다양한 크기 및 활자체를 갖는 인쇄체 한글 영상의 문서화에 관한 연구

A Study on Documentization of Printed Hangul Image with Multi-size and Multi-style

김장욱, 김경숙, 손영선
동명정보대학교 정보통신공학과

Jang-Wook Kim, Kyung-Suk Kim, Young-Sun Sohn
Department of Information & Communication Engineering,
Tongmyong University of Information Technology
(yssohn@tmic.tit.ac.kr)

요 약 문

본 논문에서는 CCD카메라로 입력 받은 다중 크기 및 활자체로 구성된 한글문서의 화상 데이터를 편집기에서 수정 가능한 문자로 변환시키는 시스템을 구현하였다. 먼저 Dynamic 이진화 처리 과정을 거친 화상을 흑백 화소의 누적분포에 따라 문자단위로 분할한 후, 다양한 크기로 분할된 문자를 표준패턴 크기로 표준화 시켰다. 한글을 자소 간 공백 위치의 특징에 따라서 6가지 유형으로 분류한 후, 퍼지 이론을 접목시킨 원형 패턴 벡터 알고리즘을 사용해서 표준벡터와 입력된 글자의 특징벡터를 비교하여 문자로 인식하게 하였다. 각 6가지 유형에서 서로 다른 자소로 결합된 문자들을 30개 선정하여 여러 가지 활자체 및 크기에 적용해 본 결과, 모두 문서화가 가능함을 알 수 있었다.

Key Words : 문자인식, 퍼지이론, 원형 패턴벡터, 문서화

I. 서 론

오늘날 고도화된 정보사회에서 많은 양의 문서정보를 신속하게 처리하기 위하여 정보의 입력수단을 키보드가 아닌, 스캐너나 카메라 등의 영상 입력 장치를 사용하여 문서를 자동으로 입력시켜 주는 시스템 개발이 요구되어지고 있고, 이미 상용화된 실정이다[1].

컴퓨터 과학의 한 분야인 인공지능분야에서는 정보를 일일이 입력해야 하는 불편함을 컴퓨터에 시각 기능을 부여하여 해결하려는 연구가 행하여 지고 있다[2].

대량의 문서가 자동으로 입력되어 편집 가능한 상태로 저장된다면 많은 시간의 절약과 영상을 문서화 함으로써 저장 용량이 축소되

는 장점이 있으며, 멀티미디어 접속장치를 활용하여 장애인들도 능동적으로 문화생활에 접할 수 있는 기회가 확대되리라 기대되어진다.

본 논문에서는 CCD카메라로 입력 받은 다양한 크기 및 활자체로 구성된 한글 영상을 퍼지이론을 적용한 원형 패턴 벡터 방식으로 인식하여 수정 가능한 문서로 변환해 주는 시스템을 구현하였다.

Ⅱ. 전체 시스템 알고리즘

본 논문에서 구현한 시스템은 그림 1과 같이 CCD카메라로 영상을 입력받아 명도에 따라 동적으로 임계값을 할당하는 이진화 처리 과정에서 구하여진 영상을 수평·수직방향으로 투영(projection)하여 문자 단위로 분할한다. 다양한 크기를 갖는 활자 및 영상으로 인해 각각 다르게 분할된 문자는 확대 또는 축소시켜 표준화 시킨 후, 한글의 6가지 유형으로 분류하였다. 분류된 문자는 원형패턴벡터 방식을 사용하여 인식되어지는데 이 방식은 한글의 다양한 서체를 인식하기에는 문제점이 있기 때문에, 퍼지이론을 접목시켜 글자의 특징 벡터와 기준점에서 벡터 요소들 간의 거리차를 이용하여 다양한 활자체의 글자라도 같은 글자로 인식하게 하였다. 또한 인식된 문자들을 에디터에 입력되게 함으로써, 입력받은 영상 파일을 수정 가능한 문서 파일로 변환시켰다.

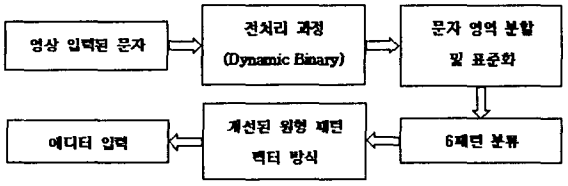


그림 1. 전체 시스템 구성도

Ⅲ. 문자 분할 및 표준화

1. 문자 분할

동적 이진화 처리 과정을 거친 영상에서 문자영역을 분할하기 위하여 영상을 수평방향으로 투영하면서 행 단위로 분할하고, 한 행에서 다시 수직방향으로 투영하여 문자단위로 분할하는 알고리즘을 사용하였다. 그림 2에 보여지듯이, 수평방향의 투영에서는 검은 화소의 누적분포에 따라 문자열을 분할하였고, 수직방향의 분할에서는 자음과 모음간의 공백으로 인해 문자 분리 현상이 일어났지만 문자의 평균 글자 폭을 적용시켜 해결하였다[3].

