Journal of Korean
Data & Information Science Society
2003, Vol. 14, No.3 pp. 525~533

Feature Extraction Method for the Character Recognition of the Low Resolution Document[1)]

Daehak Kim[2)] · Hyoungchul Cheong[3)]

Abstract

In this paper we introduce some existing preprocessing algorithm for character recognition and consider feature extraction method for the recognition of low resolution document. Image recognition of low resolution document including fax images can be frequently misclassified due to the blurring effect, slope effect, noise and so on. In order to overcome these difficulties in the character recognition we considered a mesh feature extraction and contour direction code feature. System for automatic character recognition were suggested.

Keywords : 문자인식, 저해상도, 중앙값 필터링, 영상 정규화, 기울기 보정, 피쳐 추출, 비쥬얼 프로그래밍

1. 서 론

오늘날의 멀티미디어(multimedia)시대에 있어서 영상을 컴퓨터를 이용하여 인식하고 처리하는 영상처리(image processing)분야는 의학, 통신 등 광범위한 분야에 걸쳐 응용되고 있다. 또한 인터넷과 정보통신 분야의 발달로 인해 특정한 자료나 이미지 파일의 전송이 훨씬 간편해지고 있다. 이러한 통신기술의 발달과 더불어 신속하고 편리하게 정보를 교환하기 위한 수단으로 널리 사용되어 왔던 팩스를 컴퓨터에 부착하여 함께 사용 할 수 있게 되어 팩스의 활용 분야가 확대되고 그 활용도가 증가되고 있는 실정이다.

그러나 팩스 등을 통하여 수신된 저해상도영상은 180DPI(dots per inch)정도의 저해상도로 고속전송 되어 영상의 선명도가 매우 낮고 문자의 모양에 심한 왜곡이 생길

---

1) 본 연구는 2001학년도 대구가톨릭대학교 연구비 지원에 의한 것임
2) 경북 경산시 하양읍 금락리 330-1 대구가톨릭대학교 정보통계학과 교수
E-mail : dhkim@cataegu.ac.kr
3) 경기도 평택시 용이동 평택대학교 정보통계학과 교수
E-mail : jhc@ptuniv.ac.kr

수도 있다. 또한 문자열을 추출할 때도 가로줄이 기울어지거나 문자가 추출될 때 흐려짐(blurring)현상으로 인하여 문자들 간에 접촉이 상당히 많이 생길 수 있고 영상자료(image data)로서의 흐린 이미지를 동반한 자료의 가공처리가 그리 쉬운 일만은 아니다. 따라서 저해상도 영상자료의 내용을 시스템이 자동으로 인식하여 텍스트 정보로 변환한 후 컴퓨터에서 사용 가능한 각종 파일로 가공처리 할 수 있는 방안이 마련된다면 이 또한 효과적으로 활용될 수 있을 것이다. 선진국의 경우 저해상도 문서의 영상인식 소프트웨어가 이미 상품화되어 그 활용도가 상당히 높지만 국내에서는 효과적인 한글 저해상도 문서인식을 위한 연구의 필요성이 요구되고 있다. 특히 저해상도 문서의 인식에서는 인쇄된 문자(optical character)의 인식과 달리 문자들 간의 빈번한 접촉현상으로 인하여 추출된 개별 문자 영역에는 많은 잡음이 첨가될 수 있으므로 그 인식을 위한 피쳐(feature)추출도 기존의 문자 인식에 비해 잡음에 둔감해야 필요가 있다. 인쇄된 문자의 인식은 Kahan등(1987)에 의해 연구된 바 있다.

본 논문에서는 저해상도 문서의 영상정보를 고성능으로 인식하기 위하여 기존의 문자인식의 사전단계(preprocessing)에서 주로 사용되는 알고리즘들을 소개하고 흐려짐 현상이 있는 저해상도 문서인식에서 문자 영상에 포함된 잡음이나 굴곡부분들에 대해 크게 영향을 받지 않으면서도 문자인식이 가능한 피쳐(feature)을 추출하는 방법을 고려하였다. 또한 이를 활용한 자동인식시스템의 구성도를 제안하였다.

