論文94-31B-8-15

우리나라 우편 봉투 영상에서의 주소 영역 추출을 위한 지식 기반 시스템

(A Knowledge-Based System for Address Block Location on Korean Envelope Images)

金 基 喆*, 李 晟 瑍**

(Ki-Cheol Kim, Seong-Whan Lee)

要 約

본 논문에서는 우리나라 우편 봉투 영상의 구조를 분석하여 수취인 주소 영역을 자동적으로 추출하는 지식 기반 시스템을 제안한다. 제안된 시스템은 우편 봉투 영상의 효과적인 구조 분석을 위하여 적응적 이진화, 연결 요소 추출, 기울어진 영상 교정 등의 전처리를 수행한 다음, 다양한 우편 봉투에 대한 통계적인 특징 분석을 통하여 얻은 지식에 기반을 둔 반복적인 처리 과정에 의해 우편 봉투 영상에서 수취인 주소 영역을 추출하고, 수취인 주소, 성명, 우편번호 부분을 추출한다. 현재까지 발표된 수취인 주소 영역을 추출하는 시스템의 대부분은 우편봉투 영상을 여러개의 수취인 주소 후보 영역으로 분할한 다음 후보 영역 중에서 하나의 주소 영역을 선택하는 방식으로 구성되어있다. 그러나, 우리나라 우편 봉투에서의 주소 영역들은 서로 근접한 위치에 필기되기 때문에 본 논문에서는 이러한 분할과 선택의 과정을 처리하지 않고 연결 요소를 병합, 분할하고 검증하는 과정을 반복함으로써 수취인 주소 영역을 추출하는 반복적인 과정을 제안한다.

서울 우편집중국으로부터 제공받은 다량의 실제 우편 봉투에 대해 실험한 결과, 제안된 시스템이 우리나라 우편 봉투 영상에서의 주소 영역 추출에 매우 효과적임을 알 수 있었다.

Abstract

In this paper, we propose a knowledge-based system for locating Destination Address Block(DAB) by analyzing the structure of Korean envelope images. In the proposed system, the preprocessing steps such as adaptive binarization, connected component extraction, and deskewing are carried out first for the effective structure analysis of the envelope image. Then, DAB containing address, name and zipcode parts of the input envelope image is extracted by an iterative procedure based on the knowledge acquired from the statistical feature analysis of the various envelope images. Most of the system for locating address blocks on envelopes have extracted DAB by segmenting an envelope image into several candidate blocks followed by selecting one among the candidate blocks. Because it is very difficult to segment a Korean envelope image into several blocks due to the specific writing habits that the addresses on the envelope are written in close proximity to each other, the proposed iterative procedure determines DAB by splitting or merging the connected components, and verifies the determined DAB without segmentation and selection.

Experiments with a great number of the live envelopes provided from Seoul Mail Center in Korea were carried out. The results reveal that the proposed system is very effective for address block location on Korean envelopes.

*學生會員, **正會員, 忠北大學校 컴퓨터科學科
(Dept. of Computer Science, Chungbuk Nat'l Univ.)

※ 본 시스템은 제 1회 문자 인식 워크샵에서 전시된 시스템의 기능을 수정 보완하였음을 밝힙니다.

接受日字 : 1993年 6月 20日

Ⅰ. 서 론

현대 사회에서 정보 전달의 수단으로 중추적인 역할을 담당하고 있는 우편물의 양은 날마다 폭발적으로 증가하고 있으며 이러한 우편물을 발송하기 위하여 각 지역별로 우편물을 분류하는 데는 많은 인력과 시간이 필요하기 때문에 보다 신속하고 정확한 우편 소통을 위하여 우편물 자동 분류에 대한 연구가 진행되고 있다. 이 연구 분야에서 우편물의 구조를 분석하여 수취인 주소 영역을 추출하는 것은 문자 인식 기술과 더불어 우편물 자동 분류에 반드시 필요한 핵심 기술이라 할 수 있다. 그러나, 우편물에는 수취인 주소 영역 뿐만 아니라 발신인 주소, 우표, 소인, 광고 문안, 그림 등 주소 영역 이외의 영역이 포함되어 있으며 필기구의 종류, 필기자의 필기 습관, 우편물의 표면 상태 등에 따라 다양한 변형을 포함하고 있기 때문에 정확한 수취인 주소 영역의 추출에는 많은 어려움이 있다.

국외에서는 1970년대 초부터 대학과 연구소를 중심으로 우편 자동화에 대한 연구가 진행되어 현재에는 시스템의 실시간화와 성능 향상 등 실용화에 주력하고 있는 반면, 국내에서의 연구는 초기 단계여서 1990년에 서울 우편집중국을 개국하여 독일 AEG사의 우편 자동 분류 시스템과 운반 시설을 완비하고 우편물을 자동으로 분류함으로써 우편 작업의 생산성 향상에 노력하고 있다. 그러나, 이 시스템은 고가의 수입품일 뿐만 아니라 엽서와 규격 봉투만을 처리할 수 있으며, 일정 각도 이상으로 기울어져 필기된 우편번호는 처리할 수 없는 문제점이 있다. 특히, 우편번호만을 인식 대상으로 하고 있어서, 우편번호가 기재되지 않으면 인식할 수 없으며 우편번호를 잘못 기재한 경우 정확하게 분류할 수 없기 때문에 우편번호만을 인식하여 우편물을 자동 분류하는 데는 그 한계가 있다. 따라서, 우편번호 뿐만 아니라 필기된 주소를 동시에 인식하여 자동 분류를 위한 정보로 사용함으로써 오분류를 최소화하고, 우편 영상에 발생할 수 있는 변형이나 필기 변형을 효과적으로 흡수할 수 있는 우편 자동 분류 시스템의 개발에 관한 연구가 절실한 실정이다.