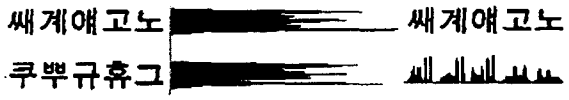


그림 2. 수평·수직 방향 히스토그램 결과

2. 크기 표준화

분할된 문자영역은 다양한 크기의 한글을 입력 받기 때문에 평균 글자 폭을 적용하더라도 다른 크기로 분할된다. 따라서, 동일한 조건에서 문자를 인식하기 위해서는 크기의 표준화 과정이 필요하다. 본 시스템에서는 각각의 분할된 문자를 48×48픽셀 크기로 표준화 시키기 위하여 좌표를 사상(Mapping)시켰다. 즉, 분할된 문자의 좌표 값이 x, y이고, 표준 크기의 좌표 값이 x', y'이라고 가정하면, 확대 비율은 식(1)과 같이 계산되어진다.

$$(x', y') = ((x \times T_x), (y \times T_y)) \quad (1)$$

여기서, T_x, T_y는 확대인수를 나타내는데, 이것은 분할된 문자의 크기와 표준 크기와의 비율로 계산되어진다. 즉, T_x는 분할된 문자 크기와 표준크기와의 가로 길이의 비율, T_y는 세로 길이의 비율을 나타낸다.

확대되어진 영상은 할당받지 못한 빈 픽셀(hole)들이 생기게 되므로, 이는 빈 픽셀에 이웃하는 픽셀들의 평균값을 할당하는 보간법(Bilinear interpolation)을 사용하여[4] 크기를 표준화 시킬 수 있었다.

Ⅳ. 문자 인식

1. 6형태 분류

한글은 그림 3에서와 같이 두 개 또는 그 이상의 자음과 모음들이 수직 또는 수평방향으로 결합된 형태에 따라 6가지 유형으로 분류된다[5].

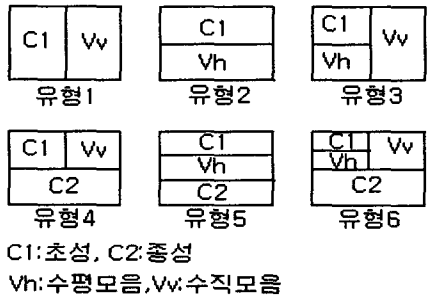


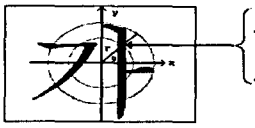
그림 3. 한글의 6가지 유형

입력된 한글이 그 구조상의 특성에 기초하여 6형태 중 하나로 분류되어지면 인식을 위한 탐색 공간이 매우 축소되므로, 문자인식 과정 전에 표준화된 글자들을 각 유형별 자소간의 공백 특성을 조건으로 6가지 형태로 분류하였다.

2. 퍼지를 이용한 원형패턴벡터알고리즘

2-1. 원형패턴벡터를 이용한 특성벡터 추출

원형 패턴 벡터 방식은 그림 4와 같이 입력문자에 대하여 무게 중심을 구하고 그 무게 중심을 원의 중심으로 정하여 반지름 r이 평균 거리인 원을 반시계 방향으로 일정한 각도씩 회전시키면서 문자와 교차하는 부분의 특징점들을 추출하는 방식이다[6].



$$\begin{cases} x = r \cdot \cos\theta \\ y = r \cdot \sin\theta \end{cases} \quad (2)$$

그림 4. 특징점의 위치

이 방식을 사용하여 표준화된 문자의 특성 벡터를 추출하며, 서로 다른 패턴들 사이의 구별력을 크게 하기 위하여 영상에 반지름이 r1, r2, r3인 3개의 원을 3도 간격으로 각도가 2π가 될 때까지 회전시킨다. 한 원당 120개의 데이터를 얻으므로, 3개의 원에서는 총 360개의 데이터를 사용한다. 만약 영상에 어떤 포인트가 i번째 원에 속하는 검정 화소 들의 수라면 RGB0의 값으로, 흰색 화소들의 수라면 RGB255의 값으로 나타내었다. 무게 중심 $(\overline{X},\overline{Y})$이 극 좌표계의 원점이라 할 때 (즉, $x_0 = \overline{X}, y_0 = \overline{Y}$), 패턴 영상은 식(3)과 같이 표현될 수 있다[7].

$$f'(r,\theta) \begin{cases} 255 & \text{(배경의 흰색화소)} \\ 0 & \text{(문자 부분의 검정화소)} \end{cases} \quad (3)$$

$$\text{단, } 0 \le r \le n,\ 0 \le \theta \le 2\pi,\ n = \sqrt{x_1^2 + y_1^2},$$
$$x_1 = Max(\overline{X}, M - \overline{X}),\ y_1 = Max(\overline{Y}, M - \overline{Y})$$

2-2. 퍼지 멤버쉽 함수를 이용한 문자인식

그림 5, 6을 보면 다중 활자체의 문자들이 가로, 세로 48×48픽셀 크기로 표준화 되어도 중심점에서 조금 이동된 문자들, 전처리 과정에서 손실된 픽셀 등에 의하여 기준점에서 특징 벡터들 간의 거리차가 조금씩 달라지는 것을 알 수 있다.

