을 이용한 다중 크기의 다중 활자체 한글 문서 인식", 한국 정보과학회 논문지, Vol. 19, pp. 69-79, 1992.

[7] 김창윤, "신경망 모델을 이용한 인쇄체 한글 문자의 인식에 관한 연구", 광운대학교 대학원 석사학위 논문, 1992.

[8] Simon Kahan, Theo Pavlids, Henry S. Baird, "On the Recognition of Printed Characters of Any Font and Size", IEEE Transactions on Pattern Analysis and machine Intelligence, Vol. PAMI-9, No. 2, pp. 274-288, 1987.

정 민 철(Min-Chul Jung) **[정회원]**

- 1993년 2월 : 인하대학교 전자재료공학과(공학사)
- 1995년 9월 : 미국 뉴욕주립대 전기,컴퓨터공학과(공학석사)
- 2001년 6월 : 미국 뉴욕주립대 전기,컴퓨터공학과(공학박사)
- 2002년 9월 ~ 현재 : 상명대학교 컴퓨터시스템공학과 조교수

<관심분야>
영상처리, 컴퓨터 비전, 인공신경망, 인공지능.