영문 명함 영상에서의 문자 영역 추출에 관한 연구

이 지 훈*, 류 재 욱*, 이 준 행*, 신 철 수*, 김 광 백**
*신라대학교 컴퓨터정보공학부
**신라대학교 컴퓨터공학과

A Study on Character Area Extraction of An English Calling Card Image

Ji-Hoon Lee*, Jae-Uk Ryu*, Jun-Hang Lee*, Chol-Soo Shin*, Kwang-Baek Kim**
*School of Computer and Information Engineering, Silla University
**Dept. of Computer Engineering, Silla University

요 약

본 노문에서는 명함 영상에서 문자 영역을 추출하기 위해서 전처리 과정을 수행하여 잡영을 제거한다. 잡영이 제거된 명함 영상을 3배로 축소하여 가로 스미어링을 적용하여 문자열의 후보 영역을 추출하고 문자열과 비문자열의 영역으로 분리한 후, 문자열 영역에 세로 스미어링을 적용한다. 추출된 문자열 영역과 세로 스미어링된 문자열 영역에 대해 OR연산을 수행하여 문자의 특징이 분리되는 것을 제거하고 윤곽선 따라가기 알고리즘을 적용하여 문자의 영역을 추출한다. 제안된 방법을 실제 영문 명함의 개별 문자 추출에 적용한 결과, 기존의 영문 명함 추출 방법보다 개선되었다.

1. 서 론

최근에 명함은 한 개인의 얼굴을 나타내는 것과 같은 역할을 한다. 더욱이 명함에 대한 활용도가 높아지면서 명함에 대한 효율적인 관리가 필요하게 되었다. 기존의 관리방법은 명함 꽂이나 명함집 등을 이용한 것이 대부분이고, 복잡하고 많은 문제를 불러 일으킨다. 예를 들어 동일인이 명함 수정 후 예전 명함과 현재 명함, 두 개를 가지게 된다면 어느 명함이 최신의 것인지 애매하게 된다.

이와 같이 명함 관리 방법을 해결하기 위해서는 명함 인식 프로그램을 통한 데이터베이스 관리가 필요하다. 그리고, 스캐닝 기능을 가지고 있는 입력장치를 통해서 문자를 인식하는 기술이 증가하고 있고, 이에 따라 좀 더 빠른 문자 추출 기술이 요구되고 있다. 또한, 명함 영상의 특성상 이미지와 문자의 위치, 그리고 명함의 구성이 각각 다양하기 때문에 명함 영상에서 문자를 추출하는 것은 매우 중요한 기술이라 할 수 있다[1].

따라서 본 논문에서는 영문 명함 영상에서 문자 영역을 추출하는 방법을 제안하고, 기존의 영문 명함의 문자 추출 방법과 비교 분석한다.

2. 영문 명함 영상의 전처리

스캐너를 통해 입력된 영문 명함 영상은 스캐너 자체의 기계적인 특성에 따라 잡영이 포함되거나 이웃 문자들이 서로 접촉되어 있기 때문에 정확히 문자를 추출 할 수 없는 경우가 발생한다. 따라서 영문 명함 영상의 전처리 과정을 수행하여 잡영 등을 제거하는 과정이 중요하다[2,3].

입력된 영문 명함 영상을 그레이 스케일의 영상으로 변환하여 이진화 한 후 잡영과 선을 제거한다. 잡영과 선을 제거하는 과정은 전체 영상의 잡영을 제거하기 위해서 연결요소(Connected Component)가 일정한 크기 이하의 개수를 가진 연결 요소를 제거하고 영상 내의 문자열 영역과 이미지 영역만을 나타낸다. 또한 문자열에 포함되지 않는 가로 및 세로 선은 연결요소의 상·하·좌·우 방향 성분이 대각 방향의 성분보다 월등히 많은 비율로 나타나는 경우에는 잡영으로 판단하여 제거한다.

2.1 문자열 영역추출

일반적으로 영문 명함 정보에서 필요로 하는 것은 로고나 이미지 등이 아닌 문자 영역이다. 따라서 본 논문에서는 영문 명함 영상에 대해서 문자열과 비문자열 영역으로 분리하고, 문자열 영역에서 개별 문자를 추출한다.

