저해상도 팩스 표지 영상의 구조 분석

임 영규, 이 성환
고려대학교 컴퓨터학과/인공시각연구센터
E-mail : {yklim, swlee}@image korea.ac.kr

Structure Analysis of Low Contrast Fax Cover Pages

Young-Kyu Lim and Seong-Whan Lee
Dept. of Computer Science and Engineering/Center for Artificial Vision Research,
Korea University

요 약

팩스가 보편적인 정보 전달 매체로 자리잡게 됨에 따라 기업체나 관공서 뿐만 아니라 가정에서도 많은 작업이 팩스를 통해 이루어지게 되었다. 이에 따라 팩스 문서의 분석 및 인식의 필요성이 증가하게 되었다 팩스 문서는 표지와 내용의 두 부분으로 이루어지는데 팩스 문서의 처리를 위해서는 성명, 주소등을 포함하는 팩스 표지의 분석이 중요하다. 따라서 본 논문에서는 팩스 표지 영상의 구조 분석 방법을 제안한다 제안한 팩스 표지 구조 분석 방법은 팩스 표지가 헤드, 송/수신 정보, 메시지로 구성된다는 점에 착안하여 위치 정보를 이용한 영역 분리에 중점을 두었으며, 팩스 표지의 종류를 몇 가지로 분류하여 도표 형태의 팩스 표지도 분석이 가능하도록 하였다. 문자 인식에서는 팩스 문자 인식에 우수한 성능을 보이고 있는 자소 기반 한글 문자 인식기를 사용하였다. 또한 한글의 자소 모델에 기반한 후처리 방법을 개발하여 인식 오류를 교정하였다.

1. 서론

현대 사회에서 팩스가 보편화된 정보 전달 수단으로 자리잡음에 따라 팩스를 이용한 업무도 많아지고 팩스 문서가 실생활에서 중요한 자리를 차지하게 되었다. 팩스는 송신 되어지는 문서를 영상 형태로 전달 받기 때문에 영상 자체의 처리가 가능하다. 따라서 팩스 문서를 영상 형태에서 자동으로 분석하고 인식하려는 시도가 많이 이루어지고 있다.

팩스 문서는 표지와 내용 부분으로 이루어지는데 팩스 표지에 나타나는 여러가지 정보의 인식을 통해 자동 팩스 라우팅 같은 다양한 응용 분야에 적용이 가능하다. 그러나 팩스 표지 영상은 200dpi 이하의 저해상도 영상이며, 문서 자체의 훼손, 기울어짐 또는 통신망상에서의 노이즈로 인해 변형이 많이 발생하므로 인식에 어려움이 많다.

기존의 팩스 표지 영상의 구조 분석에 관한 연구들은 팩스 수신자 밑에 이중 밑줄이 존재한다고 가정하여 비트 연산을 통해 이중 밑줄을 추출하여 수신자를 인식하거나, 팩스 표지내에서 필드를 구분해주는 세미콜론의 탐색과 키워드 인식을 통한 방법을 사용하였다[1, 2]. 그러나 현재 많이 쓰이고 있는 팩스 표지는 수신자 밑에 이중 밑줄이 있는 형태가 거의 없으며, 또한 세미콜론이나 키워드 인식의 문제도 팩스 표지가 저해상도라는 것을 감안하면 인식 성능을 보장하기 어렵다.

따라서 본 논문에서는 기존 방법들의 문제점을 개선한 저해상도 팩스 표지 영상내에서 수신자를 추출하는 방법을 제안한다. 제안한 방법은 팩스 표지 영역 분리를 통해 팩스 표지를 헤드와 송/수신 정보, 메시지 영역으로 분리 후 수신자 부분이 들어있는 송/수신 정보 영역의 정밀 탐색을 통해 수신자를 찾는다. 팩스 문자 인식은 저해상도 팩스 문자 인식에서 우수한 성능을 보여주고 있는 자소 단위 한글 문자 인식기를 사용하며 후처리를 통해 인식 오류를 교정한다.

