한국지능정보시스템학회논문지 제10권 제3호 2004년 12월(pp.1~10)

측면윤곽 패턴을 이용한 접합 문자 분할 연구

정민철
상명대학교 공과대학 컴퓨터시스템공학과
(mjung@smu.ac.kr)

본 논문에서는 영문 인쇄체의 접합 문자를 분할하는 새로운 알고리듬을 제안한다. 본 논문에서 제안하는 문자 분할의 접근 방식은 특징을 기반으로 한 접근 방식(feature-based approaches)과 인식을 기반으로 한 접근 방식(recognition-based approaches)의 단점을 보안한 새로운 문자 분할 방법이다. 접합 문자의 측면 윤곽 특징을 정의하고, 그 측면 윤곽 특징을 이용하여 문자 인식의 도움 없이도 접합 문자 내의 문자를 일차 내정하여 분할 한 후 다시 측면 윤곽 특징을 이용하여 문자 분할을 최종 확정한다. 또한 본 논문에서는 분할 비용을 정의하는데, 분할 비용은 최적의 경로로 문자 분할을 수행하도록 한다. 제안된 문자 분할의 성능은 U.S. 메일에서 주소를 자동으로 인식하여 메일을 자동으로 도착지별로 분류하는 시스템(Envelope Reader System)을 이용해 구해졌다. 3359개의 메일이 실험되어졌는데, 제안된 문자 분할 알고리즘에 의해 분류율이 68.92%에서 80.08%로 성능이 향상되었다.

논문접수일 : 2004년 5월 게재확정일 : 2004년 12월 교신저자 : 정민철

1. 서 론

문서를 인쇄할 때나 또는 스캐닝할 때 각 문자는 이웃하는 문자와 접합(touching)되어 원래 개개의 문자 패턴과는 전혀 다른 새로운 패턴을 형성한다. 이러한 접합 문자(touching characters)를 원래 개개의 문자로 분리하는 것을 문자 분할(character segmentation)이라 하며, 문서 자동 처리와 인식 분야에서는 현재까지도 완결되지 않은 문제이다. OCR (Optical Character Recognition) 시스템이 문자를 자동으로 인식할 때 접합 문자는 문자 인식률을 크게 저하시키고 에러율은 높게 한다. 그 이유는 문자인식기(character recognizer)는 단일 문자를 입력으로 받아 특징 벡터를 추출하여 그 문자를 분류하는 데, 두 문자 이상이 접합된 문자열은 단일 문자의 특징 벡터 추출을 매우 어렵게 하기 때문이다. 부정확한 문자 분할은 분류 오류의 가장 큰 원인들 중 하나이다. 사실, 문자 인식을 연구하면서 관찰한 결과 인식 오류의 반 이상은 접합 문자들이 원인이 되었다. 문자가 정확히 분할되는 한, 각 문자 이미지의 질 저하는 OCR 시스템의 전체적인 인식률에 크게 영향을 미치지 않는다. 만약 인식 결과를 신뢰할 수 없다면 그 이유는 바로 접합 문자들을 제대로 분할할 수 없기 때문이다[1]. 문자 인식기(character recognizer)는 문자를 인식하기 위해 문자 분할(character segmentation)을 전 처리 단계에 필요로 하는데, 문자 분할(character segmentation)은

높은 성능을 위해 문자 인식(character recognition) 결과를 필요로 한다[2]. 이 딜레마를 해결하기 위해서는 문자 분할과 문자 인식, 이 두 문제를 동시에 해결하는 방법이 필요하다. 이를 위해 본 논문에서는 문자 분할 전에 접합 문자 내에 있는 소속 문자를 인식하고 문자를 분할하는 새로운 문자 분할 방법을 제시한다.

