

기하학적 패턴 벡터를 이용한 한·영 글꼴 문자인식

석 영수, 홍 창희, 조 정락, 강 기섭, 민 종규, 이 응주
동명정보대학교 정보통신공학과

Hangul and English Text Font Recognition Using Geometrical Pattern Vector

Young-Soo Suk, Chang-Hee Hong, Jung-Rak Cho, Gee-Sub Kang,
Jong-Kyu Min and Eung-Joo Lee

Dept. of Information Communication Eng., TongMyong Univ. of
Information Technology
E-Mail : ejlee@tmic.tit.ac.kr

요약

본 논문에서는 문서 위의 문자를 Off-Line방식으로 컴퓨터에 저장할 수 있도록 기하학적 패턴 벡터를 이용하여 한·영문자 및 글꼴을 인식하는 알고리즘을 제안하였다.

일반적으로 문서에서는 여러 가지 글꼴에 따라 글자의 형태가 다르므로 대표적인 한·영 세 가지 글꼴을 기하학적 패턴(Geometrical Pattern Vector)을 이용하여 크기와 이동에 인식하도록 하였다.

이진 입력 한영혼용 영상에서 잡음을 제거하고 수평·수직 투영 기법을 이용하여 한 문자를 분할하여 문자의 폭에 따라 기하학적 패턴을 추출한다. 추출한 패턴은 각 합계를 계산하여 기준 패턴 합계와 비교한 후 기준 패턴 문자와 글꼴을 인식하게 된다.

마지막으로 제안한 알고리즘의 성능을 평가하기 위해 크기, 이동 변형이 있는 대표적인 한·영 글꼴(신명조, 궁서, 고딕)체와 영어 Time New Roman체를 대상으로 모의 실험을 수행하였다. 제안한 알고리즘은 기존의 원형 패턴 알고리즘보다 문자인식률과 글꼴 그리고 영어의 대·소문자를 구별하는 우수함을 보였다.

I. 서론

컴퓨터의 사용으로 문서 위의 정보를 컴퓨터에 저장하여 효율적으로 처리하고 관리하기 위해서는 문서 위의 문자를 인식하여 정보를 저장한다. 혼란하는 대부분 단일 글꼴 문자나 변형이 없는 문자에 대한 문자인식은 높

은 평가를 받고 있으나, 문서 위에 다른 글꼴의 문자를 사용하거나 크기를 변형했을 경우 현저히 인식하지 못하는 문제점을 가지고 있다. 그러므로 글자의 글꼴에 따라 인식 할 수 있는 알고리즘이 요구되고 있다.

문서에 포함된 많은 문자들을 자동으로 판독하여 컴퓨터에 입력시켜 주는 Off-Line방식과 On-Line방식이 있다. On-Line방식은 레이저 펜을 써서 인식하는 것과 Off-Line방식은 스캐너나 디지털 카메라로 찍은 사진을 이용하여 인식하는 방식으로 나눌 수 있다. 본 논문의 알고리즘은 Off-Line방식을 이용하였다.

문서에서 문자열과 각 개별 문자로 분할하는 방법에서는 투영 프로파일 및 계층적 군집화 기법을 이용한 방법이 있다. 그러나 계층적 군집화 기법을 이용한 방법은 문자의 어절 단위를 분할하는 알고리즘으로서 한 글자로 분할하지 못하는 문제점이 지적된다.[1]

문자인식에서는 기록 블록을 이용한 방법은 한 글자로 분할한 후 역전파 신경회로망을 이용하여 인식하기 때문에 학습과 한글 유형 분류기가 필요 하는 단점이 있다.[2]

최근에 원형 패턴 벡터를 이용한 문자인식은 단일 글꼴 문자를 회전과 크기, 이동에 무관한 원형 패턴을 이용하여 한글에서만 적용한 문자인식 알고리즘으로서 영어를 인식하지 못하는 단점을 가지고 있다.[3]

본 논문에서는 한·영혼용 문서에서 각 개별 문자로 분할하고 이를 한글과 영어의 대·소에 대한 문자와 글꼴 인식 알고리즘을 제안하였다.