본 논문에서는 우리나라 우편 봉투 영상의 구조를 분석함으로써 수취인 주소 영역을 자동적으로 추출할 수 있는 지식 기반 시스템을 제안한다. 그림 1은 제안된 지식 기반 시스템의 전체적인 구성을 나타낸다. 제안된 시스템에서는 우편 봉투 영상의 효과적인 구조 분석을 위하여 적응적 이진화, 연결 요소 추출, Hough 변환을 이용한 기울어진 영상 교정 등의 전처리를 수행하였다. 현재까지 발표된 수취인 주소 영역을 추출하는 시스템의 대부분은 우편 봉투 영상을 여러개의 수취인 주소 후보 영역으로 분할하고, 지식을 이용하여 후보 영역 중에서 수취인 주소 영역을 선택하는 방식으로 수취인 주소 영역을 추출하였는데, 이러한 방법으로 수취인 주소 영역을 추출할 경우 우리나라 우편 봉투의 구조나 필기자의 습관을 고려해볼 때, 우리나라 우편 봉투에서의 주소 영역들은 서로 근접한 위치에 필기되기 때문에, 우편 봉투 영상을 수취인 주소 후보 영역으로 분할하기 어려울 뿐만 아니라 분할된 수취인 주소 후보 영역에 발신인 주소나 우표 및 소인 영역이 포함될 가능성이 상당히 높다. 따라서 본 논문에서는 수취인 주소 후보 영역을 설정하지 않고 다양한 우편 봉투의 통계적인 특징 분석을 통하여 얻은 지식을 이용하여 연결 요소를 병합, 분할하고 검증하는 과정을 반복함으로써 우편 봉투 영상에서 수취인 주소 영역을 추출하고, 수취인 주소, 성명, 우편번호 부분을 추출하였다.

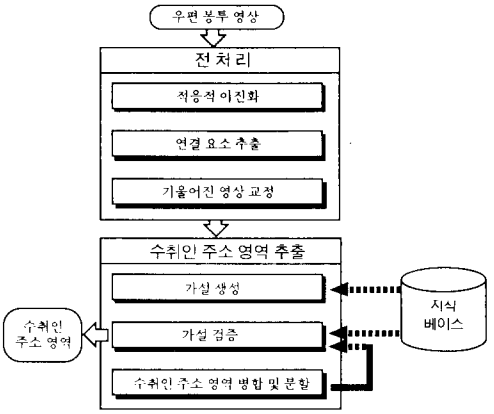


그림 1. 제안된 지식 기반 시스템의 전체적인 구성.

Fig. 1. Overview of the proposed knowledge-based system.

본 논문의 구성은 다음과 같다. Ⅱ장에서는 우편 봉투에서 수취인 주소 영역의 추출과 관련된 연구를 소개하며, Ⅲ장에서는 본 연구에서 사용한 효과적인 우편 봉투의 구조 분석을 위한 전처리 기법을 기술한다. Ⅳ장에서는 우편 봉투에서 통계적인 특징을 추출하여 규칙 형태로 표현된 지식 베이스를 구축하는 방법에 대하여 소개하고, Ⅴ장에서는 구축된 지식 베이스를 이용하여 우편 봉투 영상에서 수취인 주소 영역을 분할하고, 분할된 수취인 주소 영역에서 수취인

주소, 성명, 우편번호 부분을 추출하는 방법을 소개한다. Ⅵ장에서는 실험 및 결과 분석에 대하여 기술하고, 마지막으로 Ⅶ장에서는 결론 및 앞으로의 연구 방향을 제시한다.

Ⅱ. 관련 연구

본 장에서는 우편 봉투 영상에서 수취인 주소 영역을 추출하는 기존의 연구를 종합하여 우편 봉투 영상의 구조 분석을 위한 전처리, 영역 단위 분할에 의한 후보 수취인 주소 영역 설정, 후보 영역 중에서 수취인 주소 영역의 선택 등 세 부분으로 나누어서 기술한다.

우편 봉투 영상의 구조 분석을 위한 전처리는 우편 영상에서 문자 부분만을 강조하거나 처리하고자 하는 데이타량을 감축하거나 영상의 변형을 교정하는 등 우편 봉투 영상을 효과적으로 분석하기 위해 수행하는 과정으로, 영상 향상(Image enhancement), 이진화, 기울어진 영상 교정 등이 사용된다. 영상 향상은 영상의 유용성을 증가시키기 위해 선택된 특징을 강조하거나 억제하는데 사용되는 기법으로 마스크를 사용하는 공간 영역(Spatial-domain) 방법과 Fourier 변환이나 히스토그램 수정 등을 이용하는 주파수 영역(Frequency-domain) 방법 등 영상 처리에 관한 기법들이 이용된다.[1] 이진화는 명도(Gray scale) 영상을 0 또는 1의 이진 영상으로 변환하는 과정으로 전체적 이진화(Global binarization)는 화소들의 밝기 분포 히스토그램에서 전경과 배경에 큰 피크가 존재하는데, 이 두 피크 사이의 중간값을 임계값으로 하여 이진화를 수행한다. 적응적 이진화(Adaptive binarization)[2]는 인접 화소에 대한 밝기의 차이에 따라 적응성을 갖는 장점이 있다. 기울어진 영상을 교정하는 데는 Hough 변환에 의한 방법[2-4], 연결 요소의 히스토그램 분석 방법[5] 등이 있다.