2. 문자 분할의 접근법

인쇄체 문자 분할에 관련된 기존의 접근 방식은 크게 두 가지로 나눌 수 있는데 첫째는 특징을 기반으로 한 접근 방식(feature-based approaches)이고, 둘째는 인식을 기반으로 한 접근 방식(recognition-based approaches)이다[3]. 특징을 기반으로 한 접근 방식은 오픈 루프(open-loop) 또는 분할 후 인식(segmentation-*to*-recognition) 방법으로 문자 분할이 문자 인식 전에 수행된다. 이 방식에 사용되는 방법은 수직 투영 윤곽 분석, 윤곽선 분석, 피치(pitch) 추정법 등이 있다. 수직 투영 윤곽 분석에서 사용되는 특징은 봉우리(peak)와 계곡(valley)의 갯수, 높이, 폭 등으로 수직 투영 윤곽에서 최소의 밀도를 가지는 부분이 분할되는 영역으로 선택된다[4]. 윤곽선 분석에서 사용되는 특징은 상부 윤곽선의 최소점과 하부 윤곽선의 최대점으로 이러한 점들이 분할되는 점으로 선택된다[5]. 피치 추정 법에서 사용되는 특징은 개개 문자의 폭으로써 단일 문자의 폭대로 접합 문자를 분할하는 방법이다[6, 7]. 이 방법은 모든 문자가 같은 폭을 가지는 한글이나 커리어(courier) 폰트 같은 고정 피치(fixed pitch)를 갖는 영문 폰트에는 효과적이나 가변 피치(variable pitch)를 가지는 대분부의 영문 폰트에는 사용될 수 없다. 인식을 기반으로 한 접근 방식은 클로즈 루프(closed loop) 또는 분할과 인식(segmentation-*and*-recognition) 방법으로 문자 분할과 문자 인식이 협력되어 수행된다. 슬라이딩 윈도우 방법은 문자 크기의 윈도우가 접합 문자를 따라 움직일 때 윈도우에 들어가는 영역의 이미지가 인식의 조건에 일치하면 문자 인식을 위한 처리를 한다[8, 9]. 회귀적 분할과 인식 방법은 먼저 예정 분할점을 정한 후 회귀적 분할과 인식 처리를 한 후, 언어학적 문맥을 이용하여 분할점을 결정한다[10]. 특징을 기반으로 한 접근 방식(feature-based approaches)은 아무런 문자 인식이나 구조의 지식 없이 문자 분할을 수행함으로 에러 발생이 필연적이며 일단 발생된 에러는 수정할 수 없다. 인식을 기반으로 한 접근 방식(recognition-based approaches)은 문자 인식의 결과를 문자 분할에 다시 이용하는데 많은 처리 시간이 필요하다. 본 논문에서는 위 두 방법의 단점을 보안한 새로운 문자 분할 방법을 제시한다. 접합 문자의 측면 윤곽 특징을 정의하고, 그 측면 윤곽 특징을 이용하여 문자 인식의 도움 없이도 접합 문자 내의 문자를 일차 내정하여 분할 한 후 측면 윤곽 특징을 이용하여 문자 분할을 최종 확정한다. 문자 분할 시에는 분할 비용을 고려하여 최적을 경로로 분할을 수행한다.

3. 측면 윤곽 패턴을 이용한 문자 분할

본 논문에서는 접합 문자의 새로운 특징으로 측면 윤곽을 정의하고 이용하였다. 두 개 또는 그 이상의 단일 문자가 접합되어 접합 문자가 되면 그 접합 문자는 단일 문자와는 전혀 다른 새로운 패턴이 된다. 그러나 접합 문자의 맨 왼쪽 측면 패

턴과 맨 오른쪽 측면 패턴은 단일 문자일 때와 동일하며 이는 문자 접합에 영향을 받지 않는다. 이러한 측면 윤곽 패턴을 분석하여 접합 문자 내에 있는 문자를 문자 분할 전에 인식할 수 있다. 영문자 및 숫자는 그림 1에서 알 수 있듯이 독특한 네 방향의 측면 윤곽 패턴을 가진다(한글도 독특한 측면 윤곽패턴을 가지나 본 연구에서는 한글은 제외한다). 대부분의 문자들은 이러한 네 방향의 독특한 측면 윤곽 패턴 중 한 방향의 측면 윤곽 패턴만을 가지고도 인식될 수 있다.