영문 명함 영상에 있는 문자들은 가로 방향으로 일정한 크기와 일정한 간격으로 구성되어 있으므로 문자들간의 간격을 제거하는 가로 스미어링 기법을 적용하여 문자들을 연결시켜서 문자열의 후보 영역으로 추출한다. 하지만, 스미어링 기법은 상당한 시간을 초래하므로, 영문 명함 영상에서의 문자 추출에 소요되는 시간을 단축하기 위해서 영상을 3배로 축소한다. 이 방법은 시간과 공간적인 절약과 영상의 특징점을 추출하는데 적용된다. 영문 명함 영상의 전처리 결과는 그림 1과 같다.

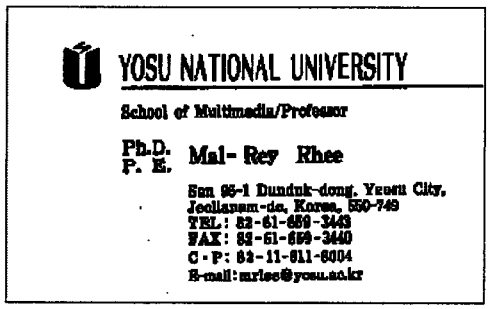

그림 1. 명함 영상 전처리

스미어링 기법은 문지르기 기법이라고도 하며, 문자간의 여백을 설정값 이하로 하여 문자간의 여백을 제거하는 기법이다[4]. 문자간의 여백을 제거하지 않는다면 문자열을 추출하는데 어려움이 있으므로 가로 스미어링을 이용하여 문자열의 후보 영역을 추출한다.

윤곽선 따라가기 알고리즘[5]을 이용하여 문자열의 후보 영역에 대해서 블록화 한다. 이때 추출되는 블럭 영상의 속성은 경계 위치를 나타내는 상·하·좌·우의 좌표값과 블록 영상의 가로와 세로의 길이, 흑화소 밀도(블럭의 총 화소수/블럭의 면적)이며 블럭 영상의 세로 길이와 가로 길이의 비율이 설정값 이상이면 비문자 블럭으로 처리하여 이미지로 저장하고, 설정값 미만이면 문자열의 블록으로 판단하여 위치 좌표를 저장한다. 설정값에 따라 문자열과 비문자열로 윤곽선 따라가기 알고리즘을 적용하여 추출된 영상은 그림 2와 같이 윤곽선 따라가기 알고리즘을 적용하여 문자열과 비문자열의 블록을 추출한 결과이다.

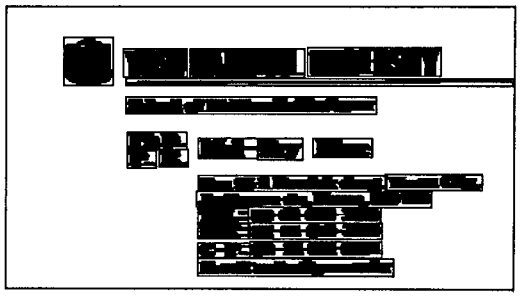

그림 2. 문자열/비문자열 분리영상

윤곽선 따라가기 알고리즘의 단계는 그림 3과 같다.

```
단계 1. 명함 화상을 위에서 아래로, 왼쪽에서
        오른쪽으로 탐색하며,
        IF (새로운 검은 화소가 존재)
            THEN GOTO 단계 2
        ELSE IF (조사할 다음 화소가 존재)
            THEN GOTO 단계 1
        ELSE GOTO 단계 5
단계 2. 시작화소 저장
단계 3. 모든 주변 화소에 대하여
IF(진행방향에서 새로운 검은 화소 존재)
            THEN 처음으로 발견된 화소에
                현 화소와의 위치적 관계 기술
                새로 발견된 화소 위치로 이동.
                GOTO 단계 3
        ELSE GOTO 단계 4
단계 4.
        IF (현 화소가 시작 화소)
    THEN GOTO 단계 1
        ELSE 현 화소가 발견된 화소임을 표시,
            전 화소로 이동.
            GOTO 단계 3
단계 5. 종료
```