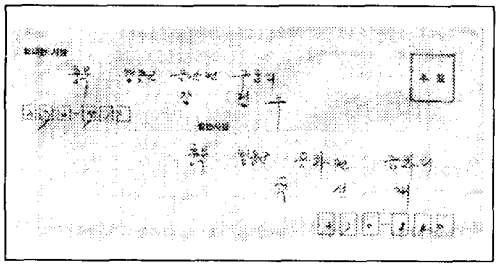
우편 봉투 영상을 영역 단위로 분할하는 방법은 크게 하향식(Top-down)과 상향식(Bottom-up)으로 분류할 수 있다.[6] 하향식은 영상의 일반적인 특성에 근거하여 영상을 점점 작은 영역으로 분할하는 방법으로 미리 알고있는 지식을 영상 분할에 사용할 수 있다는 장점이 있는 반면에 복잡한 형태의 영상이나 기울어진 영상에서 영역을 분할하는 데는 부적합하다는 단점이 있다. 하향식으로는 런 길이 평활화(Run length smoothing) 방법[5,7], 투영 윤곽(Projection profile) 분석 방법 등이 있다. 런 길이 평활화 방법은 먼저 수평 방향의 평활화를 수행하고, 수직 방향의 평활화, 두 결과에 논리적 AND 연산 적용 그리고 마지막으로 부가적인 수평이나 수직 방향의 평활화를 수행함으로써 필요한 영역을 분할한다. 투영 윤곽 분석 방법은 대부분의 문서 영상이 수직 또는 수평 방향의 블럭으로 구성되므로 반복적으로 적용할 경우 매우 효율적으로 각 영역을 분할할 수 있는 특성이 있다. 상향식은 가장 기본이 되는 작은 부분에서 출발하여 유사한 특성을 갖는 부분을 단계적으로 병합함으로써 원하는 형태를 생성하는 방법으로 연결 요소(Connected component) 분석 방법[2,5], 전이(Transition) 분석 방법[8] 등이 있다. 연결 요소란 임의의 흑화소나 백화소에 대하여 8방향 또는 4방향의 인접 화소가 같을 경우 이 화소를 모두 연결하여 얻은 화소의 집합을 말한다. 연결 요소 분석 방법에서는 유사한 특성을 갖는 연결 요소를 모아서 하나의 연결 요소로 만들고 이를 분석하여 문자 영역만을 추출한 후 단어, 문자열, 절을 만들기 위한 병합을 수행한다. 이 방법을 사용하면 활자체나 문자 크기의 변화에 관계없이 임의의 방향에 위치한 문자열을 분리할 수는 있다는 장점이 있다. 전이 분석 방법은 수직 또는 수평 주사 방향을 따라 이동하면서 백화소에서 흑화소로 또는 그 반대로 바뀌는 점의 수나 그 거리의 특징을 분석하여 규칙을 생성하고 이 규칙에 따라 문자를 추출하고 문자를 병합하여 문자열과 영역을 추출하는 방법이다.

분할된 각 후보 영역에서 수취인 주소 영역을 선택하는 데에는 주로 우편 봉투를 분석하여 얻은 지식을 사용하는 지식 기반(Knowledge-based) 방법을 이용하는데, 이러한 지식을 표현하고 표현된 지식을 이용하는 방법에 따라 모델 기반(Model-based) 방법, 규칙 기반(Rule-based) 방법[2,8], 모델 기반과 규칙 기반을 결합한 방법[9] 등으로 나눌 수 있다. 모델 기반 방법은 지식을 각 영역에 대한 프레임의 형태로 표현하는 방법이다. 이 방법에서 지식은 문제에 대한 문제 영역 지식(Domain knowledge)과 문제 영역 지식을 다루기 위한 메타 지식(Meta knowledge)으로 구성되는데, 지식을 사용하여 수취인 주소 영역을 선택하기 위하여 가설을 생성하고 선택한 후, 검증하고 평가하는 과정을 반복한다. 규칙 기반 방법은 지식을 각 구성 요소에 대하여 IF ~ THEN ~ 형태로 표현하고 이를 이용하여 수취인 주소 영역을 선택하는 방법이다. 여기에서 IF 부분은 각 후보 영역의 특징(영역내의 문자 수, 문자열의 수, 위치, 크기 등)을 검사하는 조건을 나타내는 것이며, THEN 부분은 IF 부분의 조건이 만족되었을 때 수행해야 할 처리나 특정 후보 영역이 수취인 주소 영역일 신뢰도를

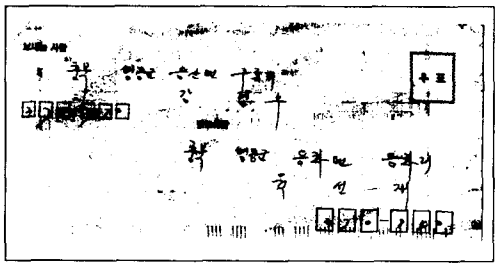
나타내는 것으로 확률 이론이나 퍼지 집합이론이 이용된다.

Ⅲ. 우편 봉투 영상의 전처리

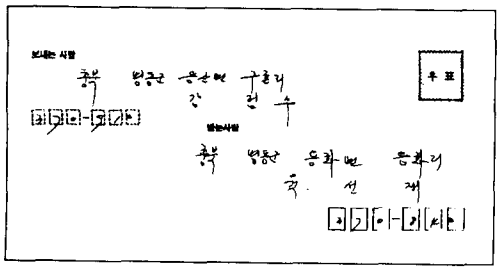
본 연구에서는 전반적으로 처리 시간을 단축하고 영상의 기울어짐에 무관하게 수취인 주소 영역을 추출할 수 있도록 적응적 이진화, 연결 요소 추출, 기울어진 영상 교정 등의 전처리를 수행하였다.



(a)



(b)



(c)

그림 2. 이진화의 예
(a) 우편 봉투의 명도 영상 (b) 전체적 이진화를 거친 우편 봉투 영상 (c) 적응적 이진화를 거친 우편 봉투 영상

Fig. 2. An example of the binarization.
(a) gray scale image of an envelope (b) envelope image with global binarization (c) envelope image with adaptive binarization.

1. 적응적 이진화

전체적 이진화는 하나의 임계값을 결정하기가 매우 어려울 뿐만 아니라 우편 봉투 영상의 주소 영역이나 배경이 특정한 색을 갖는 경우가 많기 때문에 우편 봉투 영상의 이진화에는 효과적이지 못하다. 따라서, 9 x 9 윈도우의 인접 화소에 대한 명도의 차이에 따라 적응성을 갖는 Srihari 등[2]의 적응적 이진화를 이용하여 우편 영상을 이진화하였다. 그림 2(b)와 (c)는 그림 2(a)의 우편 봉투의 명도 영상을 각각 전체적 이진화, 적응적 이진화에 의해 이진화한 영상의 예이다.