2. 영상 인식의 전처리 구성

1절에서 설명하였듯이 180DPI 정도만의 해상도로 고속 송신된 문서인식의 어려움을 극복하고 또한 문자들 사이에 자주 발생하는 접촉문제를 해결하기 위하여 본 절에서는 자주 사용되는 영상인식의 전처리(preprocessing)과정을 소개하고 본 논문에서 적용한 방법을 설명하였다.

2.1 히스토그램 평활화

영상개선(image enhancement)에 주로 사용되는 일반적 방법으로는 히스토그램 평활화(histogram equalization)와 레벨 정규화(gray level normalization)등이 있다. 히스토그램 평활화는 주어진 영상을 변환하여 결과영상의 명도가 균일한 분포를 가지도록 하는 방법으로 명도가 한쪽으로 치우쳐 있거나 균일하지 못한 영상에 적용되는 방법이다. Gose등(1996)은 레벨 정규화의 다양한 방법을 서술한 바 있다.

[그림 1] 수신된 저해상도 문서의 일부

다음 [그림1]은 팩스를 통하여 얻어진 저해상도 문서의 일부를 보여주고 있다. 이 영상이미지의 명도 히스토그램과 평활화를 통한 영상의 명도 히스토그램을 [그림 2]에 나타내었다. [그림 2]에서 나타나듯이 평활화를 거친 영상이미지히스토그램에서는 오른쪽 끝(그레이 레벨 255)에 많이 분포함으로서 평활전 보다 명도가 많이 균등하게 분포되어 있음을 알 수 있다.

[그림 2]. 명도 히스토그램(원영상(왼쪽), 평활화된 영상(오른쪽))

2.2 이진화 과정

인식대상 문자영역을 세그멘테션 하기 전에 영상자료는 반드시 이진화(양자화)과정을 거쳐야 한다. 각 픽셀의 농담을 이산적인 정수값으로 변환하는 이진화과정은 0부터 255까지의 그레이 수준을 적절한 경계값(threshold)을 이용하여 흑과 백의 두가지 색으로 나타낸다. 다음 [그림 3]은 이진화 과정을 거친 획득영상을 나타낸다. 이진화 과정의 결과는 명도의 수준이 0(흑)과 255(백)에만 분포하고 있으며 희게 나타난 부분이 검게 나타난 부분보다 훨씬 많은 도수를 나타냄을 알수 있다.

[그림 3]. 이진화 과정의 결과

2.3 기울기 보정 알고리즘

획득된 영상이 기울어져 있을 경우 기하학적 회전 변환을 통하여 기울지 않은 영상을 만들 수 있다. 문자가 기울어진 정도를 나타내는 기울기(slope)를 찾아내는 방법으로 투영결과를 이용하는 방법, Hough변환 결과를 이용하는 방법 그리고 최단 이웃군집(nearest -neighbour clustering)방법 등이 알려져 있다. Daniel등(1994)은 흑백영상에 대하여 기울어진 기울기를 구하는 방법을 연구한 바 있다. 기울어진 각(θ)를 찾아내면 각 픽셀(pixel) (a, b)에 대해 다음과 같은 식(2.1)로 기울기가 보정된 픽셀 (a', b')를 얻을 수 있다.

$$a' = a\sin(\theta) - b\cos(\theta),\quad b' = x\cos(\theta) + y\sin(\theta) \tag{2.1}$$

2.4 이진영상 잡음제거

[그림 5]에서 볼수 있듯이 이진화 영상에는 많은 잡영(noise)들이 포함되어 있다. 이들 잡영(noise)들을 효율적으로 제거해야 문자 영역의 세그멘테이션과 인식과정에서 오류를 최소화 할 수 있다. 잡영제거 알고리즘들은 이동평균방법을 이용할 수도 있지만 본 연구에서는 중앙값필터링(median filtering)방법을 이용하여 잡영을 제거하였다. 다음 [그림 4]는 중앙값필터링 방법을 나타내고 있다. 중앙값필터링은 변환 하고자 하는 픽셀과 이를 둘러싸고 있는 8개 픽셀의 중앙값을 취한 명도로 변환하는 방법이다.