2. 팩스 표지 영상 구조 분석

팩스 표지는 저해상도 문서이며 형태와 종류가 다양하기 때문에 문자 인식을 통한 구조 방법만으로는 좋은 결과를 나타낼 수 없다 본 논문에서는 다양한 팩스 표지에 적합한 구조 분석을 위해 팩스 표지를 몇 가지로 분류하여 각각에 적합한 분석 방법을 적용하였다. 제안된 방법은 그림 1과 같은 단계를 거쳐 수행된다

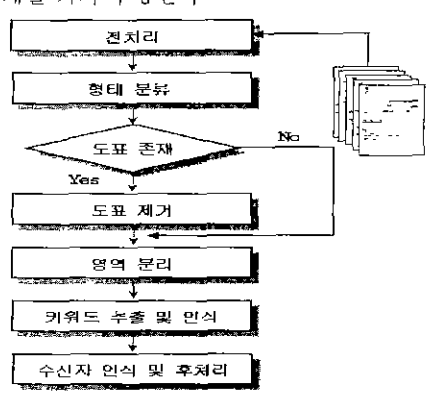

그림 1 제안된 팩스 표지 영상의 구조 분석 방법

2.1 전처리 과정

전처리 과정은 팩스 표지 영상에 나타나는 잡영을 제거하고 다양한 형태의 변형으로부터 원 영상에 가깝게 복원하는 역활을 한다. 특히 기울어져서 입력된 팩스 표지 영상을 바로잡는 기울어짐 교정은 전처리 과정의 중요한 역활 중 하나이다 기울어짐을 교정하기 위해서는 일반적으로 허프 변환을 많이 쓴다. 허프 변환의 수행시간은 변환하는 점의 갯수에 비례하므로, 속도의 향상을 위해서 연결 요소 분석을 통해 나온 최소 외접 사각형의 제일 끝점을 기준으로 허프 변환을 수행한다[3].

2.2 팩스 표지 종류 인식 및 도표 처리

전처리 과정을 거친 팩스 표지 영상은 표지 내의 도표의 존재 유무에 따라 도표 형태의 팩스 표지와 일반적인 형태의 팩스 표지로 분류된다. 기존의 방법에서는 도표 형태의 팩스 표지는 인식이 불가능하였지만, 제안된 방법에서는 인식이 가능하다.

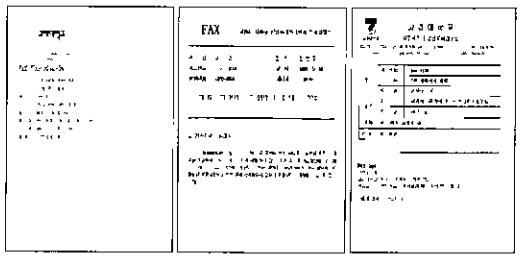

그림 2. 다양한 팩스 표지의 예

팩스 표지는 일반적인 논문이나 잡지와는 달리 그림과 도표가 복잡하게 존재 하지 않으며, 대개 그림 은 팩스 표지의 상단 부분에 많이 나타난다[4]. 먼저 연결 요소의 크기 정보를 이용하여 도표 후보들을 추출한다. 연결 요소 분석 과정에서 나타나는 크기가 큰 연결 요소들은 그림, 그래픽, 도표이다. 추출된 후보들을 위치 정보와 도표가 지니는 특징인 비슷한 길이의 수직, 수평선의 존재 유무로 검증한다. 도표 형태는 식 1과 2의 수평, 수직 히스토그램을 통해 파악을 한다

$$HL(X[i]) = \begin{cases} True & if\ X[i] > T_h \\ False & otherwise \end{cases} \quad (1)$$

$$VL(Y[i]) = \begin{cases} True & if\ Y[i] > T_v \\ False & otherwise \end{cases} \quad (2)$$

수식에서 임계값 T_h, T_v 는 수직, 수평 방향 선 성분을 구분할 수 있는 최소 크기이다. 도표의 형태가 파악되면 글자들을 둘러쌓고 있는 모든 선을 없애서 글자 부분만 남긴다. 이후 연결 요소 분석을 한번 더 수행한다.

2.3 팩스 표지 영상 영역 분리

일반적으로 많이 쓰이고 있는 팩스 표지는 어느 정도의 동일한 형식이 나타난다 팩스 표지는 그림 3에서 보는 봐와 같이 회사 로고나 주소 등이 들어 있는 헤드와 팩스 송수신자에 관한 내용이 들어 있는 송/수신 정보 영역과 여러 가지 전달하고자 하는 내용을 담고 있는 메시지로 나눌 수 있다 팩스 표지 영역 분리는 이렇게 세 부분으로 나누어진 팩스 표지를 분석하여 수신자가 들어 있는 송/수신 정보 영역만을 추출하는데 목적이 있다. 송/수신 정보 영역만을 탐색함으로써 불필요한 부분의 탐색을 피하게 되므로 전체 시스템의 수행 속도를 향상시킬 수 있으며, 인식과정에서의 오류도 줄일 수 있다.