2. 연결 요소 추출

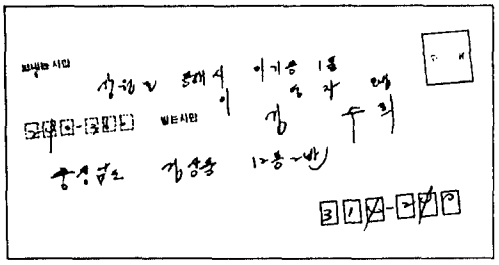
연결 요소는 임의의 흑화소에 대하여 8방향의 인접 화소가 흑화소일 경우 흑화소들을 모두 연결하여 추출하고, 연결 요소를 포함하는 최소 사각형의 위치, 최소 사각형의 종횡비, 사각형 내 흑화소의 밀도 등의 특징을 추출하여 기울어진 영상 교정이나 수취인 주소 영역을 추출할 때 이용한다. 그림 3(b)는 그림 3(a)의 기울어진 영상에서 연결 요소를 추출하여 연결 요소를 포함하는 최소 사각형을 표현한 영상이다.

3. 기울어진 우편 봉투 영상의 교정

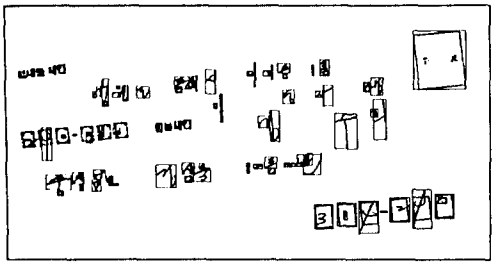
우편 봉투 영상의 기울어짐은 그 정도가 적을 경우에도 우편 봉투 영상 인식 시스템의 성능을 급격히 저하시킨다. 또한 문서 영상의 구조 분석 방법 중 비교적 빠르기 때문에 널리 사용되고 있는 투영 윤곽 분석이나 런 길이 평활화 방법 등은 기울어진 영상에는 적용하기 어렵기 때문에 영상의 기울어짐을 교정하는 것은 우편 봉투 영상의 구조 분석에서 필수 과정이라 할 수 있다. 본 연구에서는 기울어진 우편 봉투 영상의 교정을 위하여 Hough 변환을 이용하였다. 기본적으로 Hough 변환이란 임의의 각도에 위치한 선분과 그 선분의 기울어진 각도를 검출하는데 사용하는 방법으로 영상 공간 (x, y)를 파라미터 공간 (ρ ,θ)의 곡선으로 대응시키는 $\rho = x \cdot \cos\theta + y \cdot \sin\theta$ 변환으로 이루어지며, 이 변환에서 (ρ , θ)에 따라 그 교차의 수를 누적한 2차원 배열에서 가장 큰 값을 갖는 θ 가 원래 영상의 기울어진 각에

대응된다. [3, 4, 10]

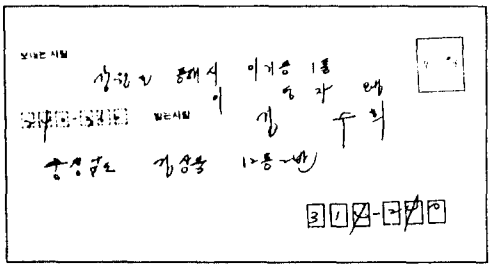
문자열의 경우에는 실제 선분은 아니지만 밀도가 낮고 두꺼운 선분의 일종이라 가정할 수 있으며, 같은 문자열의 Hough 변환 결과는 같은 ρ 값에서 누적되도록 ρ 의 간격을 적절히 조정해야 한다. 본 연구에서는 연결 요소를 추출하여 너무 작거나 너무 큰 연결 요소는 제외하고, 남아있는 연결 요소의 평균 높이를 계산하여 ρ 의 간격을 결정하였다. 한편, Hough 변환을 이용하면 영상의 기울어진 정도를 비교적 정확하게 찾을 수 있는 반면에 영상의 모든 화소를 고려하기 때문에 너무 많은 계산량이 요구된다. [3] 따라서 본 연구에서는 연결 요소를 포함하는 최소 사각형의 중점을 기준으로 Hough 변환을 수행함으로써 처리 시간을 단축하였다. 본 연구에서는 Fletcher와 Kasturi [3] 의 방법을 이용하여 우편 봉투 영상의 기울어진 각을 추정하였으며, 우편 봉투 영상의 기울어진 각이 일정한 각 이상이면 2차원 회전 변환을 이용하여 우편 봉투 영상의 기울어짐을 교정하였다. 그림 3(c)는 그림 3(a) 우편 영상의 기울어짐이 교정된 이진 영상이다.



(a)



(b)



(c)

그림 3. 기울어진 우편 영상 교정의 예
(a) 기울어진 우편 영상 (b) 연결 요소가 추출된 영상 (c) 기울어짐이 교정된 영상

Fig. 3. An example of deskewing the skewed envelope image.
(a) a skewed envelope image
(b) envelope image in which connected components are extracted
(c) deskewed envelope image

Ⅳ. 우편 봉투 영상의 통계적인 특징 추출과 지식 베이스 구축

우편 봉투 영상은 크게 수취인 주소, 발신인 주소, 우표 및 소인, 그림 및 기타 영역 등으로 구분할 수 있으며, 수취인 주소 영역은 다시 수취인 주소, 성명, 우편번호 부분 등으로 구분할 수 있다. 본 연구에서는 각 영역과 부분에 대해 다음과 같은 특징을 추출하였다.

- 각 영역과 부분에 대한 특징: 상대적인 위치, 폭, 높이.
- 연결 요소, 단어, 문자열에 대한 특징: 각 요소의 갯수, 폭, 높이, 화소의 밀도, 단어간격, 문자열 간격.

위와 같은 우편 봉투에 대한 특징을 다량의 우편 봉투에 대해 추출하고, 이를 통계적으로 분석하여 수취인 주소 영역을 추출하기 위한 정보를 다음과 같은 규칙 형태의 지식으로 표현하여 지식 베이스를 구축하였다.