그림 3. 윤곽선 따라가기 알고리즘 과정

2.2 개별 문자 영역 추출

추출된 문자열 영역에서 문자 영역을 추출한다. 문자 영역의 추출은 영문 명함 추출의 속도를 개선하는데 중요하며, 사용되는 기법으로는 세로 스미어링을 이용하고 윤곽선 따라가기 알고리즘을 적용한다. 개별 문자 영역을 추출함에 있어서 윤곽선 따라가기 알고리즘만을 적용하면 정확한 문자의 영역을 추출 할 수가 없다[3]. 세로 스미어링은 "j "와 같은 문자에 대해서 하나의 문자로 추출하기 위해 적용한다. 추출된 문자열 블록과 세로 스미어링된 문

자열 블록간의 OR 연산을 수행하여 문자의 특징이 분리되는 것을 제거하고 윤곽선 따라가기 알고리즘을 적용하여 개별 문자 영역을 추출한다. 세로 스미어링 기법을 적용한 영문 명함 영상의 결과는 그림 4와 같고, 추출된 문자열 영역과 세로 스미어링한 영역에 대해 OR연산을 수행하여 문자의 특징들이 분리되는 것을 제거한 후, 윤곽선 따라가기 알고리즘을 적용한 결과는 그림 5와 같다. 본 논문에서 제안된 영문 명함의 문자 영역 추출 구성도는 그림 6과 같다

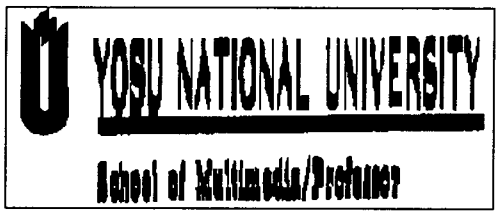

그림 4. 세로스미어링 기법을 적용한 영상

그림 5. 윤곽선 따라가기 알고리즘을 적용한 영상

3. 실험 및 결과

제안된 영문 명함 추출방법의 성능을 평가하기 위해서 1500×800 픽셀 크기의 명함 40개를 대상으로 IBM 호환 기종의 pentiumⅢ에서 C++ Builder 5.0 으로 실험하였다. 기존의 추출방법[2]와 제안된 방법간의 문자열 및 문자 영역 추출 결과는 표 1과 같다.
표 1에서 제안된 추출 방법은 40개의 영문 명함 영상에서 약 98%의 문자 영역이 추출되었고, 기존의 명함 영상 추출 방법[2]보다 제안된 명함 추출 방법이 문자열 영역과 개별 문자 영역의 추출률이 개선된 것을 확인할 수 있다. 기존의 명함 추출 방법에서는 " j " 와 같은 문자 영역이 정확히 추출되지 않았으나 제안된 방법에서는 추출된 문자열 영역과 세로 스미어링된 문자열 영역간의 OR연산을 수행하므로 " j "와 같은 문자들이 정확히 추출되었다.

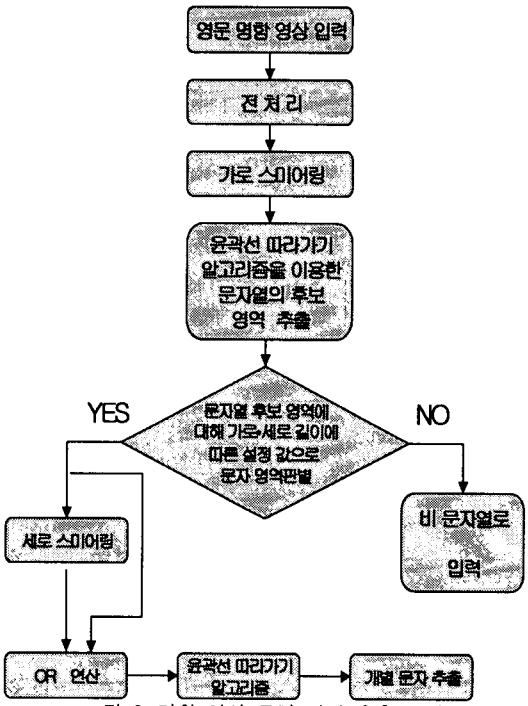

그림 6. 명함 영상 문자 영역 추출 구성도

표1. 문자열 영역과 문자 영역의 추출 결과 비교

	문자열 영역 (총 503개)	문자 영역 (총 6112개)
제안된 명함 추출결과	502	5988
제안된 방법의 오류	1	124
추출률 (OR연산)	0.998	0.979
기존의 명함 추출결과[2]	501	5622
기존의 명함 추출방법 오류	2	490
추출률 (AND연산)	0.996	0.919

