한국지능정보시스템학회
02 춘계학술대회 논문집, pp. 410~413

스미어링 기법과 윤곽선 추적 알고리즘을 이용한 영문 명함 영상에서의 문자 추출

조아현* · 이혜현* · 류재욱* · 김광백**

The Extraction of Character from an English Name Card by Using Smearing Method and Contour Tracking Algorithm

Ah-hyun Cho* · Hye-hyun Lee* · Jae-uk Ryu* and Kwang-Baek Kim**

요 약

본 논문에서는 영문 명함 영상에서 개별 문자 추출 방법을 제안한다. 30개의 원본 명함 영상을 대상으로 스미어링 기법과 윤곽선 추적 알고리즘을 이용하여 영문 명함의 개별 문자들 추출하였다. 본 논문에서는 3×3 마스크를 이용하여 가장 작은 값으로 3 배 축소하는 방법을 적용하여 스미어링하는 시간을 단축시키고 문자들간의 간격을 제거하여 윤곽선 추적 알고리즘을 이용하여 문자열 후보 영역을 추출하였다. 그리고 추출된 후보 영역의 가로 및 세로의 비율과 면적을 이용하여 문자열과 비문자열로 분리하고, 문자열 영역에서 4 방향 윤곽선 추적 알고리즘을 이용하여 개별문자를 추출하였다. 30개의 명함 영상을 실험한 결과, 309개의 문자열 중에서 280개가 추출되었고 개별 문자는 4504개중에서 4110개가 추출되었다.

Key words : 스미어링, 윤곽선 추적 알고리즘, 문자열 영역, 개별 문자

1. 서론

최근에는 자신을 소개하기 위해 명함을 많이 보유하고 있고, 이런 명함을 효율적으로 관리하기 위해서 명함 꽂이와 명함 집 등이 많이 판매되고 있는 추세이다. 그러나 많은 명함을 보유하고 있을 경우에는 명함 꽂이나 명함 집은 복잡성을 가중시키므로 많은 문제를 가진다. 사람이 명함을 수정할 경우에는 명함 꽂이나 명함 집에서 두 개의 동일한 명함을 보유하게 되고 이후에 어느 것이 현재의 것인지 인지할 수 없다. 이러한 문제 등을 해결하기 위해서는 명함 인식 기술이 필요하다. 명함 인식을 위해서는 문자영역을 빠른 시간에 정확하게 추출하는 기술이 요구된다. 명함 영상내의 문자들은 가로 방향으로 규칙적인 간격으로 구성되어 있다. 본 논문에서는 같은 줄에서는 문자의 크기가 일정하다는 정보를 이용하여 원 영상을 3배 축소하고 스미어링 기법[1,2]을 이용하여 문자 사이의 빈 여백을 제거하고 문자들을 뭉쳐서 윤곽선 추적 알고리즘[3,4]으로 문자열의 후보 영역을 추출한다. 추출된 문자열 후보 영역에서 가로 및 세로의 비율과 면적을 이용하여 문자열과 비문자열 영역을 분리하고 문자열 영역에 대해서 4 방향 윤곽선 추적 알고리즘을 이용하여 개별문자를 추출한다.

2. 영문 명함 문자 영역 추출

본 논문에서는 입력된 명함 영상을 3 배 축소하여 가로 스미어링과 4 방향 윤곽선 추적 알고리즘을 이용하여 문자열 후보 영역을 추출하여 문자열과 비문자열 영역을 분리하고, 추출된 문자열 영역에서 4 방향 윤곽선 알고리즘을 적용하여 개별문자를 추출한다.

2.1 스미어링 기법과 윤곽선 추적을 이용한 문자열 추출

원본 명함 영상은 정규화 되지 않은 크기와 간격의 문자들로 구성되어 있다. 그러나 한 문자열 블록 내에서는 문자의 크기와 간격의 변화가 거의 없다. 따라서 정확한 문자를 추출하기 위해서는 문자열의 추출이 중요하다. 본 논문에서는 3×3 마스크를 이용하여 가장 작은 값으로 3배 축소하는 방법을 적용하고 스미어링 기법을 이용하여 문자들간

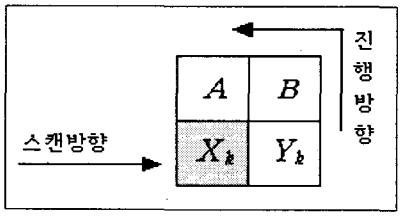
* 신라대학교 컴퓨터정보공학부
** 신라대학교 컴퓨터공학과

의 간격을 제거하고 문자들을 뭉쳐서 문자열을 분리하도록 한다. 그러나 스미어링 기법은 많은 시간이 소모되기 때문에 큰 영상에 적용하기에는 어려움이 있다[5]. 이를 해결하기 위해 원본 영상을 3배 축소하여 스미어링 기법을 적용한다.

