런 길이를 이용한 필기체 한글 자획의 방향 성분 추출

정민철
상명대학교 공과대학 컴퓨터시스템공학과
e-mail주소: mjung@smu.ac.kr

Extraction of Directional Strokes in Handwritten Hangul using Runlength

Minchul Jung
Dept. of Computer System Engineering, Sangmyung University

요 약

본 논문은 수평 런 길이와 수직 런 길이를 이용해 필기체 한글 문자의 자획 두께를 구하고, 그 자획 두께를 이용해 입력 문자의 자소를 수평 성분과 수직 성분으로 분리하는 기술을 제안한다. 수평 성분과 수직 성분 분석은 각도와 관계없이 자획 두께와 수평 런 길이의 변화량만을 이용해 구한다. 분리된 수평 성분 자획과 수직 성분 자획은 오프라인 필기체 한글 인식을 위한 요소 기술 중 하나인 자소 분리를 위한 특징이 된다.

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 문서를 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 밝기인 정적 정보만을 이용할 수 있는 오프라인 필기체 문자인식은 인식률을 높이기 위해서 아직 풀어야할 많은 문제점이 존재한다. PDA 등에 사용되는 온라인 필기체 한글 인식 제품은 이미 상품화된 것도 있으나, 오프라인 필기체 한글 인식은 아직 초보적인 연구 단계에 머물러 있으며, 인식률은 저조하며 오인식률은 높아 실용화하여 상품화하기에는 아직 미흡한 점이 많다. 이에 본 논문에서는 오프라인 필기체 한글 문자 인식률을 높이기 위한 새로운 방법을 개발하여 구현한다. 한글 문자는 초성 자음, 중성 모

음, 종성 자음이 조합하여 하나의 문자를 이루는 계층적 구조를 가진다. 따라서 문자를 인식하기 위해 자소 분리를 하고, 자소 분리를 위해 자획 분리를 하여, 자소 간의 접촉 문제와 문자 간의 접촉 문제를 동시에 해결한다. 본 논문에서 제안하는 자획의 방향 성분 추출은 수평 런 길이를 활용한다. 수직 자획이나 경사 자획의 수평 런 길이는 자획 두께가 되며, 수평 자획의 수평 런의 개수가 자획 두께이다. 이를 활용해 자소의 자획을 수직과 수평으로 구분 분리한다. 경사 성분은 수직이나 수평에 포함되어 분리된다.

2. 연결 성분 분석

연결 성분 분석(connected component analysis)은 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 수평 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.

3. 수평 런 길이 와 수직 런 길이 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.

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

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

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

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 2에서 보인 문자 '국'은 그림 1에서 같이 수평 런 길이 빈도수 히스토그램과 수직 런 길이 빈도수 히스토그램으로 나타낼 수 있다.