II. 한·영 텍스트 영역의 특징

1. 텍스트 영역에 대한 특징

일반적으로 한영혼용인 문서는 한글 문자열로 문단을 세그멘테이션(Segmentation)하고 문단에 각 글자로 세그멘테이션해야 한다. 처음으로 입력된 영상에 수평 방향 투영 기법으로 문자열을 추출 후, 수직 방향 투영 정보를 이용하여 개별 문자 영역을 세그멘테이션한다.

대한전자공학회 문자인식

(a) (b)

그림1. 한글 문자열에 대한 수평 정보; (a) 입력된 한글 영상, (b) 각 라인에 수평 투영 결과 영상.

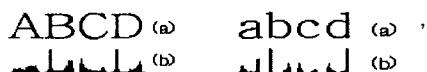
그림1의 (b)영상은 입력된 영상(a)의 문자열들을 수평 투영 기법을 이용한 결과 영상으로서 글자의 검정 화소가 있는 부분은 나타나고, 검은 화소가 없는 부분은 나타나지 않고 골(Valley)로 이루고 있다. 이 골을 이용하여 문자열로 분할한다.

대한전자공학회 (a)



(b)

그림2. 한글 문자열에 대한 수직 정보; (a) 입력된 한글 영상, (b) 각 글자에 수직 투영 결과 영상.



(a) (b)

그림3. 영어 문자열에 대한 수직 정보; (a) 입력된 영어 영상, (b) 각 글자에 수직 투영 결과 영상.

그림2와 3의 (b)결과 영상은 분할된 문자열(a)에 수직 투영 기법을 이용한 것으로서 각 글자마다 골을 이루고 있다. 수직 정보를 이용하여 각 개별 문자로 분할한다.

2. 한글과 영어의 글꼴에 대한 특징

인쇄된 문서를 스캐너로 입력받은 영상에서는 여러 문자의 글꼴이 있지만 그 중 대표적인 글꼴(신명조, 고딕, 궁서)체를 구별 할 수 있다.



그림4. 한글 '원'자의 각 글꼴(신명조, 고딕, 궁서)들.

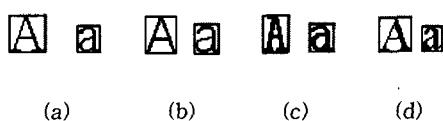


그림5. 영어 대·소문자 영상들; (a) 신명조체, (b) 고딕체, (c) 궁서체, (d) Time New Roman체.

그림4는 한글 '원'자의 글꼴에 대한 영상으로서 각 글꼴에 따라 모양이 다르고, 사각형 영역의 크기가 다른 특징을 가지고 있으며, 그림5는 영어의 대·소문자도 글꼴에 따라 다른 모양과 사각형 크기를 가지고 있다.

한글과 영어의 대표적인 글꼴의 특징 정보를 이용하여 각 글꼴을 구별한다.

III. 텍스트 영역에서의 문자 분할과 한·영 문자 및 글꼴 인식 알고리즘

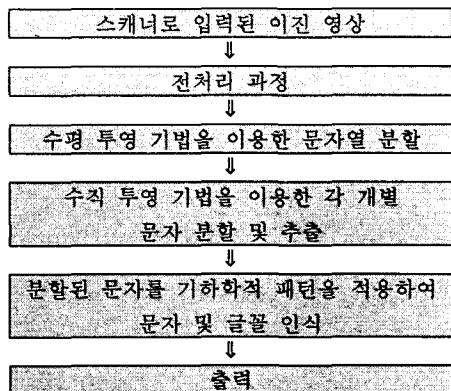


그림6. 제안한 문자 및 글꼴 인식 알고리즘 구성도. 본 논문에서 제안하는 문자 및 글꼴 인식 알고리즘은 입력영상에 이진 형태론(Binary Morphology)을 처리한 후 입력 영상에 잡음을 제거하고, 글자의 문단이 기울어져 있으면 기울기 보정하는 전처리 과정을 거친 후 수직·수평 히스토그램 분포를 이용하여 한 글자를 분할한 후 기하학적 패턴을 추출한다. 추출된 기하학적 패턴은 수직 라인 패턴과 수평 라인 패턴 그리고 주어진 글자의 중심으로 하는 세 개의 원주 상에 위치한 분포 값을 나타내는 것으로 추출한 실험 문자에 각 패턴의 합계와 기준 패턴의 합계를 비교하여 일치시 기준 패턴의 문자와 글꼴을 인식하는 알고리즘으로 그림6과 같다.