3배 축소된 영상에서 문자간격을 제거하기 위해 가로 스미어링을 수행한다. 스미어링은 흑화소(black)를 수평으로 일정크기의 픽셀만큼 늘려주는 방법으로 스미어링의 크기는 동적으로 수행하여 검지한다. 가로 스미어링은 문자들의 특징을 묶어주는데 효율적이다.

가로 스미어링된 영상에서 문자열 후보 영역을 추출하기 위해서 윤곽선 추적 알고리즘을 이용한다. 4 방향 윤곽선 추적은 그림 1의 2×2 마스크를 이용한다. 음영이 들어간 경계부분을 기준으로 반시계 방향으로 진행되어 이미지에서 경계를 만나기 전까지는 왼쪽에서 오른쪽으로 위에서 아래로 스캔한다. 4 방향 윤곽선 추적 알고리즘은 X_k를 시작점으로 A와 B에 대응하는 두 픽셀을 고려하여 마스크 진행방향을 결정한다. X_k가 지나간 자리가 영상의 윤곽선이 된다. A와 B가 모두 배경일 경우 마스크는 X_k를 기준으로 진행방향으로 회전하고 A가 경계일 경우 기준점 X_k는 A로 이동하면서 마스크는 한 픽셀 앞으로 전진한다. 또 B가 경계일 경우 마스크는 Y_k를 기준으로 시계방향으로 이동한다. 즉 X_k는 B경계점으로 이동한다. A와 B가 모두 경계일 경우는 X_k는 가까운 A로 이동하고 B 또한 이동해야할 경계이므로 Y_k는 B를 피해 X_k로 이동한다. 표1은 A와 B의 값에 따른 X_k와 Y_k의 진행방향인 X_{k+1}와 Y_{k+1}를 보여준다. 여기서 A와 B는 0과 1로 표시되며 0은 배경 픽셀을 의미하고, 1은 경계 픽셀을 의미한다.

윤곽선 추적에 의해 추출된 문자열 후보 영역에서 문자열을 추출하기 위해 추출된 후보 영역의 가로 및 세로의 비율과 면적을 이용하여 문자열과 비문자열로 분리한다. 본 논문에서는 30개의 명함 영상을 실험하여 추출된 문자열 후보 영역의 가로와 세로의 비율이 수치상으로 1.5 이상이고 면적이 8000보다 적으면 문자열로 추출하였다.



[그림1] 윤곽선 추출을 위한 2×2 마스크

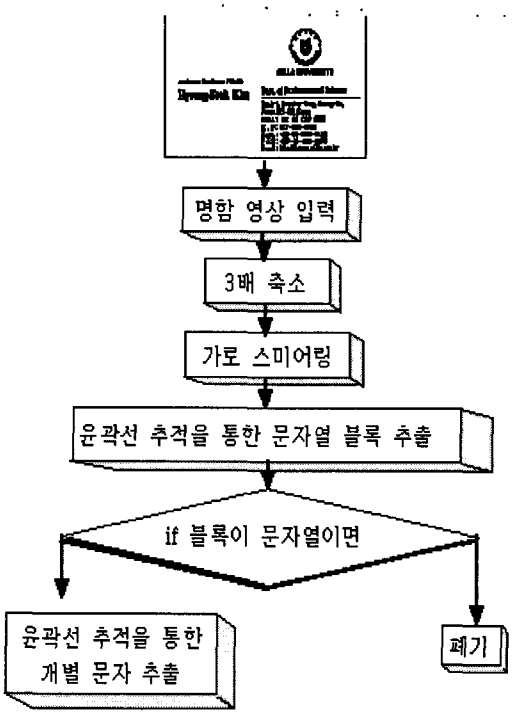
2.2 윤곽선 추적을 이용한 개별 문자 추출

윤곽선 추적 알고리즘은 영역의 윤곽선 지점을 추출하여 분할되는 영역을 하나의 특징 영역으로 구분하여 문자를 추출하는 방법이다. 개별 문자 추출에서도 윤곽선 추적은 반시계 방향으로 그림1과 같이 2×2 마스크를 이용하여 개별 문자를 추적한다. 윤곽선 추적은 문자열 영역의 경계 픽셀 중의 하나를 시작점에 위치시키고 그림 1의 A와 B에 대응되는 두 픽셀을 고려하여 마스크의 다음 진행 방향을 결정한다. 개별 문자를 추출하기 위한 2×2 마스크의 A와 B에 따른 진행 방향은 표 1과 같다.