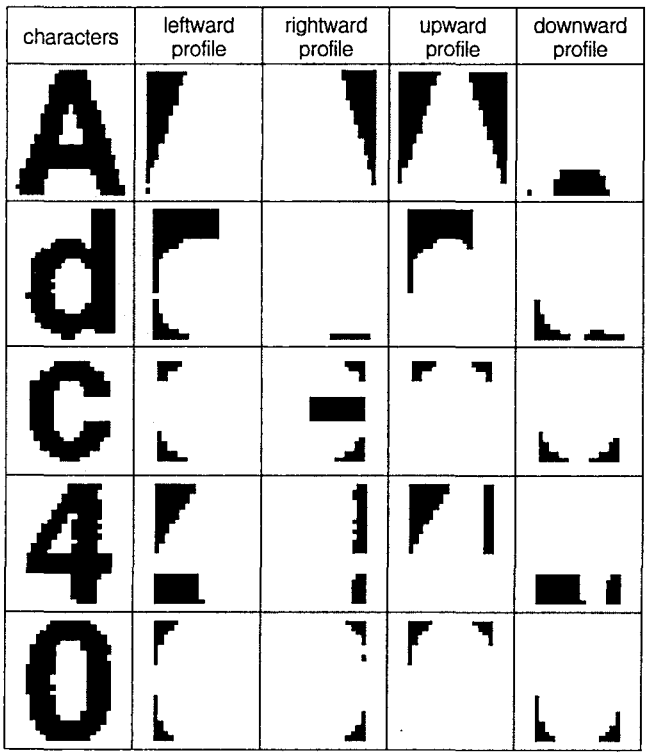
그림 1에서 보이는 네 방향의 측면 윤곽 패턴은 다음과 같은 공식으로 정의되어진다.

- 위쪽 방향의 측면윤곽(upward profile): 이미지 $I(x, y)$를 위쪽(y_0)에서 아래쪽으로 수직으로 스캔하면서 흑색 픽셀을 만날 때까지(y_1)의 백색 픽셀 갯수의 합

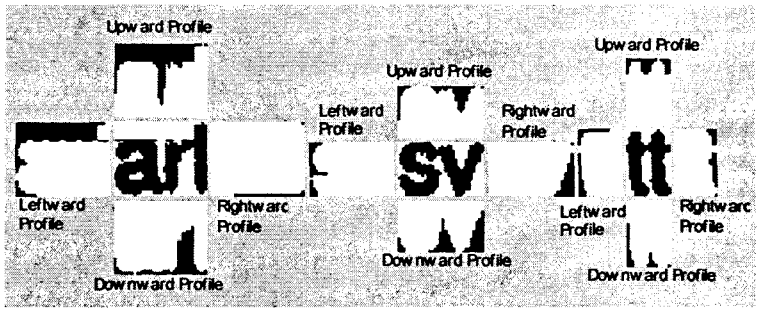
$$P_{up}(y) = \sum_{y=y0}^{y1} I(x, y) \quad (1)$$

- 아래쪽 방향의 측면윤곽(downward profile): 이미지 $I(x, y)$를 아래쪽(y_2)에서 위쪽으로 수직으로 스캔하면서 흑색 픽셀을 만날 때까지(y_3)의 백색 픽셀 갯수의 합

$$P_{down}(y) = \sum_{y=y2}^{y3} I(x, y) \quad (2)$$



[그림 1] 측면 윤곽 패턴의 예



[그림 2] 접합 문자의 측면 윤곽 패턴의 예

- 왼쪽 방향의 측면윤곽(leftward profile): 이미지 $I(x, y)$를 왼쪽(x_0)에서 오른쪽으로 수평으로 스캔하면서 흑색 픽셀을 만날 때까지(x_1)의 백색 픽셀 갯수의 합