12	3	6
2	9	5
25	7	15

2	3	5	6	7	9	12	15	25

	7	

[그림 4].메디언 필터링의 구조도

[그림 5]는 2회의 중앙값필터링 방법을 거친 잡영이 제거된 영상을 보여주고 있다. 이 그림을 통하여 중앙값필터링 방법을 통하여 효율적으로 잡영이 제거되었음을 알 수 있다.

[그림 5]. 원영상(왼쪽)과 잡영이 제거된 영상(오른쪽)

3. 문자인식을 위한 피쳐추출

2절에서와 같은 영상의 사전처리가 완료되고 나면 이제 문자인식을 고려하여야 한다. 그러나 문자인식을 위하여 먼저 문자가 포함되어 있는 영역의 세그멘테이션(segmentation)과 나아가 피쳐 까지 추출(feature extraction)하여야 문자로 인식할 수 있다. 본절에서는 문자의 영역추출과 피쳐추출 방법을 고려하였다.

3.1 문자영역 세그멘테이션

문자영역을 세그멘테이션 하는 방법에는 기능적 해석법(Functional layout analysis)과 구조적 해석법(structural layout analysis)을 이용할 수 있다. 기능적 해석법은 대상 문서의 사전 layout규칙 정보를 이용하여 구조적인 블록에 라벨링을 수행한다. 반면 구조적 해석법은 영상의 포맷에 따른 사전 정보 없이 독립된 단어, 문자열, 로고 심볼 등을 추출한다. Hori등(1995)은 “박스 유도 추론(box driven reasoning)"에 근거한 로버스트 구조적 방법을 연구한 바 있다. 구조적 해석방법에는 하향(top-down)방법과 상향(bottom-up)방법이 있다. 하향방법은 문서를 크게 한 두 개의 열(column)블록으로 구분하고 각 열블록은 다시 문단 블록으로 나누어지며 각 문단 블록들은 다시 문자열로 나누어지게 된다. 상향방법은 문자영역에서 단어영역으로 합쳐지고 이들은 다시 더 큰 영역으로 합쳐진다. 하향방법의 알고리즘들이 비교적 안정되게 세그멘테이션 해내고 있으며 상향방법은 계산이 복잡해질 수 있다. 이 두 방법을 결합시킨 방법들도 많이 발표되었다.

3.2 문자영상 정규화

획득된 영상에는 다양한 폰트와 다양한 크기의 문자들이 포함되어 있다. 따라서 이

들을 학습시킨 표본 문자와 비교하기 위해서는 입력 문자 영역을 학습시킨 표본문자의 크기로 정규화 시켜야 한다. 정규화 알고리즘은 크게 선형정규화와 비선형정규화로 구별된다.

선형정규화는 문자 영상의 가로 및 세로 길이의 비율에 따라 영상 매핑을 하며, 문자 영상의 확대 및 축소를 빠르게 수행할 수 있다. 그러나 확대나 축소 비율이 클 경우 문자 정보를 소실할 수 있다. 디지털화된 $[X\times Y]$크기의 문자영상 $f(x,y)$를 $[M\times N]$크기의 정규화영상 $f'(m,n)$으로 선형정규화 할 때 다음의 관계가 성립된다.

$$x=m/S_{x,}\quad y=n/S_y \tag{3.1}$$

이때 $S_x=X/M$ 이고 $S_y=Y/N$ 로서 원영상과 정규화 영상의 가로의 비와 세로의 비를 각각 나타낸다.