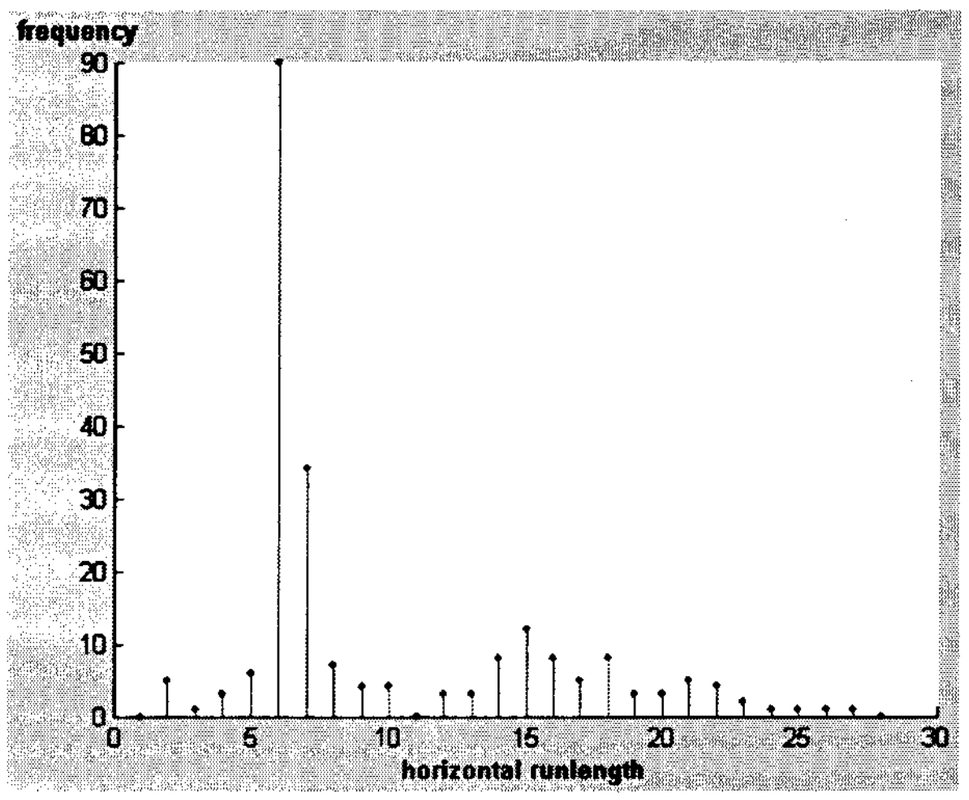

그림 1. 문자 '국'의 (a) 수평 런 길이 빈도수 히스토그램, (b) 수직 런 길이 빈도수 히스토그램

그림 1(a)에서 보듯이 문자 '국'에는 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 1(b)에서 보듯이 모두 206개의 수직 런이 있으며(수직 런 길이 2픽셀이하는 제외), 그 중 수평 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.

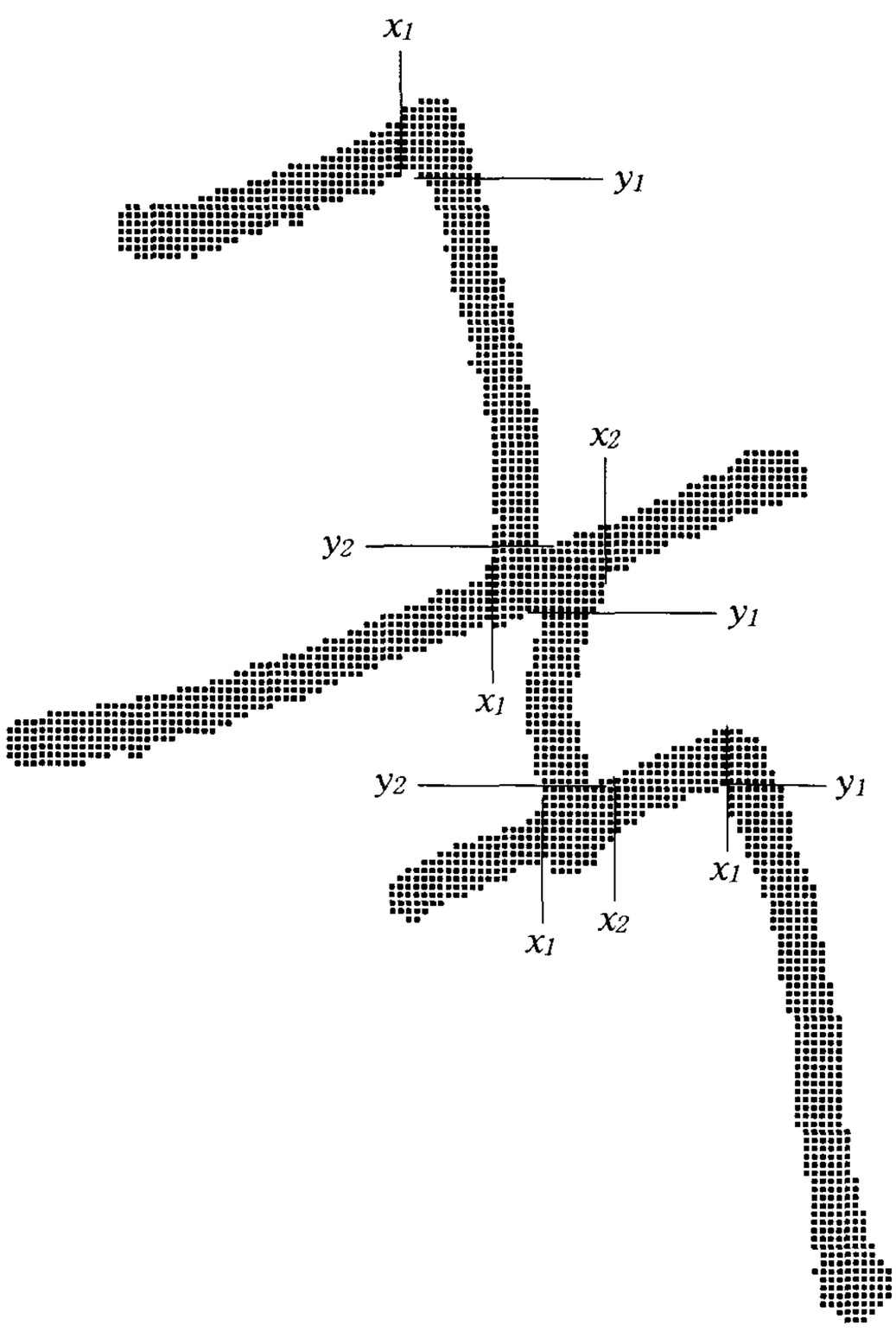

그림 2. 문자 '국'의 수직 자획의 시작 교점(x_1, y_1)과 수평 자획의 시작 교점(x_2, y_2)