1. 전처리 과정

인쇄된 문서를 스캐너로 입력받은 영상을 흰 배경색과 검은 글자색으로 나누는 2진 형태론적 영상 처리하였다. 전처리 과정에서는 입력된 2진 영상에서 잡음을 제거하고 팽창 연산(Dilation Operator)을 수행하여 원 영상의 크기를 유지하는 연산자이다.

문서 영상이 기울어져 있을 경우에는 문서의 기울어진 각도를 계산해 내고 영상 회전 변환을 Sin함수와 Cos함수를 이용하여 기울기 보정하였다.[4]

원형패턴 문자인식 표한글 문자 폰 트 태복	원형패턴 문자인식 표한글 문자 폰 트 태복
-------------------------------	-------------------------------

그림7. 기울어진 영상을 기울기 보정한 결과; (a) 기

울어진 영상, (b) 기울기 보정한 결과 영상.

2. 각 문자 분할과 인식

전처리 과정이 끝난 영상에서 텍스트 영역의 특징 정보를 이용하여 수평 투영 프로파일을 구하고, 수평 프로파일에는 글자 영역의 라인과 골(Valley)로 분리되어 배경과 텍스트 라인으로 분할하였다. 분할한 텍스트 라인에서 최상단 좌표와 최하단 좌표를 구한다.

다음으로 최상단 좌표와 최하단 좌표 사이에 수직 투영 프로파일을 구하고 텍스트 영역의 수직 투영 결과 정보를 이용하여 글자와 골로 분리하여 글자의 좌측 좌표와 우측 좌표를 구한다.

다시 좌측 좌표에서 우측 좌표까지 최상단에서 아래로, 최하단에서 위로 글자의 프로파일을 검출하고, 검출된 글자 영역의 상단 좌표와 하단 좌표를 구한다.

최종적으로 구한 각 좌표를 원 영상에 적용하여 영상의 좌측부터 각 개별 글자로 추출하였다.

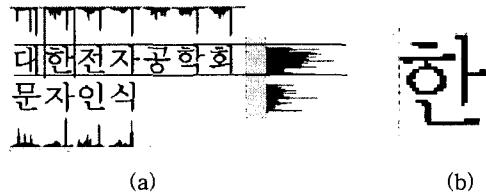


그림8. 제안한 분할 알고리즘 결과 영상; (a)문자열에 대한 수평수직 투영 결과, (b)추출한 개별 글자 영상. 분할된 글자의 글꼴과 인식에 무관한 특성을 추출하기 위해 기하학적 패턴 벡터 생성 알고리즘을 사용하였다. 기하학적 패턴 벡터는 그림9와 같은 수평과 수직 그리고 원형을 투영하여 문자 검은 화소의 분포를 추출하는 알고리즘이다. 수평·수직 투영 기법을 이용하여 한 개별 문자로 분할된 그림8의 (b)영상에 기하학적 패턴을 적용한다.

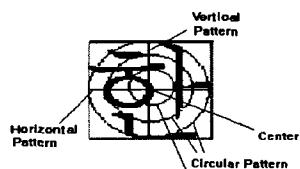


그림9. 기하학적 패턴 벡터의 생성 예.

우측 좌표값 (Y_R)에서 좌측 좌표값 (Y_L)을 빼면 글자의 폭을 구할 수 있고, 하단 좌표값 (X_B)에서 상단 좌표값 (X_T)을 빼면 글자의 높이를 구할 수 있다. 구한 글자의 폭과 높이를 반으로 나누면 식(3)과 같이 글자의 중심(Center)을 구할 수 있다. 구한 중심을 수직과 수평 투영 라인 패턴과 원형의 중심으로 이용한다.

$$CH = X_B - X_T, CW = Y_R - Y_L \quad (2)$$

$$C(X_C, Y_C) = (CH/2) + X_T, (CW/2) + Y_L \quad (3)$$

식(2)의 CH는 글자의 높이와 CW는 글자의 폭을 나타

내는 값이다. 식(4)와 (5)는 글자 영역에 수직과 수평 투영하여 패턴을 추출하는 식을 나타내고 있다.

