인쇄 및 필기 한글 우편영상에서의 수취인 주소 영역 추출 방법

(Destination Address Block Location on Machine-printed and Handwritten Korean Mail Piece Images)

정 선 화 † 장 승 익 †† 임 길 택 †† 남 윤 석 †
(Seon-Hwa Jeong) (SeungIck Jang) (Kil-Taek Lim) (Yun-Seok Nam)

요 약 본 논문에서는 우리나라 우편영상에서 수취인 주소 영역을 추출하는 방법을 제안한다. 우편영상에 기입된 주소가 프린터나 타자기 등에 의해서 인쇄된 주소일 수도 있고 사람에 의해서 필기된 주소일 수도 있다. 즉, 인쇄체 우편영상과 필기체 우편영상 모두에 적용될 수 있는 수취인 주소 영역 추출 방법을 제안한다. 제안 방법에서는 이진화된 우편영상으로부터 연결요소를 추출하고 연결요소를 결합하여 문자열을 생성한다. 그 후 문자열을 군집화하고 생성된 군집 중 몇 개의 군집을 선택함으로써 수취인 주소 영역을 결정한다. 우리나라 우편봉투에 기입되는 정보의 유형별 기입 위치 패턴에 따라 우편영상을 총 9개의 균등 영역으로 분할한 후 각 영역의 중심을 초기값으로 갖는 9개의 군집을 생성하였고 k-Means 방법을 사용하여 군집화를 수행하였다. 군집화 과정에서 사용되는 거리함수로 우편영상의 폭 대 높이의 비율이 반영된 수정된 맨하탄 거리를 사용하였다. 제안 방법의 성능을 알아보기 위하여 실제 우편물 영상 1,988개를 사용하여 실험한 결과 약 93.56%의 우편영상에서 수취인 주소 영역을 정확하게 추출할 수 있었다.

키워드 : 인쇄체 주소, 필기체 주소, 수취인 주소 영역 추출, 주소 인식 시스템

Abstract In this paper, we propose an efficient method for locating destination address block on both of machine-printed and handwritten Korean mail piece images. The proposed method extracts connected components from the binary mail piece image, generates text lines by merging them, and then groups the text lines into nine clusters. The destination address block is determined by selecting some clusters. Considering the geometric characteristics of address information on Korean mail piece, we split a mail piece image into nine areas with an equal size. The nine clusters are initialized with the center coordinate of each area. A modified Manhattan distance function is used to compute the distance between text lines and clusters. We modified the distance function on which the aspect ratio of mail piece could be reflected. The experiment done with live Korean mail piece images has demonstrated the superiority of the proposed method. The success rate for 1,988 testing images was about 93.56%.

Key words : Machine-printed address, Handwritten address, Destination address block location, Address reading system

1. 서 론

우리나라 우편물 처리업무는 그림 1에 제시한 바와 같이 크게 수집, 선별·정리·소인, 발송구분, 도착구분, 순로구분, 배달업무 등으로 나눌 수 있다. 여기서, 순로구분이란 집중국에서 발송 및 도착 구분 과정을 거쳐 집배국으로 운송된 우편물을 각 집배원별로 구분하고 이를 다시 배달 순서대로 정렬하는 작업으로서, 현재 우리나라는 우편집중국을 건설하고 우편물 자동구분기를 도입하여 우편물의 발송구분 및 도착구분 작업을 자동화하고 있으나 순로구분 작업은 아직까지 수작업에 의존하고 있는 실정이다[1].

순로구분의 자동화를 위해서는 수취인 주소를 자동으로 인식하고 인식된 결과를 코드화된 배달지 정보로 변

† 비 회 원 : 한국전자통신연구원 연구원
sh-jeong@etri.re.kr
ysnam@etri.re.kr
†† 정 회 원 : 한국전자통신연구원 연구원
sijang@etri.re.kr
ktlim@etri.re.kr
논문접수 : 2003년 6월 14일
심사완료 : 2003년 10월 15일

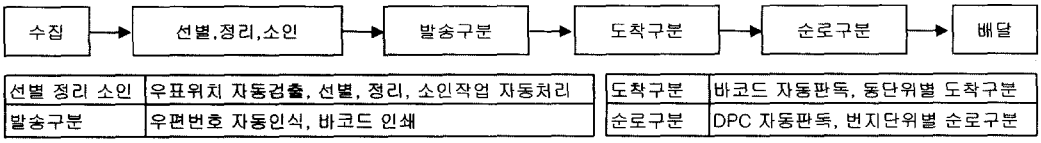


그림 1 우편물 처리 과정

환한 뒤 미리 정해진 배달 순서에 따라 우편물을 정렬하는 과정을 거쳐야 한다. 이러한 일련의 순로구분 자동화 처리 과정에서 가장 핵심이 되는 기술 중에 하나가 주소 인식 기술이다. 주소 인식 기술은 주로 문자 인식과 관련되어 연구[2-7]되어 왔으나, 주소 인식의 성능 향상을 위해서는 인식 기술뿐만 아니라 인식 전에 수행되는 영상 전처리 기술 또한 매우 중요하다. 주소 인식에서 영상 전처리는 인식하고자 하는 주소 영역 추출, 문자열 분리 및 문자 분리를 포함한다. 우편영상에 주소 인식 기술이 적용되는 경우 인식하고자 하는 주소 영역 추출 문제는 수취인 주소 영역 추출 문제가 된다.