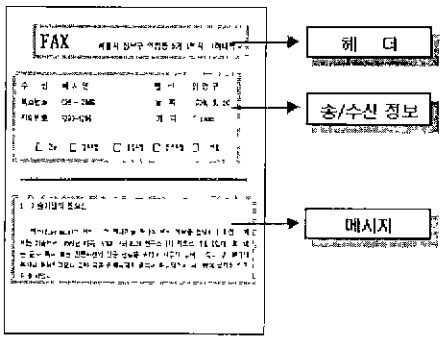

그림 3. 팩스 표지의 구조

대부분의 팩스 표지에는 각 영역을 분리 시켜 주는 영역 구분자가 존재한다. 이러한 구분자는 대체로 선분 형태로 되어있다. 팩스 표지에서 보이는 선들은 팩스의 전송 중에 일부분이 손실되거나 잡영이 첨가되어 연결된 선을 이루지 않기 때문에 찾기가 쉽지 않다. 따라서 run length smoothing 작업을 통해 중간에 손실된 부분을 이어준다. 선 성분은 연결요소의 길이가 길고 높이가 임계값 이하이면 선으로 간주한다.

2.4 키워드 추출 및 인식

팩스 표지의 영역 분리가 이루어진 후 우리가 원하는 정보는 송/수신 정보 영역에 모여 있게 된다 송/수신 정보 영역에는 팩스 송수신에 관한 여러 가지 항목들이 있는데 각 항목들을 각각의 블록으로 묶는다. 블록화 작업은 식 3에 의해 이루어진다. 각 항목들을 하나의 블록으로 묶을 경우 블록내에는 각 항목을 가르치는 키워드와 그와 관련되는 내용이 같이 묶이게 된다. 키워드는 하나의 블록내에서 항상 제일 왼쪽 편에 위치하고 세미콜론 같은 필드 구분자에 의해 내용 부분과 분리가 된다.

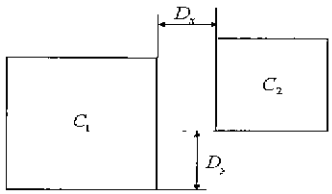

그림 4. 연결 요소의 위치 관계

$$MG(C1, C2) = \begin{cases} True & if\ D_x < T_x \text{ and } D_y < T_y \\ False & otherwise \end{cases} \quad (3)$$

블록화 작업이 이루어지고 나면 키워드의 위치를 파악하여 인식하는 과정을 수행한다. 팩스 표지에서 키워드는 송/수신에 관계되는 단어로 "수신", "발신", "성명" 등이 이에 해당한다 팩스 글자는 200 dpi 이하의 저해상도이기 때문에 인식에 상당한 어려움이 있다. 본 논문에서는

팩스 문자 인식에 우수한 성능을 보이는 자소 단위 한글 인식기를 이용하여 키워드를 인식한다[5]

2.5 수신자 정보 인식 및 후처리

팩스 표지 내에 수신자는 해당하는 키워드의 우측편에 위치하게 되는데, 키워드 바로 다음에는 일반적으로 세미콜론이 많이 나타난다. 세미콜론은 글자의 평균 크기와 비교하여 구분한다. 세미콜론이 인식되면 세미콜론 오른쪽에 있는 수신자의 성명을 인식한다.

그러나 팩스 표지를 보면 수신자 부분에 성명뿐만 아니라 직책, 부서명, 학교명 등이 같이 있는 경우가 많이 있다. 이와 같은 경우에 수신자의 성명만을 추출하는 방법이 필요하다. 대개 한국인의 성명은 두 자에서 네 자 정도이기 때문에 단어 길이가 여기에 해당하는 것을 이름 후보들로 추출한 다음, 수신자 사전과 매칭을 통해 일정한 임계값 이상의 매칭율을 보이면 성명으로 인식한다.

인식된 성명은 인식기 성능의 한계로 인해 오류가 발생하기 때문에 후처리 과정이 필요하다. 미리 구축된 수신자 사전과 한글 자소에 대한 거리 함수를 이용하여 인식 오류를 교정한다

3. 실험 및 결과 분석

본 팩스 표지 영상 구조 분석기는 펜티엄 166MHz PC 상에서 MS Visual C++ 1.52를 사용하여 구현되었다. 팩스 표지 분석에 사용된 영상은 팩스 모뎀을 통해 수신된 82개를 사용하였다. 이중 도표 형태의 팩스 표지는 모두 8개였다. 문자 인식기로는 팩스 문자 인식에 우수한 성능을 보여준 바 있는 자소 단위 한글 문자 인식기를 사용하였다.