그림 5. 바탕체 문자의 특징점 추출

그림 6. 돋움체 문자의 특징점 추출

이 점을 고려하여 문자인식을 위한 멤버쉽 함수는 그림 7과 같이 특징 벡터가 나타나는 각 지점에 소속정도의 유연성을 크게 해주는 사다리꼴 모양으로 적용하였고, 활자체가 다른 동일한 글자의 경우에 특징 벡터가 중심점에서 조금 이동하거나 기준점과 특징 벡터간

의 거리차가 적고 같은 멤버쉽 구간 내 약간의 픽셀들이라도 존재한다면 같은 글자의 특징 벡터라고 인정하여 주는 방식으로 적용하였다. 만약 표준 벡터와 유사한 벡터들이 복수개 존재할 경우에는 소속 정도가 높은 것을 선택하여 다른 문자의 특징 벡터들과 구분하였다[8].

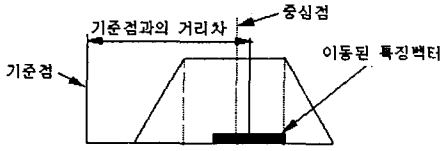


그림 7. 특징 벡터에 대한 멤버쉽 함수

V. 시스템 실행

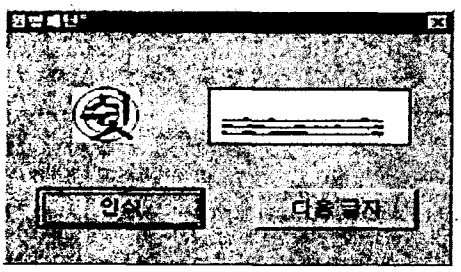


그림 8. 원형 패턴 벡터 결과

입력한 영상파일을 이진화 하면, 다양한 크기의 문자들을 표준화시킨 다음 한 문자씩 특징 벡터를 추출한다.

그림 8과 같이 추출되어진 특징 벡터를 해당 문자가 분류되어 있는 6패턴 내 문자들의 특징 벡터와 비교하면서 멤버쉽 값의 소속 정도가 가장 높은 것으로 인식하여 그림 9에서 보여지는 간단한 에디터 창에 입력되도록 구현하였다.

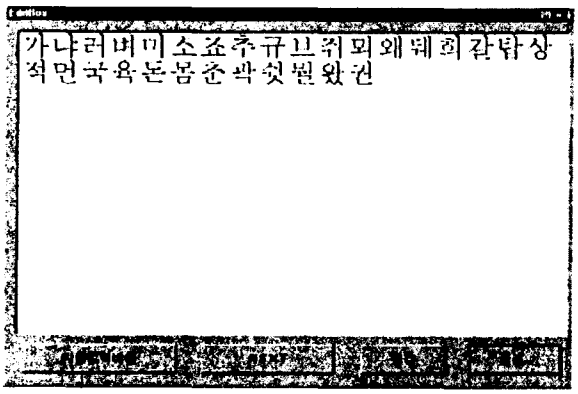


그림 9. 인식 결과 (에디터에 입력된 문자들)

VI. 결론 및 향후과제

본 논문에서는 CCD카메라를 이용하여 문자 영상을 입력 받아 수정 가능한 문서 형태로 변환해주는 시스템을 구현하였다. 원형 패턴 벡터 알고리즘과 퍼지이론을 접목시킨 개선된 문자인식 알고리즘으로 다양한 크기 및 활자체를 갖는 한글 인식이 가능하였다.

향후과제로는 현재 통용되는 모든 한글과 영문, 숫자에도 제안된 방법을 적용시켜 문서화하는 것이라 고려되어진다.

VII. 참고문헌

[1] 이인동, 권오석, 김태균, "문서 영상에서 문자와 비문자의 분리추출방법", 한국정보과학회논문지, 제17권3호, 1990년 5월.

[2] 김홍배, 송종수 외 3명, "멀티폰트 한글 문서인식 시스템의 개발", 한국인공지능개발연구조합, 1992년 11월.

[3] 최봉희, 이인동, 김태균, "문자영역 추출 과정에서의 오분리의 교정", 한국정보과학회 논문지, 제21권1호, 1994년 1월.

[4] 장동혁, "디지털 영상 처리의 구현", PC어드밴스

[5] 김계경, 김진호, 천성일, 최홍문, "붙은 글자들이 포함된 인쇄체 한·영 혼용 문서에서의 효과적인 문자인식 알고리즘", 전자공학회논문지, 제33권11호, 1996년 11월.

[6] 정지호, 김희태, 최태영, "원형 패턴 벡터를 이용한 인쇄체 한글인식", 제11회 신호처리 합동 학술대회, 제11권1호, pp.113-116, 1998년

[7] 이성환, 박희선, "고리 투영을 이용한 위치, 크기 및 회전 변형에 무관한 한글 문자 인식", 인지과학회 논문지, 제3권1호, pp.139-160, 1991년 6월.

[8] 本多中二, 大理有生, "퍼지공학 입문", 응보출판사, 1999.