본 논문에서는 순로구분의 주 대상인 일반 소형 통상 우편영상에서 수취인 주소 영역을 추출하는 방법에 관하여 기술하였다. 우편봉투에 기입되는 정보의 유형은 나라마다 다양하다. 우리나라의 경우 일반 소형 통상 우편봉투 앞면에는 그림 2에서 볼 수 있듯이 발신인 주소, 수취인 주소, 우표 및 소인, 광고성 로고 및 문구, 바코드 등이 존재한다. 이들을 문자 영역과 비문자 영역으로 분류하면 발신인 주소와 수취인 주소 그리고 광고성 문구 등이 기입된 영역을 문자 영역으로 분류할 수 있고, 우표 및 소인, 광고성 로고 그리고 바코드 등이 부착되거나 인쇄된 영역을 비문자 영역으로 분류할 수 있다. 수취인 주소는 문자 영역에 포함되므로 수취인 주소 영역 추출 문제를 개념상 우편영상으로부터 문자 영역을 추출하는 문제와 문자 영역에서 수취인 주소 영역을 결정하는 문제로 나누어 볼 수 있다.

문자 영역에서 수취인 주소 영역을 결정하는 문제는 우편주소 기재 요령을 참고할 수밖에 없다. 발신인 주소와 수취인 주소는 주소라는 공통된 특성을 가지므로 인식을 수행한다하더라도 발신인 주소와 수취인 주소를 구별하기 어렵다. 우리나라 우편주소 기재 요령에 따르면 발신인 주소는 우편봉투의 왼쪽 상단에 그리고 수취인 주소는 오른쪽 하단에 기입하도록 되어 있다. 이 정보를 이용해서 수취인 주소 영역은 우편봉투의 오른쪽 하단의 문자 영역으로부터 찾아질 수 있다. 수취인 주소 영역을 추출하는 방법에 관한 기존 연구[8-15]를 살펴보면, 대부분 우편영상으로부터 문자 영역을 추출하는 방법에 주안점을 두고 있으며 수취인 주소 영역 위치는 각 나라의 우편주소 기재 요령을 반영하여 추출된 문자 영역으로부터 결정되고 있다.

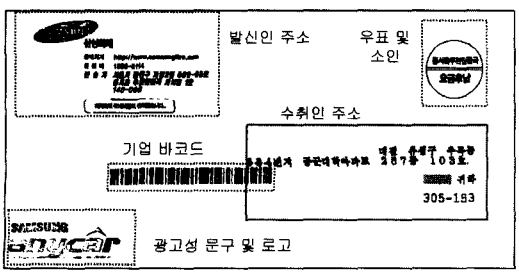


그림 2 일반 소형 통상 우편영상의 예

본 논문의 구성은 다음과 같다. 2장에서 기존 연구들의 비교 분석과 함께 제안 방법과 기존 연구의 차이점을 기술하였으며 3장에서는 제안 방법에 대한 구체적인 설명을 그리고 4장에서는 실제 우편물 영상에 대하여 측정된 제안 방법의 성능 및 오류 분석 결과를 제시하였다. 마지막으로 5장에서 결론 및 향후연구를 기술하였다.

2. 기존 연구

우편영상에서 수취인 주소 영역 추출에 관한 기존 연구는 문자 영역 추출 문제와 수취인 주소 영역 결정 문제를 분리하여 해결하는 방법과 동시에 해결하는 방법으로 분류될 수 있다[8-15]. 전자에 속한 방법들은 우편영상에서 수취인 주소 영역 후보로 정의된 영역을 몇 개 추출한 후 수취인 주소 영역에 가장 적합한 영역을 수취인 주소 영역으로 선정하였다[8,12,13,15]. 후자에 속한 방법들에서는 우편영상에서 문자 영역 추출 과정 없이 바로 수취인 주소 영역을 추출하였다[9-11,14].