한편 비선형 정규화는 밀도 정보를 반영하여 문자 영상자체를 변형시켜 정규화한다. 영상을 비선형 정규화 시키면 정규화 이전의 초기 영상에서 밀도가 높은 부분은 확대가 되고 밀도가 낮은 부분은 축소된다. 비선형 정규화 알고리즘은 다음과 같다.

단계 1] 먼저 문자영상 $f(x,y)$에 대하여 수평 방향(x축)과 수직 방향(y축)의 선밀도를 계산한다. 선밀도는 수평 또는 수직 방향으로 주사선을 통과시켰을 때 만나는 "스트로크"의 개수를 의미한다.

단계 2] 계산된 선밀도에 대하여 선분의 가장자리나 획의 삐침 등에 의해 나타나는 불필요한 정보의 유입을 막기 위하여 중앙값 필터링을 수행한다.

단계 3] 이제 수평 방향(x축)과 수직 방향(y축)의 점밀도를 계산한다. 점밀도는 수평 또는 수직 방향으로 주사선을 통과 시켰을 때 만나는 흑색 화소의 개수이다.

단계 4] 선밀도와 점밀도를 합한 전체 밀도를 구한다. 수평방향의 전체밀도를 $D_h(y)$, 수직방향의 전체밀도를 $D_v(x)$로 표시하자. 그리고 이들을 수평, 수직방향에 대하여 합한 밀도를 $DH=\sum_{y=0}^{Y-1}D_h(y)$, $DV=\sum_{x=0}^{X-1}D_v(x)$로 각각 둔다.

단계 5] 이제 모든 m,n에 대하여 (3.2)을 만족시키는 x,y를 구하여 정규화매핑을 시도한다.

$$D_v(x)\ \geq\ (m-0.5)\frac{DH}{M}\quad ,\quad D_h(y)\ \geq\ (n-0.5)\frac{DV}{N} \tag{3.2}$$

3.3 피쳐 추출(feature extraction)

하나의 피쳐만을 이용해서 팩스 문서와 같은 저해상도의 이미지로부터 고성능으로 문자를 분류 해내는 것은 어렵다. 아직 모든 종류의 문자에 잘 적용되는 이상적인 피쳐를 찾아내지는 못하고 있다. 최근까지 문자인식을 위해 많은 피쳐들이 발표되었다. Leondes(1998)는 비선형 주성분분석을 이용한 피쳐 추출방법을 제안하였다. Fukunaga(1990)는 통계적 패턴인식에 대하여 연구하였으며 비모수적 판별분석을 위한 피쳐 추출을 제안한 바 있다. 많은 피쳐 추출방법 중에서 성능이 비교적 안정된 피쳐들인 다음과 같은 피쳐들을 이용하여 저해상도 문서의 문자 인식용 피쳐들을 추출 할 수 있다.

3.3.1 메쉬피쳐(mesh feature)