<표 1> 2×2 마스크의 A, B에 따른 진행방향

	a	b	X_{k+1}	Y_{k+1}
좌측	0	0	A	B
우측	0	1	B	Y_k
전진	1	0	A	X_k
우측	1	1	X_k	A

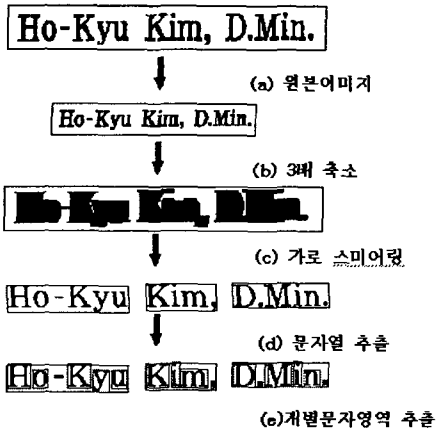
표 1의 A와 B에 따른 진행 방향을 이용하여 윤곽선 추적으로 개별 문자를 추출한다. 본 논문에서의 영문 명함 영상 추출 구성도는 그림2와 같다.



[그림 2] 명함 문자 영역 추출 구성도

3. 실험 및 결과

실험 환경은 IBM 호환 기종 펜티엄III 환경에서 C++ Builder 5.0으로 구현하였다. 30개의 영문 명함 영상 대상으로 실험하였다. 그림 3은 입력 명함 영상에 대한 문자열 영역 추출과 개별 문자 영역 추



[그림3] 명함 문자열 추출 단계

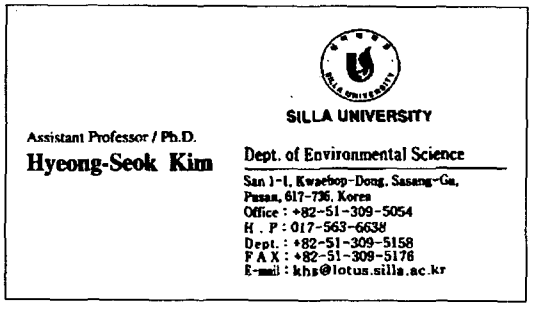
출과정을 나타내었다. 표 2는 문자열 및 비문자열의 추출 개수를 나타내었다. 표 2의 문자열 추출 개수는 30개의 영문 명함에 대해서 가로와 세로의 비율이 1.5 이상이고 면적이 8000보다 적으면 문자열 영역으로 추출한 결과이다. 표 1에서 알 수 있듯이 총 309개의 문자열에서 280개의 문자열이 추출되었다. 표 3은 추출된 280개의 문자열 영역에서 개별 문자를 추출한 결과이다. 개별 문자 추출은 총 4504개의 문자 중에서 4110개의 문자가 추출되었고 최종 개별 문자추출률은 91%이다. 개별 문자 추출에 실패한 경우는 영문 명함에서 문자 영역의 크기가 적고 글자 간격이 거의 없는 경우에는 2개의 개별 문자가 뭉쳐서 추출된 경우이다. 그림 4와 5는 제안된 방법으로 문자열과 비문자열 그리고 개별 문자를 추출한 결과의 한 예를 나타내었다.