IF *condition*(*block*)
THEN *block is the region with probability P*

여기에서 *condition*(*block*)은 어느 특정 영역 block의 특징을 검사하는 조건이고, *P*는 IF 부분의 조건을 만족함으로써 얻게 되는, *block*이 우편 봉투에서의 특정 영역 *region*일 확률을 나타내는 것으로, 검사하는 조건을 c라 하고 r^i를 우편 봉투의 각 영역이라 할 때 특정 영역 *block*을 r^i영역이라고 할 사후 확률(Posteriori probability) $P(r^i|c)$를 나타낸다.

본 연구에서 사후 확률 $P(r^i|\ c)$는 식 (1)의 Bayes 정리에 기반을 두고 계산하였다.

$$P(r^i|c) = \frac{P(c|r^i)P(r^i)}{\sum_j P(r^j)P(c|r^j)} \quad (1)$$

Ⅴ. 수취인 주소 영역 추출

우편 봉투 영상의 연결 요소로부터 수취인 주소 영역(Destination Address Block: DAB)을 추출하기 위하여 IV 장에서 기술한 지식 베이스를 이용한다. 우리나라 우편 봉투에서의 주소 영역들은 필기자의 습관에 따라 서로 근접한 위치에 필기되기 때문에, 기존의 연구에서와 같이 우편 봉투 영상을 수취인 주소 후보 영역으로 분할하기 어려울 뿐만 아니라 분할된 수취인 주소 후보 영역에 발신인 주소나 우표 및 소인 영역이 포함될 가능성이 상당히 높다. 따라서 본 연구에서는 수취인 주소 후보 영역을 설정하지 않고 다양한 우편 봉투의 통계적인 특징 분석을 통하여 얻은 지식을 이용하여 연결 요소를 병합, 분할하고 검증하는 과정을 반복함으로써 우편 봉투 영상에서 수취인 주소 영역을 추출하고, 수취인 주소, 성명, 우편번호 부분을 추출하고자 한다. 그림 4는 수취인 주소 영역 추출 알고리즘의 흐름도를 나타내며, 수취인 주소 영역과 수취인 주소 영역의 각 부분을 추출하는 알고리즘은 다음과 같다. 먼저, 연결 요소의 평균 폭과 높이를 매개변수로 하여 연결 요소를 단어 단위의 영역으로 군집화한 다음(단계 2), 각 단어 단위의 영역에 대해 지식 베이스를 이용하여 DAB를 설정한다.(단계 3) 이렇게 설정된 DAB에는 실제 DAB가 아닌데도 DAB에 포함된 영역이 있을 수 있으며 그 반대인 경우도 발생할 수 있기 때문에 DAB로 설정된 영역을 다음과 같이 검증한다.(단계 4 ~ 6) 먼저, DAB로 설정된 영역에 있는 단어 단위 영역의 평균 높이를 매개변수로 하여 DAB로 설정된 영역을 문자열 단위로 군집화한 다음(단계 4), 각 문자열 단위의 영역에 대해 지식 베이스를 이용하여 DAB의 각 부분을 설정한다.(단계 5) 이때 DAB의 각 부분으로 설정된 영역이 DAB의 각 부분일 확률이 충분히 크다면(단계 6) 현재 설정된 DAB를 최종적인 DAB로 결정하고 수취인 주소, 성명, 우편번호 부분을 추출한다.(단계 11) 만약 그렇지 않다면 조건을 검사하여 현재 설정된 DAB에 연결 요소를 병합하거나(단계 9) 분할하고(단계 10), 검증하는 과정(단계 4 ~ 6)을 반복함으로써 최종적인 DAB를 추출하고 수취인 주소, 성명, 우편번호 부분을 추출하며, 만약 위의 반복 과정이 임계값 이상 진행되면 알고리즘은 기각한다.(단계 8)

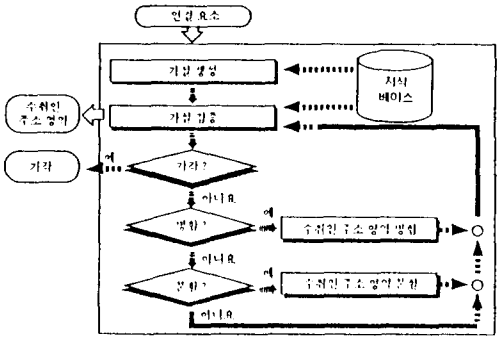


그림 4. 수취인 주소 영역 추출 알고리즘의 흐름도.
Fig. 4. Flowdiagram of the algorithm for locating DAB.

· 수취인 주소 영역과 수취인 주소, 성명, 우편번호 부분을 추출하는 알고리즘
입력: 우편 봉투 영상의 연결 요소
출력: 수취인 주소 영역과 수취인 주소, 성명, 우편번호 부분
표현 기호:
우편 봉투 ER과 수취인 주소 영역 DAB는 다음과 같이 표현된다.
ER = {DAB, RAB, PSB, GEB}
여기에서 DAB, RAB, PSB, GEB는 각각 수취인 주소, 발신인 주소, 우표 및 소인, 그림 및 기타 영역을 나타낸다.
DAB = {AP, NP, ZP}
여기에서 AP, NP, ZP는 각각 수취인 주소, 성명, 우편번호 부분을 나타낸다.
DAB(t): t 번째 반복 단계에서 설정된 수취인 주소 영역
$DAB_{AP}^{(t)}$, $DAB_{NP}^{(t)}$, $DAB_{ZP}^{(t)}$: t 번째 반복 단계의 설정된 수취인의 주소, 성명 및 우편 번호 부분
WB: 단어 단위의 영역
LB: 문자열 단위의 영역
$P(\cdot)$: $(\cdot)$이 만족될 확률
T_1, T_2, T_3, T_4, T_5, α_M, α_S: 임계값

1) t = 1로 초기화한다.
2) 연결 요소의 평균 폭과 높이를 매개변수로 하여 연결 요소를 WB로 군집화한다. WB의 갯수를

m이라 하면, WB = {WB_i|i = 1, 2, ..., m}으로 표현할 수 있다.