$$P_{left}(x) = \sum_{x=x_0}^{x_1} I(x, y) \tag{3}$$

- 오른쪽 방향의 측면윤곽(rightward profile): 이미지 $I(x, y)$를 오른쪽(x_2)에서 왼쪽으로 수평으로 스캔하면서 흑색 픽셀을 만날 때까지(x_3)의 백색 픽셀 갯수의 합

$$P_{right}(x) = \sum_{x=x_2}^{x_3} I(x, y) \tag{4}$$

위 공식들에서 y_0, y_2, x_0, x_2 는 테두리 박스(bounded box) 상의 좌표를 y_1, y_3, x_1, x_3 는 수직 또는 수평으로 스캔할 때 최초로 문자를 형성하는 흑색 픽셀을 만나는 좌표를 말한다(그림 2 참조).

그림 2은 접합 문자들의 측면 윤곽 패턴을 나타낸다. 그림 2에서 보듯이 각 접합 문자들의 맨 왼쪽 측면 윤곽 패턴과 맨 오른쪽 측면 윤곽 패턴은 문자들의 접합에 영향을 받지 않고 단일 문자의 측면 윤곽 패턴들과 동일함을 알 수 있다.

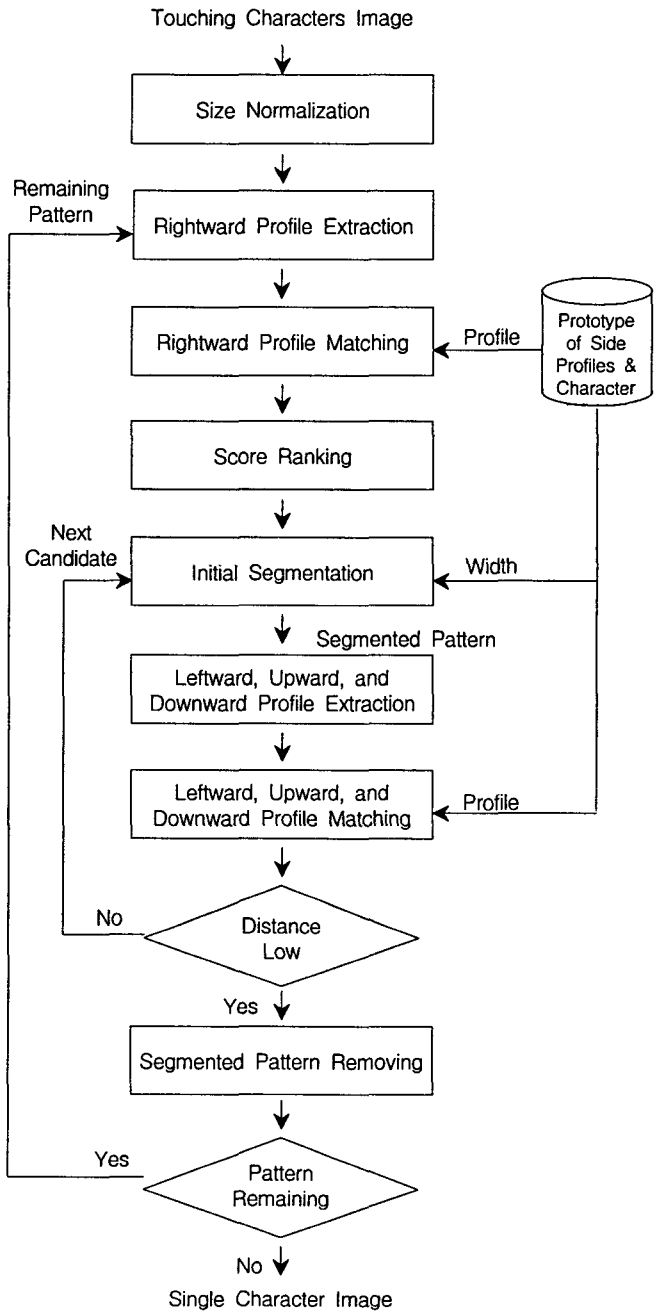
그림 3은 본 논문에서 제안한 측면 윤곽 패턴을 이용한 접합 문자 분할의 전반적인 구성도를 보인다. 첫 번째, 접합 문자를 가로세로 비율을 유지하며 지정한 세로 크기로 크기 정규화를 한다. 두 번째, 크기 정규화된 접합문자의 맨 오른쪽 측면 윤곽으로부터 히스토그램을 구한다. 히스토그램의 x 축은 픽셀의 갯수를 y 축은 문자 높이로 정의한다. 세 번째, 접합 문자의 맨 오른쪽 측면 윤곽 히스토그램과 단일 문자들의 오른쪽 측면 윤곽 히스토그램들과 비교한다. 단일 문자들의 네 방향의 측면 윤곽 히스토그램은 프로토타입(prototype)에 저장되어 있다. 접합 문자의 맨 오른쪽 측면 윤곽 히스토그램(q)와 프로토타입에 저장된 히스토그램(P)사의 거리 차, $d(q,P)$ 는 다음과 같이 구해진다.