4. 문자의 수평 성분과 수직 성분 분리

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 2는 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 분리 한다. 단, 분리된 자획의 가로와 세로 길이가 자획의 두께 보다 작으면 이를 독립된 자획으로 볼 수 없으므로 분리를 취소한다. 그림 3은 문자 '국'이 수평 성분과 수직성분으로 분리 된 것을 보인다.

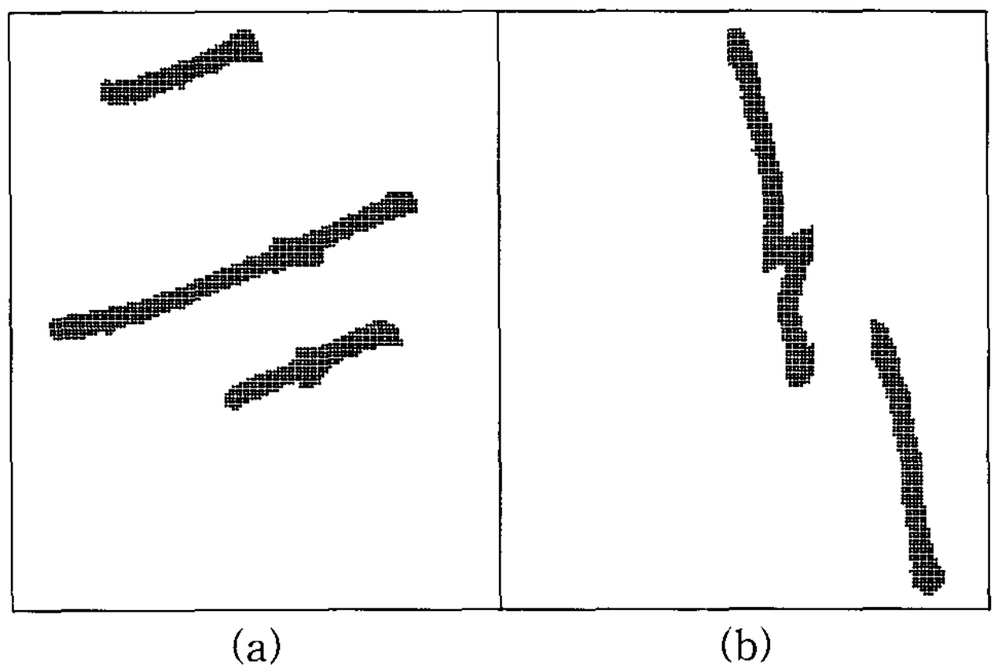

그림 3. 문자 '국'의 수평 성분(a)와 수직성분(b)

참고문헌

[1] 박정선, 이승환, "필기체 한글의 오프라인 인식을 위한 효과적인 두 단계 패턴 정합 방법", 전자공학회논문지, 제 31권, B편, 제4호, pp. 351-358, 1994.

[2] 최환수, 정동철, 공성필, "잡영과 왜곡이 심한 한글 문자의 자소분리 및 인식에 관한 연구", 한국통신학회 논문지, Vol. 22, No. 6, pp. 1160-1169, 1997.

[3] 백승복, 강순대, 손영선, "경계선 기울기 방법을 이용한 다양한 인쇄체 한글의 인식", 한국퍼지 및 지능시스템학회논문지, 제13권, 1호, pp. 1-5, 2003.

[4] 김춘영, 석수영, 정호열, 정현열, "Substroke HMM 기반 온라인 필기체 문자인식", 한국신호처리시스템학회 하계학술발표논문집, 제4권, 1호,

pp. 74-77, 2003.

[5] 정진국, "획 상대 위치 판별을 통한 온라인 필기체 한글 문자 인식에 관한 연구", 한국데이타베이스 학회 정보기술과 데이타베이스저널, 4권, 2호, pp. 65-78, 1998.

[6] 장승익, 임길택, 김호연, 정선화, 암윤석, "낱자 인식기와 자소 조합 인식기를 혼용한 인쇄체 한글 인식방법", 한국정보과학회 봄학술발표논문집, Vol. 30, No.1, pp. 244-246, 2003.