찌그러짐을 고려한 저품질 인쇄체 문자인식

김성원, 김형원, 양윤모
고려대학교 전자 및 정보공학부
전화 : 041-860-1780 / 핸드폰 : 017-540-5430

Low-Quality Printed Character Recognition Considering of image Blurredness

Seong Won Kim, Hyung Won Kim, Yun Mo Yang
Dept. of Electric & Information Engineering, Korea University
E-mail : swkim@hard.korea.ac.kr

Abstract

Character recognition has already been studied in a lot of fields. But, if input-characters have noise in practical application system, the ability decreases markedly. Special consideration should be taken into account in the recognition of blurred data.
This paper proposes low-quality printed character recognition methods that extracts blurred parts of the character image, deletes them and carry out accurate character recognition.

I. 서론

문자 인식은 시각 정보를 통하여 문자를 인식하고 의미를 이해하는 사람의 능력을 컴퓨터로 실현하려는 시도로서 이미 많은 연구가 진행되어왔다. 하지만, 문자인식의 실제 응용 시스템에서 입력문자에 잡음이 발생한 경우 그 성능은 떨어지게 된다. 입력문자에 잡음이 발생하는 원인은 매우 다양하다. 잉크젯 프린터를 사용하여 문서를 출력하는 경우, 잉크가 종이에 번져 원래 문자영상에 찌그러짐이 발생하게 된다. 문서를 팩스, 복사기 등으로 전송할 경우에도 문자영상에 열화가 발생할 수 있으며, 우편봉투와 같이 겉에 투명한 비닐이 씌워져 있는 경우에도 문자입력시 빛의 반사에 의한 잡음이 발생하게 된다. 또한 작은 크기로 문자를 출력한 경우와 저해상도의 스캐너로 문자영상을 입력받는 경우에도 잡음이 발생하게 된다. 이러한 문자영상의 찌그러짐은 문자인식에 있어서 오인식의 주원인이 되며, 문자인식의 실제 응용에 있어서 큰 문제점이 되고있다. 또한 문자에 발생하는 잡음은 발생 원인이 다양하고, 예측이 불가능하기 때문에, 표준 데이터 베이스를 만들어 대처하기가 불가능하다. 이와 같이 실용적인 문자인식 시스템에 있어서 입력 문자영상에 발생할 수 있는 여러 가지 잡음에 대처할 수 있는 저품질 문자인식은 필수적이다.

본 논문에서는 문자영상의 찌그러진 부분을 추출하는 방법을 제시하고, 찌그러짐이 발생한 영역을 보완하여 고정도의 문자인식을 수행하는 방법을 제안한다.

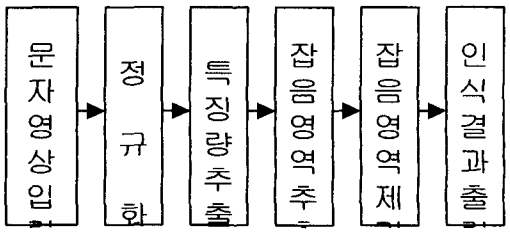


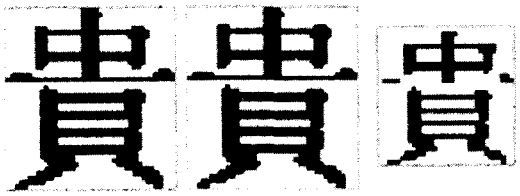
그림 1. 찌그러진 문자의 인식 흐름도

II. 특징량 추출

2.1 히스토그램을 이용한 선형 정규화

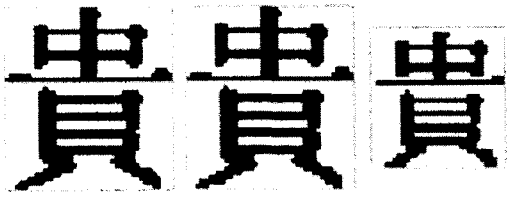
일반적인 정규화 알고리즘은 입력영상을 일정한 크기의 형상으로 선형 변형시키는 선형 정규화와 영상의 특징을 고려하는 비선형 정규화가 있다. 본 연구에서는 선형 정규화 방법을 수정, 보완하여 이용하였다. 이때, 입력 영상이 정규화 영상의 크기보다 큰 경우, 중요한 정보를 갖는 단선성분이 제거되는 경우가 있어서 오인식의 원인이 되고있다. 이런 선형 정규화의 단점을 보완하기 위하여 문자영상의 히스토그램의 변화량을 이용하였다.