$$d(q, P) = \sum_{i=1}^{N} \left| x_{iq} - x_{iP} \right| \tag{5}$$

위 식(5)에서 N 은 정규화된 문자 높이이고, x_{iq}

는 접합문자 히스토그램의 y 축 값이고, x_{iP} 는 프로토타입으로서 단일문자 히스토그램의 y 축 값이다.

네 번째, 가장 적은 거리차, $d(q,P)$ 를 가지는 프로토타입의 단일 문자가 접합 문자의 맨 오른쪽 문자의 일차 후보가 된다. 다섯 번째, 일차 후보 단일 문자의 폭으로 접합 문자를 초기 분할한다. 정규화된 단일 문자 폭은 프로토타입에 저장되어있다. 또한 이때 분할 비용도 고려한다. 분할 비용에 대해서는 다음 섹션에서 논의한다. 여섯 번째, 분할된 패턴으로부터 왼쪽 방향의 측면 윤곽, 위쪽 방향의 측면 윤곽, 아래쪽 방향의 측면 윤곽의 히스토그램을 구하고 일곱 번째, 분할된 패턴의 네 측면 윤곽 패턴의 히스토그램들과 후보 단일 문자의 히스토그램들과 각각 비교한다. 여덟 번째, 이렇게 구한 거리 차의 네 가지 합이 설정된 임계값보다 적으면 다음 단계로 넘어가고, 임계값보다 크면 네 번째로 돌아가 다음 후보 단일 문자를 가지고 위 단계들을 반복한다. 아홉 번째, 접합 문자에서 분할된 패턴을 제거한다. 제거된 패턴은 우리가 원하는 단일 문자이다. 열 번째, 접합 문자에 남아 있는 패턴이 있으면 두 번째로 올라가서 위 단계들을 반복한다. 만약 남아 있는 패턴이 없으면 접합 문자는 모두 단일 문자로 분할(제거)되었다.