그림 7은 가로 스미어링 기법을 적용한 뒤 문자간의 여백을 줄여 문자열로 판별하기 위한 영상이고, 그림 8은 개별 문자를 추출하기 위해서 추출된 문자열 영역을 세로 스미어링한 결과이다. 그림 9는 추출된 문자열 영역과 세로 스미어링한 영역을 OR 연산을 수행하여 윤곽선 따라가기 알고리즘으로 문자 영역을 추출한 결과이다.

그림 7. 가로 스미어링 기법 적용한 영상

그림 8. 세로 스미어링 기법 적용한 영상

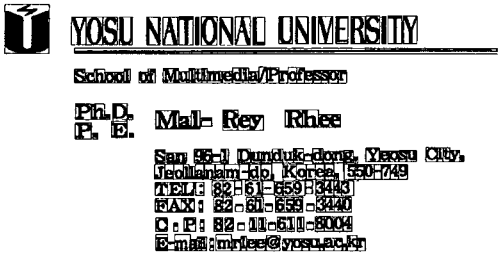

그림 9. 원 영상에서 OR 연산을 한 경우의 문자 영역 추출 결과

4. 결론 및 향후 연구과제

본 논문에서는 영문 명함 인식의 전 단계로 명함 영상에서 문자 영역을 추출하는 방법을 제시하였다. 영문 명함의 문자 추출 시간을 단축하고 추출률을 개선하기 위하여 명함 영상을 3배로 축소하고 가로 스미어링 기법을 적용하여 문자들을 연결시켜서 문자열의 후보 영역으로 추출하였다. 추출된 문자열의 후보 영역은 윤곽선 따라가기 알고리즘을 적용하여 문자열에 대한 블럭을 추출하고 블럭 영상의 경계 위치를 나타내는 상·하·좌·우의 좌표값과 블록 영상의 가로와 세로의 길이, 흑화소 밀도를 이용하여 문자열 영역과 비문자열 영역으로 분리하였다. 추출된 문자열 영역과 세로 스미어링한 영역을 OR연산을 수행하여 한 문자의 특징이 분리되는 부분을 제거하였고 윤곽선 따라가기 알고리즘을 적용하여 문자 영역을 추출하였다.

제안된 방법이 기존의 영문 명함 추출 방법보다 전처리 과정에서 잡영이 많이 줄었고, 문자열 영역의 추출률이 개선되었다. 그리고 추출된 문자열 영역과 세로 스미어링한 영역에 대해 OR연산을 수행하여 문자의 특징이 분리되는 부분을 제거하였고 윤곽선 따라가기 알고리즘을 적용하여 개별 문자의 추출률을 개선하였다.

향후 연구과제로는 문자열과 비문자열을 분리하는 방법을 체계화하는 것이 필요하고, 기울어져 있는 문자와 필기체로 쓰여진 것과 같이 문자들이 서로 붙어서 하나로 추출되는 부분들을 개선할 것이다.

참고문헌

[1] 김두식 "한글 분석 및 인식 기술의 최근 동향," 전자공학회지, 24권, 9호, pp.1058-1070, 1997.

[2] 박소연, 윤수정, 김광백, "윤곽선 추적 알고리즘을 이용한 명함 영상에서의 문자 추출에 관한 연구" 한국 정보처리 학회 추계 발표논문집, 제 8권, 제 2호, pp. 723-726, 2001.

[3] 김광백, 김철기, 김정원, 윤곽선 추적 알고리즘과 개선된 ART1을 이용한 영문 명함 인식에 관한 연구, 한국지능정보시스템학회논문지, 8권, 2호, pp.105-115, 2002.

[4] 김의정, 김태균, "오프라인 문서에서 개별문자 추출과 한자 인식에 관한 연구", 한국 정보처리 학회 논문지, Vol.4, no.5, pp.1277-1288, 1997.

[5] 원남식, 손윤구, "8-이웃 연결값에 의한 병렬 세선화 알고리즘," 정보처리학회논문지, 제2권, 5호, pp.701-710, 1996.