3) WB_i(i = 1, 2, ..., m)에 대해 식 (2)와 같이 수취인 주소, 발신인 주소, 우표 및 소인, 그림 및 기타 영역 중의 한영역으로 할당한 다음, 식 (4)와 같이 WB_i가 수취인 주소 영역일 확률이 각 영역일 확률 중에서 가장 크고 임계값 T_1보다 큰 WB_i를 수취인 주소 영역 $DAB^{(t)}$로 설정한다.

$$WB_i = \underset{ER}{\text{argmax}} \ [P(WB_i = ER)] \qquad (2)$$

위의 식에서 $P(WB_i = ER)$은 지식 베이스를 이용하여 식 (3)과 같이 4절에서 기술한 우편 봉투의 각 영역에 대한 특징들을 비교함으로써 계산한 확률이다. 즉, WB_i가 ER일 확률은 WB_i의 각 특징과 ER의 각 특징이 같은 값을 나타낼 확률을 가중치 합한 확률이다.

$$P(WB_i = ER) = \sum_u \omega_u \bullet P(WB_i\text{의 각 특징} = \text{ER의 각 특징}) \quad (3)$$

여기서, u는 비교할 특징의 갯수이고, ω_u는 가중치이다.

$$DAB^{(t)} = \{WB_i | WB_i \in DAB \text{ and } P(WB_i = ER_{DAB}) > T1\} \quad (4)$$

여기에서 i = 1, 2, ..., m이다.

4) $DAB^{(t)}$에 속하는 WB의 평균 높이를 매개변수로 하여 $DAB^{(t)}$를 LB로 군집화한다. LB의 갯수를 n이라 하면, LB = {LB_i|i = 1, 2, ..., n}으로 표현할 수 있다.

5) LB_i(i = 1, 2, ..., n)에 대해 식 (5)와 같이 수취인 주소, 성명, 우편번호 부분 중의 한 부분으로 할당한 다음, 식 (7)과 같이 수취인 주소, 성명, 우편번호 부분을 설정한다.

$$LB_i = \underset{DAB}{\arg\max}[P(LB_i = DAB)] \qquad (5)$$

위의 식에서 $P(LB_i = DAB)$는 지식 베이스를 이용하여 식 (6)과 같이 4절에서 기술한 수취인 주소 영역의 각 부분에 대한 특징들을 비교함으로써 계산한 확률이다. 즉, LB_i가 DAB일 확률은 LB_i의 각 특징과 DAB의 각 특징이 같은 값을 나타낼 확률을 가중치 합한 확률이다.

$$P(LB_i = DAB) = \sum_v \omega_v \bullet P(LB_i\text{의 각 특징} = DAB\text{의 각 특징}) \quad (6)$$

여기서, v는 비교할 특징의 갯수이고, ω_n는 가중치이다.

$$DAB_{AP}^{(t)} = \{LB_i | LB_i \in AP\}$$
$$DAB_{NP}^{(t)} = \{LB_i | LB_i \in NP\}$$
$$DAB_{ZP}^{(t)} = \{LB_i | LB_i \in ZP\} \qquad (7)$$

여기서, i = 1, 2, ..., n이다.

6) IF 현재 설정된 수취인 주소 영역의 각 부분에 대해 식 (8)이 만족되어, t번째 반복 단계에서 설정된 수취인 주소 영역의 각 부분이 실제로 수취인 주소 영역의 각 부분일 확률이 임계값 T2보다 크면

$$P(DAP_{AP}^{(t)} = AP) > T_2 \text{ and}$$
$$P(DAP_{NP}^{(t)} = NP) > T_2 \text{ and}$$
$$P(DAP_{ZP}^{(t)} = ZP) > T_2 \qquad (8)$$

THEN GOTO 단계 11)

7) t = t + 1.

8) IF t 〉 T3 THEN 기각하고 STOP.

9) IF 식 (9)와 같은 병합 조건을 만족하는 MB가 존재하여, WB_i가 (t-1)번째 반복 단계에서 수취인 주소 영역으로 설정되지 않았고, WB_i가 우편 봉투 영상의 각 영역일 확률 중에서 가장 큰 확률과 WB_i가 수취인 주소 영역일 확률과의 차가 임계값 T_4보다 작은 WB_i가 존재하면

$$MB = \{WB_i | WB_i \notin DAB^{(t-1)} \text{ and } [\max_{ER} P(WB_i = ER) - P(WB_i = DAB)] < T_4\} \quad (9)$$

여기서, i = 1, 2, ..., m이다.

THEN

BEGIN

$DAB^{(t)} = DAB^{(t-1)} \cup MB$ /* 수취인 주소 영역의 연결 요소 병합 */

$T_4 = T_4 + \alpha_M$ /* 병합 임계값 증가 */

단계 4) ~ 6)의 검증 과정을 수행한다.

END

10) IF 식 (10)과 같은 분할 조건을 만족하는 SB가 존재하여, WB_i가 (t-1)번째 반복 단계에서 수취인 주소 영역으로 설정되어 있고, WB_i가 수취인 주소 영역일 확률과 WB_i가 우편 봉투 영상의 각 영역일 확률 중에서 가장 큰 확률과의

차가 임계값 T_5보다 작은 WB_i가 존재하면

$$SB = \{WB_i | WB_i \in DAB^{(t-1)} \text{and} \\ [P(WB_i = DAB) - \max_{ER} P(WB_i = ER] < T_5\} \quad (10)$$

여기서, i = 1, 2, ..., m이다.

THEN

BEGIN

$DAB^{(t)} = DAB^{(t-1)}$ - SB /* 수취인 주소 영역의 연결 요소 분할 */

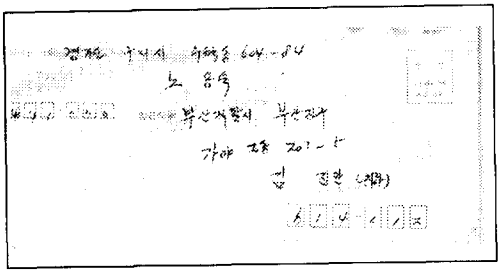
$T_5 = T_5 + \alpha_s$ /* 분할 임계값 증가 */

단계 4) ~ 6)의 검증 과정을 수행한다.