히스토그램의 변화량을 이용하는 방법은, 우선 입력 문자영상의 행, 열 방향의 히스토그램을 구하고, 앞 행(열)과 뒷 행(열)의 히스토그램과의 차이를 구하여 각 행(열)의 변화량을 구한다. 이 때에 선형 정규화에 의하여 삭제될 행(열)이 정하여 졌을 때, 삭제될 행(열)의 앞 행(열), 뒷 행(열)과의 변화량을 비교하여 변화량이 적은 행(열)을 삭제한다.



(a) 행 정규화 (b) 열 정규화 (c) 정규화 영상

그림 2. 기존의 선형정규화 방법(40×42→32×32)



(a) 행 정규화 (b) 열 정규화 (c) 정규화 영상

그림 3. 히스토그램을 이용한 선형정규화 방법

그림 2는 임의의 입력영상(42×40)에 대한 일반적인 선형 정규화를 나타낸다(본 논문에서 정규화 영상의 크기는 32×32로 한다). 그림 2(a), (b)는 각각 가로, 세로 방향의 정규화를 나타내고, 흐리게 표시된 선이 삭제가 될 행 또는 열이 되며 (c)는 정규화된 영상을 나타낸다. 이때 그림 2의 (c)에서 영상의 중간부분의 단선 성분이 사라져 이후 특징량 구성 및 인식 단계에서 악영향을 끼치게 된다. 하지만, 그림 3의 (c)는 히스토그램 변화량을 이용하여 중요 단선성분의 삭제를 방지한 것을 알 수 있다.

2.2 특징량 구성

문자영상의 정규화 이후 특징량 구성을 위하여 특징소를 추출한다. 문자의 윤곽선이 갖는 방향정보에 중점을 두어 4가지 방향성분(0^0, 45^0, 90^0, 135^0)을 추출하게 된다. 이렇게 추출된 특징소는 그 위치에 따라 1/2씩 중첩된 9개의 소영역으로 반복 분할하면 각기 서로 다른 위치별 가중치를 갖고 4차원, 36차원, 324차원의 특징벡터로 구성이 된다(그림 4). 이러한 계층구조에 의하여 구성된 특징벡터에 의해서 문자 영상의 중앙부분을 강조하고, 어느 정도의 잡음에 대해서도 안정적인 인식이 가능하다.[1][2]

본 연구에서 특징량은 324차원(9^2 × 4)으로 하였다.

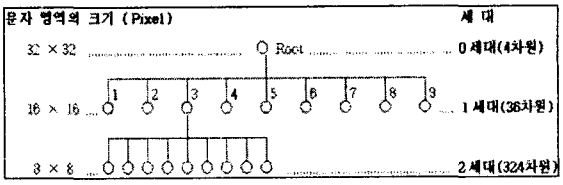


그림 4. 9진트리를 이용한 특징량 구성

III. 찌그러진 영역 추출 및 보완

3.1 문자 영상의 찌그러진 영역 추출

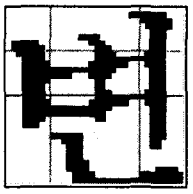
문자영상의 찌그러진 영역을 찾기 위한 방법으로 세선화 알고리즘을 사용하는 논문이 소개되고 있다[3]. 하지만, 세선화를 이용할 경우 세선화 횟수를 조절하기가 어렵고, 시간이 많이 걸린다는 단점이 존재한다.

본 논문에서는 입력문자의 특징량 구성단계(2.2절)에서 작성된 9개의 중첩된 소영역과 미리 준비한 표준 특징DB와의 거리 정보를 이용하여 찌그러진 영역을 추출하는 새로운 방법을 제안하고, 세선화에 의해 발생하는 단점을 제거하였다.

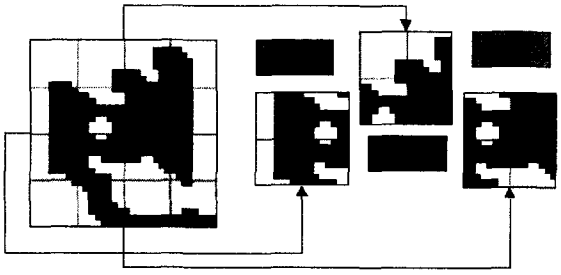
(1) 찌그러진 문자의 발생 원인과 특징

찌그러짐의 발생원인을 살펴보면 인식할 문자에 이미 찌그러짐이 존재하는 경우와 문자의 입력단계에서 잡음에 의해서 발생하는 두가지 경우가 있다. 전자의 예로는 문자를 잉크젯 프린터로 출력하거나, 크기를 작게 출력하여 발생하는 경우이며, 후자의 예로는 입