$$VerP[CH] = Img[X_T \rightarrow X_B][Y_c] \quad (4)$$

$$HoriP[CW] = Img[X_C][Y_L \rightarrow Y_R] \quad (5)$$

원형은 CW에 반으로 나눈 후 3등분 한 값을 원의 반지름 R1, R2, R3로 정한다. 그리고 글자 영역에 3° 간격 씩 시계 방향으로 각 반지름의 원을 돌면서 검은 화소는 1로, 바탕 화소는 0으로 하여 120개 원소 분포를 추출한다.

$$Cir(X_{Ri}, Y_{Ri}) = (Ri \times \sin(\theta), Ri \times \cos(\theta)) \quad (6)$$

$$S_{R_{1 \rightarrow 3}} = \sum_{i=1}^{120} Cir[R_{1 \rightarrow 3}][i] \quad (7)$$

$$S_{V,H} = \sum_{i=0}^{CV} VerP[i], \sum_{j=0}^{CH} HoriP[j] \quad (8)$$

식(6)은 원형의 3° 씩 원 영상에서의 좌표를 구하는 식을 나타내고 원형의 각 반지름에 대한 원소 합계 S_R 과 수직·수평 원소의 합계 $S_{V,H}$ 를 구하는 식(7)과 (8)을 나타내고, 각 합계를 기준의 각 패턴에 대한 합계가 일치하면 기준 패턴에 해당하는 문자를 인식하였다.

원 세 개의 반지름은 글자의 폭에 따라 동적으로 변하기 때문에 글자의 대표적인 글꼴을 분별할 수 있다. 또한 같은 글자지만 글꼴에 따라 폭이 다르기 때문에 원형의 패턴과 수직·수평의 패턴이 글꼴에 따라 달음을 알 수 있다. 그리고 영어의 대·소문자도 구별 할 수 있고 글자의 크기에 무관한 특성을 가지고 있다. 이 패턴을 이용하여 글꼴과 문자를 인식하였다.

다음은 기하학적 패턴을 적용하여 추출한 각 한글 글꼴에 대한 패턴을 나타낸 그림이다.

	신명조	궁서	고딕
원			
한			
글			
인			
식			
형			

그림10. 한글 문자를 글꼴에 따라 기하학적 패턴(수직, 수평, 원형)을 적용한 결과 패턴.

그림10은 한글의 글꼴에 따라 추출한 기하학적 패턴이 달음을 알 수 있다.

그림11은 영어 알파벳을 기하학적 패턴을 적용한 패턴 그림으로써 대문자와 소문자로 각 글꼴에 따라 추출한 기하학적 패턴이 대·소문자와 각 글꼴의 패턴이 달음을 알 수 있었다.

	신명조	고딕	궁서	Time New Roman
A				
B				
C				
a				
b				
c				

그림11. 대·소문자 영어를 글꼴에 따라 기하학적 패턴을 적용한 결과 패턴.

제안한 알고리즘으로 한 글자로 분할하고 인식한 글자를 데이터(Data)로 만들어서 저장한 후 다시 다음 문자를 분할하고 인식한 후 저장하고 문서가 끝날 때까지 알고리즘을 반복한다. 저장된 Data를 출력하는 문자 및 글꼴 인식 알고리즘을 보였다.

IV. 실험 결과 및 성능 평가

본 논문에서는 기존의 원형 패턴 벡터를 이용한 문자인식 알고리즈다 한글뿐만 아니라 영어 대·소문자 그리고 글꼴을 인식하는 알고리즘을 제안하였다. 제안한 알고리즘의 성능을 평가하기 위해 신명조체의 한영혼용 문서 그림12의 (a)와 한글 글꼴(신명조, 고딕, 궁서) 크기와 위치 이동 변화가 있는 문서(b)와 영어 알파벳의 고딕, 궁서체, Time New Roman체 글꼴 문서 그림13의 영상들을 실험하였다.