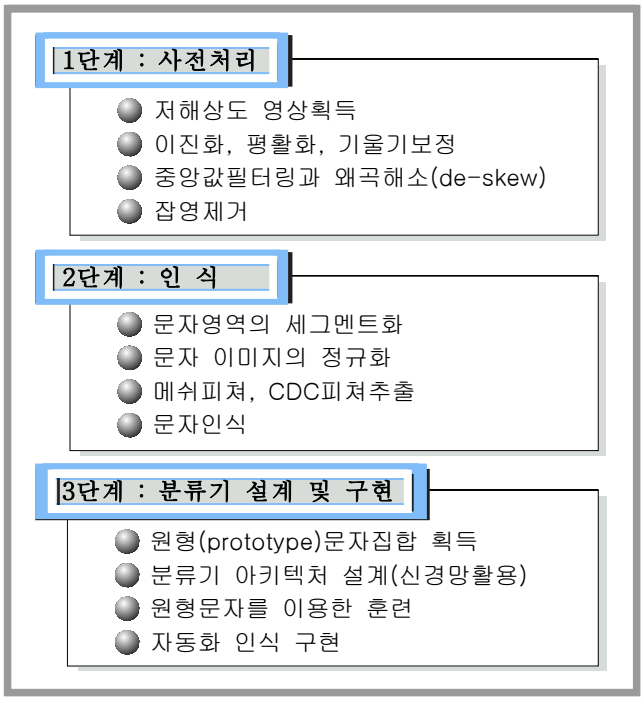
메쉬피쳐 추출방법은 이미지 영역을 몇 개의 사각형 영역으로 분할하고 분할된 각 영역에 흑색화소의 밀도(density)에 따라 결정한 값을 할당한다. 메쉬피쳐 추출의 주된 특징은 이미지를 피쳐 벡터로 변환할 때 비교적 연속적이다는 점이다. 눈으로 보기에 비슷한 이미지들은 기하학적으로 다를지라도 피쳐 공간(feature space)의 인접한 위치에 자리를 잡는다. 이렇게 되면 “filled-in hole”이나 “break in a crossbar”와 같은 국소 왜곡(local distortion)을 초래하는 잡음에 비교적 강건하게 된다. 눈으로 보기에 다른 이미지들을 동일한 벡터로 할당하지 않는다. 예로서 ″e″와 ″c″가 블러링되어 내부가 채워져 있을 때에는 기하학적으로(topologically) 동일한 구조를 갖지만 인간의 눈으로는 쉽게 구분할 수 있다. 이러한 구분 특징이 피쳐 변환에 그대로 반영된다.

3.3.2 CDC(Contour Detection Code)피쳐

CDC피쳐의 주된 특징은 각 화소에 대해 3x3크기의 마스크(mask)를 적용시켜 문자 경계선의 방향성분을 누적하여 추출할 수 있다는 점이다. 마스크는 중심 좌표를 제외한 8개 화소가 가질 수 있는 값의 종류가 256개가 된다. 각 마스크에 대해 수평(horizontal), 수직(vertical), 오른쪽 사선(right-up), 그리고 왼쪽 사선(left-up) 의 4가지의 방향성 코드를 정의하고 문자영역을 탐색한다. 예를 들어 전체 이미지 영역을 16개(4x4)의 국소 영역으로 분할하고 각 국소영역에서 각 방향성 코드의 값을 누적하여 합하면 전체 64개(4방향 x 16국소 영역 = 64피쳐)의 CDC 피쳐를 추출할 수 있다. CDC 피쳐벡터는 저해상도 영상자료에서와 같이 블러링되거나 변형된 문자들을 인식하는데 효과적으로 이용될 수 있다.

4. 시스템구성의 제안

본 논문에서 설명한 사전처리 방법과 문자인식을 위한 피쳐추출 방법을 저해상도 영상문서의 고성능인식에 적용하기 위하여 한글 자동 문서인식 시스템의 구성을 [그림 6]과 같이 3단계로 제안한다.



[그림 6]. 저해상도 문서 인식 시스템 구성도

시스템구성에 있어 다음과 같은 사항들을 추가적으로 고려하여야 한다. 첫째 사용되는 한글의 범위를 고려하여야 한다. 예를 들면 사용한글의 문자 개수는 상용 한글의 출현 빈도중 약 99%이내에 포함되는 약 1,400자 정도를 고려할 수 있고 그 외에 한글이 아닌 숫자, 영문 대소문자 그리고 특수기호 등을 포함하여 약 90자 정도를 고려함이 좋아 보인다.