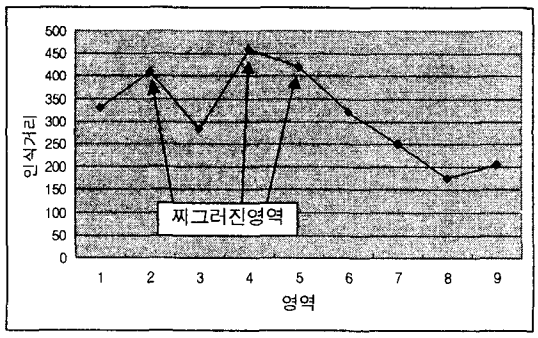
력 해상도를 낮게 함으로서 발생하는 경우가 있다. 이 때, 찌그러짐은 문자의 어느 한 부분에 집중적으로 발생을 하는 경우(그림 5.b 참조)가 대부분이며, 문자가 전체적으로 찌그러진 경우에는 인식이 불가능하다.



(a) 깨끗한 "변"字 영상



(b) 찌그러진 "변"字 영상과 찌그러진 소영역



(c) 9개 소영역의 인식 거리(1위 인식 후보문자)

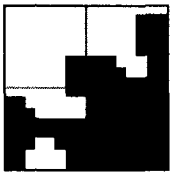
그림 5. 찌그러진 문자의 예

(2) 찌그러진 영역의 추출방법

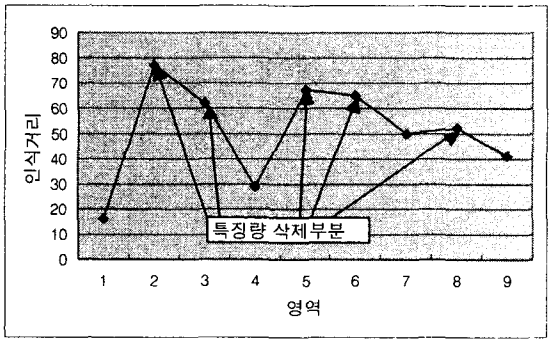
인식할 문자영상을 정규화(2.1절)하고, 9진트리를 이용하여 324차원의 특징량을 구성(2.2절)한 후, 미리 준비한 표준 특징DB와 거리(City Block Distance)비교를 함으로써 1위 인식 후보문자를 찾을 수 있다. 이 때, 인식한 1위 후보문자의 중첩된 9개 소영역(전체 문자영상의 1/4크기)에 대한 인식거리를 알 수 있고, 소영역의 인식거리가 클수록 1위 후보문자와 다르다는 것을 나타내며, 찌그러진 영역을 정확히 찾을 수 있다. (그림 5.(b)(c) 참조).

3.2 찌그러진 영역의 보완

추출된 찌그러진 소영역은 전체 크기(32×32)의 1/4(16×16)에 해당하며, 36차원의 특징량을 갖는다. 소영역은 다시 9개의 중첩된 작은 소영역(8×8)으로 나누어지며, 작은 소영역마다 4차원의 특징량을 갖게 된다(그림 4참조). 이 때, 9개의 작은 소영역에 대한 1위 인식 후보문자의 인식거리를 알 수 있게되며, 거리가 클수록 1위 인식 후보 문자와는 다르다는 것을 알 수 있다.



(a) 찌그러진 "변"字의 4번영역



(b) 찌그러진 4번영역의 9개 소영역의 인식 거리

그림 6. 찌그러진 소영역의 인식 거리

찌그러진 영역의 9개의 작은 소영역(8×8)중에서 인식거리가 큰 부분을 제거함으로서, 보다 신뢰할 수 있는 문자 인식이 가능하다. 가장 많이 찌그러진 영역에서는 5개(그림 6.b참조), 2번째로 찌그러진 영역에서는 4개, 3번째로 찌그러진 영역에서는 3개의 작은 소영역의 특징값을 제거하여 1위 인식문자를 판별한다. 예를 들면, 그림 5(b)에서 4번영역에서는 5개, 5번영역에서는 4개, 2번영역에서는 3개의 소영역을 제거한다. 이 때, 12개의 작은 소영역이 제거되므로 274차원(324 - 12×4)으로 인식이 이루어진다.

IV. 인식 실험

4.1 데이터 베이스

표준 특징DB는 14개의 문자Set으로 만들어진다. 하

나의 문자Set은 총 715문자이며 한글 상용문자, 영어 대·소문자, 숫자, 기호로 구성된다. 글자 크기는 Ms-word에서 10point로 출력하였고, Epson GT9500 스캐너를 사용하여 300dpi의 해상도로 입력받아 만들었다. 이 때, 잉크젯 프린트로 출력한 문자Set이 6개, 레이저 프린트로 출력한 문자Set이 8개이다.