[그림 3] 측면 윤곽 패턴을 이용한 접합 문자 분할의 전반적인 구성도

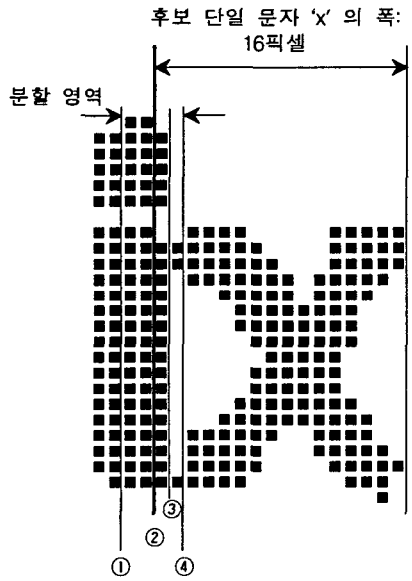
4. 문자 분할 비용

접합 문자를 크기 정규화를 하더라도 잡음, 스캔 해상도 등등에 의해 접합 문자에 속에 있는 단일 문자의 폭은 프로토타입에 저장된 같은 클래스의 단일 문자 폭과는 약간의 차이가 존재한다. 따라서 위에서 다섯 번째, 후보 단일 문자의 폭으로 접합 문자를 초기 분할을 할 때, 후보 단일 문자의 폭은 접합 문자의 분할 패스를 제공하는 것이 아니라 분할할 가능성이 있는 일정 영역을 제공한다. 최종적인 분할 패스는 이 분할 영역에서 분할 비용을 고려하여 결정된다. 본 연구에서 분할 비용은 <표 1>에서와 같이 정의한다. 즉 패턴을 구성하는 두 흑색 픽셀을 서로 분리할 때 분할 비용은 1로 증가되고 그 외에는 증가되지 않는다.

[그림 4]는 분할 비용을 고려하여 접합 문자를 분할하는 예를 보인다. 그림 4의 접합 문자 'ix' 를 입력으로 해서 앞에서 설명한 분할 과정의 첫 번째에서 네 번째까지 수행하면 접합 문자의 맨 오른쪽 문자 후보는 'x' 가 된다. 분할 비용의 고려 없이 프로토타입에 저장된 단일 문자 'x' 의 문자 폭 16 픽셀을 가지고 오른쪽에서 왼쪽으로 16 픽셀만큼접합 문자를 분할할 경우 부정확한 문자 분할이 된다(라인 ②). 따라서 프로토타입에 있는 문자 후보 'x' 의 문자 폭 16픽셀은 단지 참조 라인 ②를 지정한다. 이 참조 라인 전후로 그림과 같이 2픽셀 라인을 추가하여 분할할 가능성이 있는 분할 영역(segmentation neighborhood)을 설정한 후(라인 ①에서 라인 ④까지), 분할 영역 안에 가능한 모든 라인의 분할 비용을 계산한다. 라인 ③과 라인 ④가 같은 최소 분할 비용 3을 갖는데, 라인 ③이 참조 라인 ②에 더 근접함으로 라인 ③이 최종적으로 접합 문자 'ix' 의 분할 패스로 결정된다. 즉 최종적인 분할 패스는 최소의 분할 비용을 가지면서 참조 라인에 가장 가까운 라인이 선택된다

<표 1> 분할 비용의 정의

두 인접 픽셀의 밝기		분할 비용
흑색	흑색	1
흑색	백색	0
백색	흑색	0
백색	백색	0

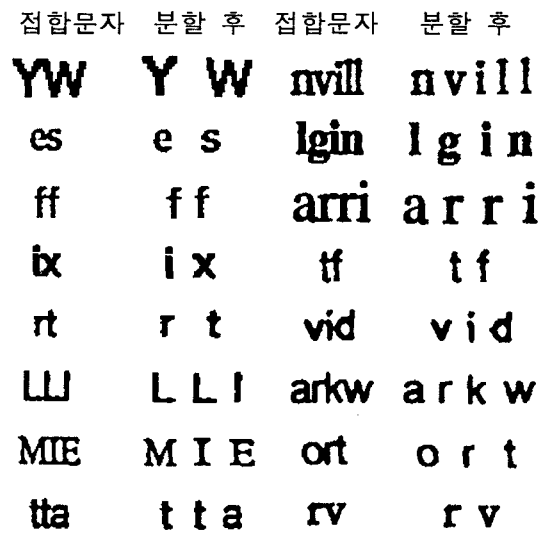


[그림 4] 접합 문자의 분할 비용(라인 ①의 분할 비용은 22, 라인 ②의 분할 비용은 21, 라인 ③의 분할 비용은 3, 라인 ④의 분할 비용은 라인 ③과 같은 3)

5. 실험과 결과

본 연구에서 제안된 문자 분할 알고리즘은 SUN Workstation에서 C언어로 구현되어져 여러 접합 문자들에 대해 실험되어졌다. 접합 문자 'rf',