실험 결과는 표 1에서 알 수 있듯이 약 92%의 수신자 인식률을 나타내었다. 표 1, 2에서 N_c는 정인식된 갯수를 나타내며, N_f는 오인식된 갯수를 나타낸다. 또한 F_r는 영역 분리 오류의 갯수, F_k는 키워드 인식 오류의 갯수를 나타내며, F_n은 성명 인식 오류의 갯수를 나타낸다. 오인식 결과를 분석해 볼 때 영역 분리가 제대로 되었지만 이를 인식에 실패한 경우가 있는 것은 팩스 문자 인식의 어려움에서 기인한다고 볼 수 있다

표 1. 수신자 성명 인식 결과

종 류	N	N_c	N_f
일반적인 표지	74	69	5
도표 형태 표지	8	6	2

표 2. 오인식 결과 분석

종 류	F_r	F_k	F_n
일반적인 표지	2	1	2
도표 형태 표지	0	1	1

그림 4는 입력된 팩스 표지에 대한 구조 분석 과정을 단계적으로 보여주고 있다. 그림 4의 (a)에서 팩스 표지 영상이 입력으로 들어오면, (b)에서 연결 요소 분석을 수행하여 송/수신 정보 영역을 분리한다. 그림 (c)는 송/수신 정보 영역에서 블록화 작업을 보여주고 있으며, (d)에서 키워드 인식을 통한 수신자 영역 추출 과정을 보여주고 있다. 수신자 영역이 추출되면 문자 인식기를 이용하여 수신자의 성명을 인식한다.

현재까지는 인쇄체 문자만 있는 팩스 표지에 대해서만

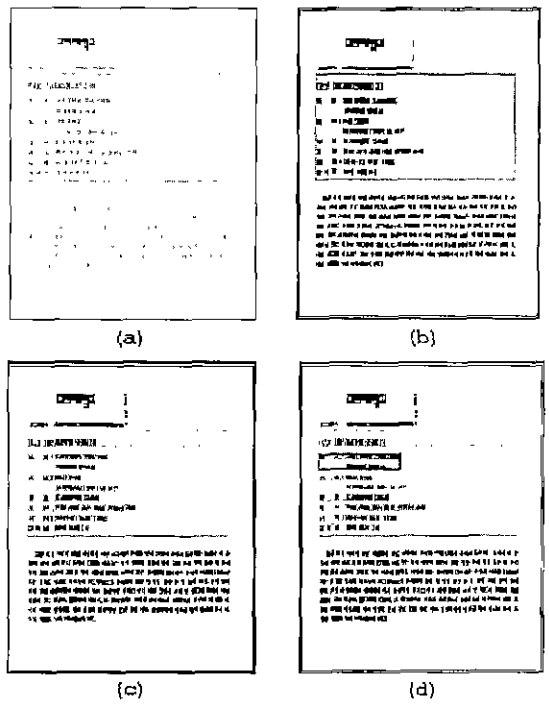

그림 5. 팩스 표지 영상 구조 분석 과정

분석을 하였으나, 앞으로는 필기체 팩스 표지 분석에까지 범위를 확장해 나갈 예정이다

감사의 글

본 연구는 1997년도 정보통신부 산학연 과제의 연구비 지원을 받았음

참고 문헌

[1] J Li and S N. Srihari, "Location of Name and Address on Fax Cover Pages," Proc of 3rd Int. Conf on Document Analysis and Recognition, Montreal, Canada, August 1995, pp. 756-759.

[2] T. Akuyama, "Addressee Recognition for Automated FAX Mail Distribution," Proc of SPIE Conference on Document Recognition(III), Vol. 2660. San Jose, California, January 1996, pp. 677-680.

[3] 김 두식, 이 성환, "한영 혼용 문서의 디지탈 라이브러리 구축을 위한 효과적인 문서 기울기 교정 및 문자 분할 방법," 한국정보과학회 춘계학술발표논문집, 대구, 제 23권 제 1호, 1996년 4월, pp 293-296.

[4] G. Ricker, A. Winkler, "Recognition of faxed documents," Proc. of SPIE Conference on Document Recognition, Vol. 2181, San Jose, California, February 1994, pp. 371-377.

[5] 김 두식, 이 성환, "저해상도 인쇄체 한글 인식을 위한 자소 기반 방법과 음절 기반 방법의 성능 비교," 한국정보과학회 추계학술발표논문집, 용인, 제 23권 제 2호, 1996년 10월, pp. 587-590