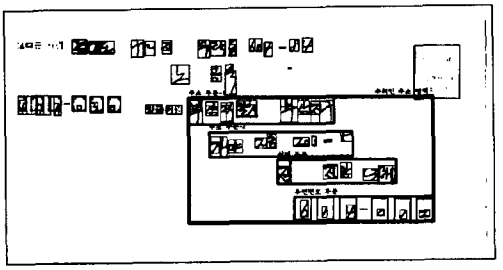
END

11) 현재 설정된 $DAB^{(t)}$를 최종적인 수취인 주소 영역으로 추출하고, 식 (7)에서 설정된 $DAB_{AP}^{(t)}$, $DAB_{Np}^{(t)}$, $DAB_{zp}^{(t)}$를 최종적인 수취인 주소, 성명, 우편번호 부분으로 추출한다.

위의 알고리즘에서 사용된 임계값은 다양한 실험을 통하여 최적의 값으로 결정하였으며,그림 5와 6은 제안된 방법에 의해 추출된 수취인 주소 영역과 수취인 주소 영역의 각 부분을 나타낸다.



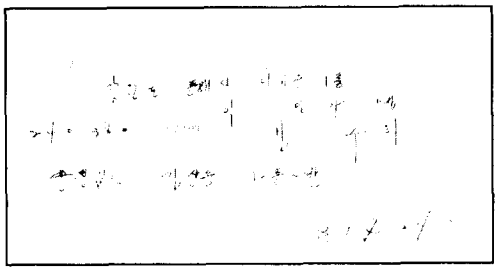
(a)



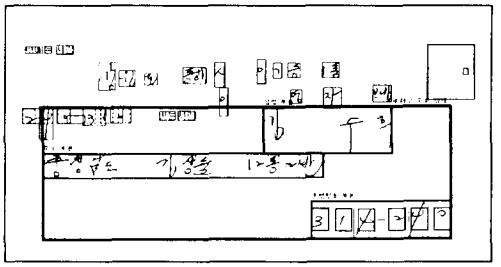
(b)

그림 5. 수취인 주소 영역 추출의 예
(a) 우편 봉투의 명도 영상 (b) 수취인 주소 영역과 수취인 주소, 성명, 우편 번호 부분이 추출된 우편 봉투 영상

Fig. 5. An example of extracting destination address block
(a) gray scale image of an envelope
(b) envelope image in which destination address block and its address, name and zipcode parts are extracted.



(a)



(b)

그림 6. 수취인 주소 영역 추출의 예
(a) 우편 봉투의 명도 영상
(b) 수취인 주소 영역과 수취인 주소, 성명, 우편 번호 부분이 추출된 우편 봉투 영상

Fig. 6. An example of extracting destination address block
(a) gray scale image of an envelope
(b) envelope image in which destination address block and its address, name and zipcode parts are extracted.

Ⅵ. 실험 및 결과 분석

실험 환경은 SPARC-2 워크스테이션의 X-Window 시스템 상에서 C 언어로 구현하였으며, 체신부 산하 서울 우편집중국으로부터 제공받은 1,000장의 실제 우편 봉투를 Microtek ScanMaker 600GS 영상 스캐너를 통해 200 DPI로 입력하여 실험하였다. 우편 봉투 영상에서 통계적인 지식을 습득하기 위하여 500 장의 우편 봉투를 사용하였고, 나머지 500 장의 우편 봉투는 제안된 시스템의 성능을 평가하기 위하여 사용하였다.

표 1은 제안된 방법에서 반복 과정을 수행하지 않은 경우와 수행한 경우의 수취인 주소 영역에 대한 추출률을 나타내며, 표 2는 올바르게 추출된 수취인 주소 영역에 대한 수취인 주소, 성명, 우편번호 부분의 추출률을 나타낸다. 수취인 주소 영역이나 수취인 주소 영역의 각 부분들에 대한 추출 요류가 발생한 우편 봉투들을 분석해 보면 대부분이 잘못된 이진화로 인해 정확하게 연결 요소를 추출하지 못했기 때문인 것으로 판명되었다. 특히, 배경이나 주소 영역이 특정한 색을 갖는 영상에 대해서는 적응적 이진화를 적용하여도 잡영이 발생하기 때문에 정확한 수취인 주소 영역을 추출하지 못하였으며, 우편 봉투 영상에 불필요한 문구나 그림이 포함된 경우에는 불필요한 문구나 그림이 수취인 주소 영역에 포함되는 경우도 발생하였다. 또한 비정상적인 위치에 필기된 주소 영역으로 인해 오류가 발생하였다. 표 3은 우편 봉투 영상 1장당 평균 처리 시간을 나타낸다.

표 1. 수취인 주소 영역에 대한 추출률
Table 1. Extraction rate for destination address block.

처리율 \ 처리 방법		비반복 처리	반복 처리
추출률		68.3 %	85.7 %
오류율	실제 영역보다 크게 추출	19.5 %	3.5 %
	실제 영역보다 작게 추출	12.2 %	2.9 %
기각률		0 %	7.9 %

표 2. 수취인 주소 영역의 주소, 성명, 우편번호 부분에 대한 추출률
Table 2. Extraction rate for address, name and zipcode parts in destination address block.

주소 부분	94.4 %
성명 부분	83.3 %
우편번호 부분	85.6 %

표 3. 우편 봉투 영상 1장당 평균 처리 시간
Table 3. Average processing time per one envelope image.