'fr', 'fi', 'tt', 'ff', 'rt', 'rp' 와 'll' 등은 접합 문자 폭이 단일 문자 폭과 적거나 같아서 접합 문자로 발견되기도 어려울 뿐 아니라 분할 패스를 발견하는 것도 어렵다[11]. 본 연구에서 제안된 문자 분할 알고리듬은 위의 접합 문자들도 성공적으로 분할하였다. 또한 접합 문자가 단일 문자의 패턴과 흡사한 'Ll'('U' 와 흡사), 'cl'('d' 와 흡사), 'rn'('m' 과 흡사)등도 성공적으로 분할하였다. 그림 5는 본 연구에서 제안된 문자분할 알고리즘으로 가변피치 폰트로 프린트된 접합 문자들의 성공적 문자 분할의 예를 보여준다.



[그림 5] 접합 문자의 성공적 분할

학술지에 보고되어지는 문자 분할의 결과들은 서로 다른 입력에 대한 결과이기 때문에 서로 성능을 비교하기가 어렵다. 또한 문자 분할 알고리즘의 성능은 "정확히" 분할된 문자의 갯수에 의해 측정될 수 있는데, "정확히" 라는 용어는 문자 분할에서 보편적이고 객관적이 측정치가 될 수 없다[3]. 본 논문에서는 새로 개발된 문자 분할 알고리즘의 성능을 비교, 측정하기위해, 비록 실험 결과가 특정 OCR 시스템 알고리즘에 의해 제약되고 영향을 받지만, 문자 분할을 전체 OCR 시스템의 내부 인자로 간주하면 OCR 시스템의 성능 향상이 문자 분할 알고리즘의 성능 향상을 의미한다고 결론짓는다. 본 논문에서 사용한 OCR 시스템은 U.S. 메일에서 주소를 자동으로 인식하여 메일을 자동으로 도착지별로 분류하는 시스템(Envelope Reader System)이다. 이와 비슷한 OCR 시스템으로는 참고문헌[12]이 있다. 이 시스템은 문자 분할을 위해 윤곽선 분석과 피치(pitch) 추정 방법을 병행하여 사용하였다. 표 2에서 보이듯이 기존의 방법을 문자 분할로 사용했을 때 3359개의 메일 중 2315개의 메일을 성공적으로 분류(68.92%)하였다. 기존 문자 분할 모듈을 본 논문에서 제안한 측면 윤곽을 이용한 문자 분할 모듈을 대체한 후 똑 같은 실험을 한 결과 2690개의 메일을 성공적으로 분류(80.08%)하였다. 즉 측면 윤곽을 이용한 문자 분할 알고리즘은 실험에 사용된 OCR 시스템의 분류 성능을 11.16% 향상시켜 기존의 문자 분할에 비해 효율적임을 증명하였다. 또한 <표 2>는 제안된 문자 분할 알고리즘이 OCR 시스템의 거부율(reject)과 에러율(error) 또한 향상시켰음을 보인다.

<표 2> 측면 윤곽을 이용한 문자 분할 알고리즘에 의한 OCR 시스템 성능 변화

Envelope Reader System	윤곽선 분석과 피치추정을 병행한 문자 분할을 내장	측면 윤곽을 이용한 문자 분할 알고리즘을 내장
Sorting (out of 3359 envelopes)	2315(68.92%)	2690(80.08%)
Reject	902(26.85%)	544(16.20%)
Error	142(4.23%)	125(3.72%)