둘째, 다중폰트 및 다중활자체의 패턴을 인식해 내기 위해서는 고성능 패턴 분류기가 이용되어야 한다. 이를 하나의 분류기로 직접 분류할 수는 있겠으나 팩스 패턴의 다양한 변화로 인해 분류기의 신뢰도를 최대한으로 확보하는 데는 어려움이 생길 수 있다. 따라서 문자인식의 전 단계에서 유형분류를 먼저 시도하고 분류된 유형에 따라 개별 문자분류기를 설계함으로써 인식성능을 극대화시킬 수 있을 것이다. 분할된 영상을 통하여 획득된 피쳐들을 이용하여 문자를 인식하기 위하여 요즘 각광을 받고 있는 신경망 이론이나 Pal등(1996)의 유전자(Genetic)알고리즘을 이용할 수 있다. 여러 개의 은닉층을 가지는 MLP(multi-layer perceptron)과 역전파(back propagation) 알고리즘을 이용하여 분류(classification)를 구현할 수 있을 것이다. p개의 입력 패턴과 n개의 출력 뉴런사이에서 목적값 t와 실제 출력값 o 사이의 오차 누적값은 에러를 최소화시키는 방향으로, 또 입력 패턴 p에 대한 뉴런 j의 출력값을 계산하기 위해서 시그모이드 함수를 사용할 수 있을 것이다.

셋째, 원하는 문자인식을 위한 비쥬얼(visual) 프로그래밍이 요구된다. 비쥬얼 프로그래밍은 많은 시간과 노력을 필요로 하나 디지털 이미지 프로세싱만을 위한 전문서적, 예를 들면 장동혁(2001), 이문호(2001) 그리고 하영호(1998)등을 이용하면 많은 도

움이 될 것이다.

5. 결론과 토의

종합적으로 저해상도의 문서 영상을 분석함에 있어서 사전처리 단계에서 도표 분리, 인식 및 한글과 영문의 효과적인 세그멘테이션 그리고 다양한 폰트의 인식 등에 대해 어려움을 가지고 있어서 고성능 상용 팩스 문서 인식 소프트웨어가 발표되지 않고 있는 실정이다. 그러나 최근의 국내 연구결과들을 보면 저해상도의 문서 인식을 위한 실험실 단위의 연구들이 진행 중이며 이를 활용한 저해상도 상용문서 인식 소프트웨어 구축에 관한 연구들이 계속 진행 중이다. 정보산업의 선진화를 위해서 한글 및 영문 혼용 팩스 문서 인식 소프트웨어의 개발에 대한 필요성은 매우 높은 실정이다.

본 논문에서 제안한 시스템 구성을 통하여 부분적이나마 한글 팩스문서의 영상인식이 고성능 피쳐추출법을 활용한 인식 시스템의 상용화가 이루어지기를 기대한다.

참고문헌

1. 이문호(2001). Visual C++ 실용영상처리, 대영사
2. 장동혁(2001). Visual C++을 이용한 디지털 영상처리의 구현, 정보게이트
3. 하영호, 임재권, 남재열 김용석(1998). 디지털 영상처리, 도서출판 그린
4. Daniel, S., George, R. and Harry, W.(1994) Automated page orientation and skew angle detection for binary document images, *Pattern Recognition*, vol. 27, no. 10, pp. 1325-1344.
5. Fukunaga(1990) *Introduction to Statistical Pattern Recognition*, Academic Press, INC.
6 Gose, E., Johnsonbaugh, R. and Jost, S.(1996) *Pattern Recognition and Image analysis*, Prentice Hall, New Jersey
7. Hori, O. and Doermann, D.(1995) Robust table-form structure analysis based on box-driven reasoning, *IEEE International conference on ICDAR*, vol. 1, pp. 218-221.
8. Kahan, S. and Pavlids, T.(1987) On the recognition of printed characters of any font and size, *IEEE Transactions. on PAMI*, vol. 9, no. 2, pp. 274-287.
9. Leondes, C.T.(1998) *Image Processing and pattern recognition*, Academic Press.
10. Pal, S.K. and Wang, P.P.(1996) *Genetic Algorithm for pattern recognition,* CRC press Inc.

[2003년 5월 접수, 2003년 7월 채택]