구 분	우편 봉투 스캐닝	적응적 이진화	연결된 요소 추출	기울어진 영상 교정	수취인 주소 영역 추출	계
처리 시간	2.10초	1.06초	0.38초	0.32초	0.04초	3.90초

Ⅶ. 결 론

본 논문에서는 우리나라 우편 봉투 영상의 구조를 분석하여 수취인 주소 영역을 자동적으로 추출하는 지식 기반 시스템을 제안한다. 제안된 시스템은 우편 봉투 영상의 효과적인 구조 분석을 위하여 적응적 이진화, 연결 요소 추출, 기울어진 영상 교정 등의 전처리를 수행한 다음, 다양한 우편 봉투에 대한 통계적인 특징 분석을 통하여 얻은 지식에 기반을 둔 반복적인 처리 과정에 의해 우편 봉투 영상에서 수취인 주소 영역을 추출하고, 수취인 주소, 성명, 우편번호 부분을 추출한다. 현재까지 발표된 수취인 주소 영역을 추출하는 시스템의 대부분은 우편봉투 영상을 여러개의 수취인 주소 후보 영역으로 분할한 다음 후보 영역 중에서 하나의 주소 영역을 선택하는 방식으로 구성되어있다. 그러나, 우리나라 우편 봉투에서의 주소 영역들은 서로 근접한 위치에 필기되기 때문에 본 연구에서는 이러한 분할과 선택의 과정을 처리하지 않고 연결 요소를 병합, 분할하고 검증하는 과정을 반복함으로써 수취인 주소 영역을 추출하는 반복적인 과정을 제안한다. 다량의 실제 우편 봉투에 대해 실험한 결과, 제안된 방법이 우리나라 우편 봉투 영상에서 수취인 주소 영역 추출에 매우 효과적임을 알 수 있었다.

앞으로의 연구 방향은 이진화로 인한 잡영의 발생을 최소화하기 위하여 이진화하지 않고 명도 영상에서 직접 수취인 주소 영역 추출에 관한 연구가 이루어져야 하며, 배경이나 주소 영역이 특정한 색을 갖을 경우 문자 부분만을 강조하는 영상 향상 기법에 관한 연구가 이루어져야 할 것이다. 또한, 수취인 주소 영역이 결정된 후 인식기가 문자를 인식할 수 있

도록 주소나 우편번호 부분을 정확하게 문자 단위로 분할하는 연구[11]가 이루어져야 하며, 인식에 있어서도 후처리에 기반을 둔 주소 및 성명 인식에 관한 연구[12]와 안정된 숫자 인식 알고리즘의 개발에 관한 연구가 지속적으로 이루어져야 할 것이다. 궁극적으로는 규격 봉투 뿐만 아니라 임의의 우편물에서 수취인 주소 영역을 추출하는 연구가 이루어져야 할 것으로 사료된다.

參 考 文 獻

[1] Y.-C. Shin, R. Sridhar, V. Demjanenko, P. W. Palumbo and J. Hull, "Contrast Enhancement of Mail Piece Images," Proc. of SPIE Conf. on Machine Vision Applications in Character Recognition and Industrial Inspection, Vol. 1661, San Jose, USA, February 1992, pp. 27-37.

[2] S. N. Srihari, C. H. Wang, P. W. Palumbo and J. J. Hull, "Recognizing Address Blocks on Mail Pieces: Specialized Tools and Problem-solving Architecture," AI Magazine, Vol. 8, No. 4, 1987, pp. 25-40.

[3] L. A. Fletcher and R. K. Kasturi, "A Robust Algorithm for Text String Separation from Mixed Text/Graphics Images," IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 10, No. 6, 1988, pp. 910-918.

[4] S. C. Hinds, J. L. Fisher and D. P. D'Amato, "A Document Skew Detection Method Using Run-length Encoding and the Hough Transform," Proc. of 10th Int. Conf. on Pattern Recognition, Atlantic City, USA, June 1990, pp. 464-468.

[5] A. K. Jain and S. K. Bhattacharjee, "Address Block Location on Envelopes Using Gabor Filters," Pattern Recognition, Vol. 25, No. 12, 1992, pp. 1459-1477.

[6] Y. Y. Tang, C. Y. Suen, C. D. Yan and M. Cheriet, "Document Analysis and Understanding: A Brief Survey," Proc. of Int. Conf. on Document Analysis and Recognition, Vol. 1, Saint-Malo, France, September 1991, pp. 17-31.

[7] A. C. Downton and C. G. Leedham, "Pre-processing of Envelope Images for Optical Character Recognition," Proc. of 9th Int. Conf. on Pattern Recognition, Rome, Italy, November 1988, pp. 27-31.

[8] J.-C. Oriot, D. Barba and J.-C. Salome, "Address Block Location Method Based on Transition Analysis Approach: Design and Evaluation on Flats Objects," Proc. of Int. Conf. on Document Analysis and Recognition, Vol. 2, Saint-Malo, France, September 1991, pp. 665-673.

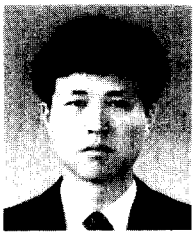
[9] N. Bartneck, "Knowledge Based Address Block Finding Using Hybrid Knowledge Representation Schemes," Proc. of the 3rd USPS Advanced Technology Conf., May 1988, pp. 249-263.

[10] J. Illingworth and J. Kittler, "A Survey of the Hough Transform," Computer Vision, Graphics and Image Processing, Vol. 44, 1988, pp. 87-116.

[11] Y. Kobayashi, K. Yanada and J. Tsukumo, "A Segmentation Method for Hand-written Japanese Character Lines Based on Transitional Information," Proc. of 11th Int. Conf. on Pattern Recognition, The Hague, The Netherlands, August 1992, pp. 487-491.

[12] 이 성환, 김 은순, "주소 및 성명에서의 한글 인식을 위한 효율적인 오인식 교정 알고리즘," 한국 정보 과학회 논문지, 제 20권 제 5호, 1993년 5월, pp. 729-738.

著 者 紹 介



金 基 詰(學生會員)
1968年 12月 26日生. 1992年 충북대학교 전자계산학과 학사. 1994年 충북대학교 컴퓨터과학과 석사. 주관심 분야는 문서구조 분석, 필기체 문자인식 등임.

李 晟 瑍(正會員) 第 31卷 B編 第 4號 參照
충북대학교 컴퓨터과학과 조교수