6. 결 론

본 논문에서는 인쇄된 영문자에서 접합 문자를 분할하는 새로운 알고리듬이 제안하였다. 이는 특징을 기반으로 한 접근 방식(feature-based approaches)과 인식을 기반으로 한 접근 방식(recognition-based approaches)의 중간적인 형태로 두 방식의 단점을 보안하고 장점을 취한다. 본 논문에서는 이를 위해 네 방향의 측면 윤곽 패턴 추출과 분할 비용을 정의했다. 실험은 특징을 기반으로 한 접근 방식인 윤곽선 분석과 피치(pitch) 추정 방법보다 높은 성능을 보임을 특정 OCR 시스템을 이용하여 간접적으로 증명하였다. 그러나 단점으로 기존의 윤곽선 분석과 피치(pitch) 추정 방법보다 처리 시간이 더 걸렸는데 이를 개선하기 위한 방법이 향후 연구되어져야 한다. 본 연구에서 제안된 문자 분할 알고리듬은 영문자에 실험되어졌지만, 한글의 문자 분할에도 적용되어질 수 있으며 그 외 다른 언어의 문자 분할에도 응용되어질 수 있다.

감사의 글

본 연구는 상명대학교 교내 연구비 지원으로 수행되었음.

참고 문헌

[1] Bokser M., "Omnidocument Technologies," *Proceedings of the IEEE*, Vol.80, No.7, pp. 1066-1078, 1992.

[2] Fujisawa H., Y. Nakano, and K. Kurino, "Segmentation Methods For Character Recognition: From Segmentation To Document Structure Analysis," *Proceedings of the IEEE*, Vol.80, No.7, pp. 1079-1092, 1992.

[3] Y. Lu, "Machine printed character segmentation-an overview," *Pattern Recognition*, Vol. 28, No. 1, pp. 67-80, 1995.

[4] Yi Lu, et al., "An accurate and efficient system for segmenting machine-printed text," U.S. Postal Service 5th Advanced Technology Conference, Washington D.C., November, Vol.3, pp.93-105, 1992.

[5] S. N. Srihari, Y.C. Shin and et al., "A System to Read Names and Addresses on Tax Forms," *Proceedings of the IEEE*, Vol. 84, No. 7, pp. 1038-1049, 1996.

[6] 장승익, "히스토그램 분석 기반의 인쇄체 문자열 분할 방법," *한국정보과학회-03 가을학술발표논문집(2)*, pp. 532-534, 2003.

[7] 권숙연, "차량 번호판의 영역 추출 및 문자 분할에 관한 연구," *한국지능정보시스템학회*, pp. 457-462, 2000.

[8] V.A. Kovalevsky, "Image Pattern Rcognition," *Springer-Berlin*, 1980.

[9] R.G. Casey and G. Nagy, "Recursive segmentation and classification of composite character patterns," *International Conference on Pattern Recognition*, pp. 1023-1026, 1982.

[10] S. Tsujimoto and H. Asada, "Resolving Ambiguity in Segmenting Touching Characters," *1st International Conference on Document Analysis and Recognition*, pp. 701-709, 1991.

[11] Lu Y., "On The Segmentation Of Touching

Characters," *2nd International Conference on Document Analysis and Recognition*, pp. 440-443, 1993.

[12] 김호연, "서장 우편물 자동처리를 위한 우편 영상 인식 시스템," *한국정보처리학회*, 10권, 4호, pp. 429-442, 2003.

Abstract

Character Segmentation using Side Profile Pattern

Minchul Jung*

In this paper, a new character segmentation algorithm of machine printed character recognition is proposed. The new approach of the proposed character segmentation algorithm overcomes the weak points of both feature-based approaches and recognition-based approaches in character segmentation. This paper defines side profiles of touching characters. The character segmentation algorithm gives a candidate single character in touching characters by side profiles, without any help of character recognizer. It segments touching characters and decides the candidate single character by side profiles. This paper also defines cutting cost, which makes the proposed character segmentation find an optimal segmenting path. The performance of the proposed character segmentation algorithm in this paper has been obtained using a real envelope reader system, which can recognize addresses in U.S. mail pieces and sort the mail pieces. 3359 mail pieces were tested. The improvement was from 68.92% to 80.08% by the proposed character segmentation.

Key words : character segmentation, Optical Character Recognition(OCR), cutting cost, side profile, character prototype

* Department of Computer System Engineering, Sangmyung University