<표 2> 문자열 및 비문자열 추출결과

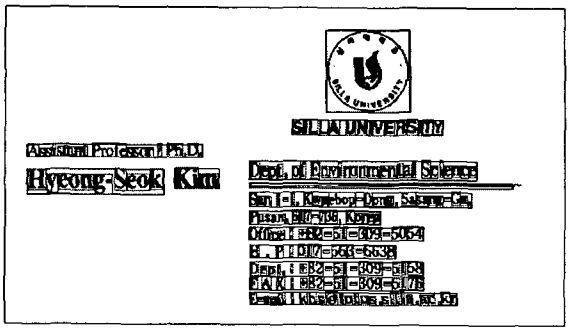
	추출 개수
문자열	280/309
비문자열	19/19

<표 3> 개별 문자 추출 결과

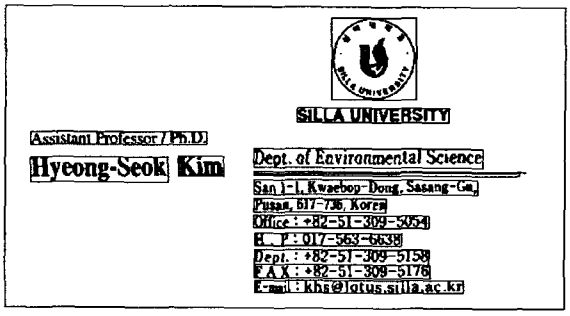
	추출 성공 수	추출 실패 수
개별문자추출	4110	394



(a) 원본 명함 영상

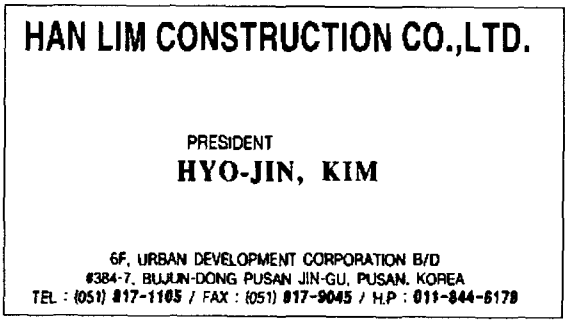


(b) 문자열과 비문자열 추출 결과

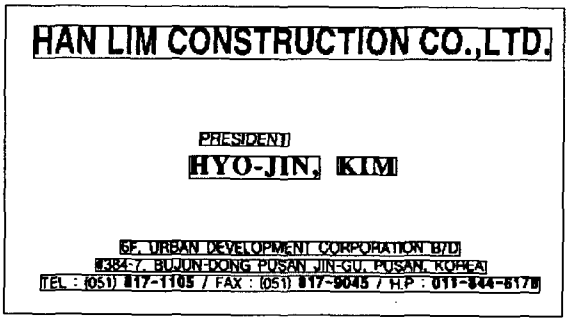


(c) 개별 문자 추출 결과

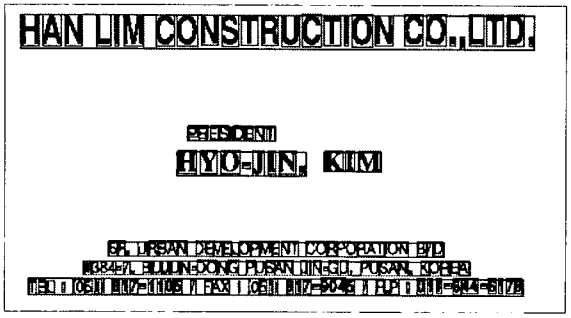
[그림 4] 영문 명함의 개별 문자 추출 결과 1



(a) 원본 명함 영상



(b) 문자열과 비문자열 추출 결과



(c) 개별 문자 추출 결과

[그림 5] 영문 명함의 개별 문자 추출 결과 2

4. 결론 및 향후 연구 방향

본 논문에서는 영문 명함에서 개별 문자를 추출하는 방법을 제안하였다. 영문 명함에서 개별 문자 추출은 문자의 불규칙한 간격과 여백을 처리하는 것이 매우 중요하므로 3×3 마스크를 이용하여 가장 작은 값으로 3 배 축소하는 방법을 적용하여 스미어링하는 시간을 단축시키고 문자들간의 간격을 제거하여 4 방향 윤곽선 추적으로 문자열의 후보 영역을 추출하였다. 추출된 후보 영역에 대해 가로 및 세로의 비율과 면적을 이용하여 문자열과 비문자열로 분리하였고, 개별 문자는 추출된 문자열 영역에서 4 방향 윤곽선 추적 알고리즘을 이용하여 추출하였다. 본 논문에서 사용한 영상의 축소와 스미어링 등의 전처리 방법은 추출 과정 중에 발생할 수 있는 오류율을 현저하게 감소시켰다.

30개의 영문 명함 영상을 실험한 결과 총 309개의 문자열에서 280개의 문자열이 추출되었고 개별 문자 추출은 총 4504개의 문자 중에서 4110개의 문자가 추출되었다. 개발 문자 추출 과정에서 문자의 크기가 적고 문자들의 간격이 거의 없는 경우에는 2개의 문자가 1개의 문자로 추출되는 문제점이 있었다.

향후 연구 방향은 문자열과 비문자열을 정확하게 분리하는 방법과 기울어진 문자들과 붙어서 하나로 추출된 문자들을 처리 할 수 있는 알고리즘을 연구하여 개선할 것이다.

참고문헌

[1] L. O'Gorman and R. Kasturi, "Document Image Analysis Systems," *IEEE Computer*, Vol.5, pp.5-8, 1992.

[2] F. M. Wahl. K. Y. Wong, and R. G. Casey, "Block Segmentation and Text Extraction in Mixed Text/Image Documents," *Computer Vision, Graphics and Image Processing*, Vol.22, pp.375-390, 1982.

[3] 김성영, 권태균, 김민환, "추적에 의한 단순화된 윤곽선 추출," *한국멀티미디어학회 춘계발표 논문집*, pp.356-361, 1999.

[4] E. K. Lim and K. B. Kim," Recognition of Car License Plate using Kohonen Algorithm," *ITC-CSCC*, pp.785-788, 2000.

[5] D. Wang and S. N. Srihari, "Classification of Newspaper Image Blocks Using Texture Analysis," *Computer Vision, Graphics and Image Processing*, Vol.47, pp.327-352, 1989.