원형패턴 문자인식 표한글 문자 폰트 태복 ABCDEFGJHIKL abcdefghijklmnop qrestuvwxyz	문자 패턴 인식 (b)
--	-----------------

그림12. 문자 입력 영상;(a) 신명조의 영한혼용 문서 영상, (b) 한글의 대표적인 글꼴(신명조, 궁서, 고딕) 문서 영상.

abcdefghijklmnop pqrstuvwxyz ABCDEFGHIJKL MNOPQRSTUV WXYZ	ABCDEFGHIJKLMNOP PQRSSTUVWXYZ abcdefghijklmnop pqrsuvwxyz	ABCDEFGHIJKL NOPQRSTUVWXYZ abcdefghijklmnop tuvwxyz
---	--	--

그림 13. 영어 알파벳의 글꼴 입력 영상;(a) 고딕의 영어 대소 입력 영상,(b) 궁서체의 영어 대소 입력 영상,(c) Time New Roman체의 영어 대소 입력 영상.
위의 영상들을 제안한 알고리즘으로 수평·수직 투영 기법을 이용하여 각 문자열과 문자를 분할한 후 기하학적 패턴 벡터를 이용하여 문자 인식뿐만 아니라 글꼴체도 인식하였고, 영어의 대·소문자도 구별할 수 있었다.
그림14는 위 영상들을 입력하여 제안한 알고리즘으로

인식한 결과그림이다.

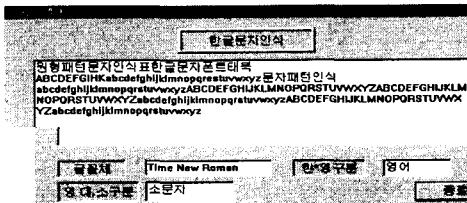


그림14. 제안한 알고리즘으로 문자인식 결과.

총 20장 영상에서 한영혼용 문서 10장과 한글, 영어만 있는 문서 각각 5장을 글꼴에 따라 약 500자에 대한 실험 결과에서 97.2%의 높은 문자인식률을 나타내었고, 영어의 대소문자 구별은 정확하게 판별하고, 글꼴도 정확하게 인식하였다.

V. 결 론

본 논문에서는 문서의 텍스트 영역에서 수평 투영 정보를 이용하여 문자열을 분할하였고, 수직 투영 정보를 이용하여 한 문자로 분할한 후 추출하였다. 또한 추출된 문자 영상이 같은 글자지만 글꼴에 따라 글자의 폭이 동적으로 변하는 특성과 폭에 따라 기하학적 패턴이 다른 특성들을 가지는 정보를 이용하여 문자와 글꼴을 인식하도록 하였다. Off-Line 방식의 HP 스캐너에 인쇄된 문서를 입력 영상으로 받은 다음 문자 인식을 시도한 결과 한 문서당 98.6%정도의 인식률을 얻을 수 있었으며 글자의 크기와 패턴에 따라 글꼴과 문자를 정확하게 인식 할 수 있다는 것을 확인하였다.

인쇄된 한글과 영어에서는 많은 글꼴이 있고, 또 새로운 글꼴이 나오고 있기 때문에 대표적인 글꼴 외에도 인식할 수 있는 연구가 요구되고, 특수문자에 대한 인식 연구가 요구된다.

참 고 문 헌

- [1] 정창부,곽희규,김수형,“투영 프로파일 및 계층적 군집화 기법을 이용한 텍스트영역의 어절단위 분할”,정보통신논문지,1999,12,Vol.3,No.1,
- [2] 김규경,김진호,찬성일,최홍문,“불은 글자들이 포함된 인쇄체 한·영혼용 문서에서의 효과적인 문자 인식 알고리즘”,전자공학회논문지,1996,11,v.33-B,n.11, pp.116-126
- [3] 정지호,최태영,“원형패턴벡터를 이용한 인쇄체한글인식”,전자공학회논문지,1999,제6권,제1호,pp.269-281[3]
- [4] 김두식,이성환,“한·영혼용 문서의 디지털 라이브러리 구축을 위한 효과적인 문서 기울기교정 및 문자분할방법”,한국정보과학회 봄 학술발표 논문집, Vol.26, No.2, pp. 482-484,1999
- [5] S.Shlien,“Multifont character recognition for typest documents,”IEEE Trans.Pattern Analysis Mach. Intell.vol.2,no.4,pp.603-620,1988.