실험에 사용한 찌그러진 문자Set은 한글 상용문자로만 이루어진 625문자이며, 7point크기로 잉크젯 프린터로 출력한 뒤, 150dpi의 해상도로 입력받아 구성하였다. (폰트 : 바탕체)

비교 실험을 위하여 표준 특징DB에 포함되지 않은 정상적 문자Set 625문자를 만들었다.(크기 10point, 300dpi, 레이저프린터)

4.2 실험방법 및 결과

문자를 입력받아 정규화, 특징량 추출 과정을 수행한 뒤 표준 특징DB와 비교하여 1순위 후보문자를 판별하고(기존방법 : 324차원), 찌그러진 영역을 검출, 보완한 뒤 1위 인식 문자를 판별한다(제안방법 : 276차원). 평가 함수로는 City -Block 거리를 사용하였다.

인식순위	찌그러진 문자Set		정상적인 문자Set	
	기존방법	제안방법	기존방법	제안방법
1위	401	441	621	621
2위	74	59	4	4
3위	29	23	·	·
~ 5위	40	44	·	·
~ 10위	16	14	·	·
그외	65	44	·	·
계	625	625	625	625

표 1. 실험결과(기존방법과 제안방법 비교)

표1에서와 같이 기존방법에 비해 찌그러짐을 고려한 경우 1위 인식은 40문자가 증가하였고(6.4%증가), 10위까지의 인식율은 92.96%를 나타내었다(기존방법 : 89.6%). 한편, 정상적인 문자Set의 경우 특징량의 차원이 줄어 들었지만(324 → 276차원) 인식율에 변화가 없어 제안한 방법이 효과적임을 알 수 있다.

또한, 5위 이내에 들어오지 못한 문자의 경우 사람의 눈으로도 판별하기 어려울 정도로 찌그러짐이 심한 경우가 대부분이었다(그림 7참조). 이런 경우 1위 인식거리 또는, 1위 인식문자와 2위 인식문자와의 거리의 비율을 기준으로 인식불가 판정(Reject)을 할 수 있다.

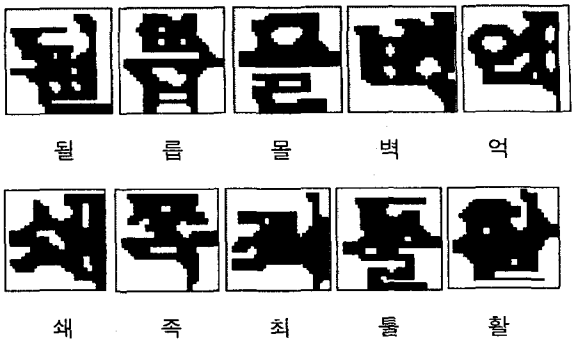


그림 7. 찌그러진 문자예(5위이후로 인식한문자)

V. 결 론

입력문자 영상에 잡음이 발생하여 찌그러진 경우, 대처할 수 있는 저품질 인쇄체 문자인식에 관하여 연구하였다. 이러한 찌그러진 문자인식에 관한 연구는 실제 응용 시스템 구현에 있어서 필수적이다.

기존에 사용하던 특징량 구성방법과 표준 특징DB와의 거리계산을 응용하여 문자영상의 찌그러진 영역을 정확히 찾아내었고, 찌그러진 부분의 소영역 중 거리가 큰 영역을 제거함으로써 보다 정확한 문자인식 방법을 제안하였다. 또한, 찌그러지지 않은 문자Set과의 비교실험을 통하여 본 논문의 타당성을 제시하였다.

향후 발전방향으로, 표준 특징DB에 사용할 문자Set을 확보하여 새로운 평가함수(Mahalanobis Distance등)를 사용한다면 찌그러진 영역의 특징량에 적절한 가중치를 부여함으로써 보다 고정도의 문자인식을 실행할 수 있으며, 폰트가 달라서 생기는 악영향도 줄일 수 있다.

또한, Reject 기준을 설정하면 우편봉투인식, 전표인식 등 실제 응용분야에서의 사용이 기대된다.

참고문헌

[1] 강선미, 이기용, 황승욱, 양윤모, 김덕진," 고속문자인식을 위한 특징량 추출에 관한 연구", 전자공학회 논문지 29B, Vol. 11, pp.1047~1056, 1992.

[2] 송효섭, 장세진, 신병주, 양윤모," 손의 형상과 움직임 방향 정보를 이용한 수화인식", 정보과학회 논문지(B), Vol. 26, No. 6, pp,804~810 1999.

[3]大町真一郎,つぶれを考慮した低品質印刷文字の高精度認識, 電子情報通信學會 論文誌,96.9.Vol.J79-D-II No9, pp1534 - 1543.