Jain과 Bhattachrjee는 명도 레벨의 우편영상에서 수취인 주소 영역을 추출하는 방법을 제안하였다[8]. 제안된 방법에서는 Gabor 필터와 다층퍼셉트론을 이용하여 각 화소를 문자와 비문자로 분류하고 문자로 분류된 화소들을 연결요소들로 표현한 후 이들의 위치 정보를 이용하여 수취인 주소 영역을 추출한다. 이 방법은 영상의 크기에 비례하여 처리속도가 증가하며 수취인 주소 영역이 몇 개의 영역으로 분할될 경우 수취인 주소 영역 추출에 실패하는 단점이 있다. Yu는 이진화된 영상으로

부터 연결요소를 추출하여 수취인 주소 영역을 추출하는 방법을 제안하였다[12]. 제안된 방법에서는 영상에서 추출한 각 연결요소들이 수평 및 수직으로 가깝거나 겹칠 경우 병합하여 문자열을 생성한 후, 문자열 개수, 영역의 크기, 문자열 정렬 형태 등을 고려하여 수취인 주소 영역을 추출한다. 그러나 우리나라 우편물의 경우 주소의 문자열 개수 및 영역 크기가 일정하지 않으므로 이 방법을 우리나라 우편영상에 적용하기 어렵다. Wolf는 대형 통상 우편물에서 수취인 주소 영역을 빠르게 찾아주는 방법에 관하여 제안하였다[13]. 제안된 방법에서는 명도 레벨의 우편영상을 일정 크기의 블록으로 분할한 후 블록의 동질성에 따라서 블록들을 병합하면서 수취인 주소영역을 추출한다. 이 방법은 Yu의 방법과 마찬가지로 수취인 주소 영역의 문자열들이 떨어져서 기입되는 경우 수취인 주소영역을 정확히 추출하기 어렵다. Chanpongsaeng은 태국 고유의 우편주소 기재 특성을 이용한 수취인 주소 영역 추출 방법을 제안하였다[15]. 그러나 이 방법은 태국 우편물의 고유 특성을 사용하여 우리나라 우편영상에 적용하기 어렵다.

Lee와 Kim은 필기 한글 우편영상에서 수취인 주소 영역을 추출하는 방법을 제안하였다[9]. 제안된 방법은 이진화된 우편영상으로부터 연결요소를 추출한 후 이들을 병합하여 문자열을 형성하였다. 그 다음 각 문자열은 지식기반 확률을 기반으로 수취인 주소 영역, 발신인 주소 영역, 우표 및 소인 영역, 그래픽 영역으로 분류된다. 이 방법은 필기체 영상에 제한된 방법으로 인쇄체와 필기체 영상을 함께 처리할 수 없다. Nakajiman은 일본 우편영상에서 수취인 주소 영역을 추출하는 방법을 제안하였다[10]. 제안된 방법은 주소기입 형태의 구분에 주안점을 두고 있다. 이는 일본 우편물의 기입 형태가 총 6종이며 형태에 따라 주소 영역의 추출이 용이하기 때문이다. Wichello와 Yan은 호주 우편영상에서 수취인 주소 영역을 추출하는 방법을 제안하였다[11]. 제안된 방법에서는 문자열을 추출하고 단순히 모든 문자열을 결합함으로써 수취인 주소 영역을 추출하였다. Xue는 중국 우편영상에서 수취인 주소 영역을 추출하는 방법을 제안하였다[14]. 중국 우편물의 수취인 주소 영역은 우편영상의 상단에 위치하고 있기 때문에, 제안된 방법에서는 수취인 주소의 마지막 문자열을 찾는데 주안점을 두고 있다. 위와 같은 방법들은 그 나라 고유의 우편주소 기재 특성을 이용하여 수취인 주소 영역을 추출하였기 때문에 우리나라 우편영상에서의 수취인 주소 영역 추출에 직접 적용되기 어렵다.

기존 연구는 입력 우편영상의 종류 - 인쇄체 또는 필기체 - 에 따라 세 분류로 나뉠 수 있다. 즉, 인쇄체 우편영상에서 수취인 주소 영역을 추출하는 방법[10], 필기체 우편영상에서 수취인 주소 영역을 추출하는 방법[8,9,14], 그리고 인쇄체와 필기체 우편영상 모두에 적용될 수 있는 방법[11-13,15]으로 구분될 수 있다. 본 논문에서는 세 번째 부류로 분류되는 기존 방법들처럼 인쇄체와 필기체 우편영상 모두에 적용될 수 있는 방법을 제안하였다. 인쇄체 우편영상과 필기체 우편영상을 구분하지 않고 입력으로 제공받는 주소 인식 시스템에서 수취인 주소 영역 추출 방법은 인쇄체 우편영상과 필기체 우편영상에 동일하게 적용될 수밖에 없다. 그 이유는 우편영상의 종류는 수취인 주소가 인쇄체인지 또는 필기체인지에 따라 결정되므로 수취인 주소 영역 추출 이전에 인쇄체 우편영상과 필기체 우편영상을 구분할 수 없기 때문이다. 따라서 제안 방법은 인쇄체와 필기체 우편영상을 모두 입력으로 제공받는 주소 인식 시스템에서 효과적으로 활용될 수 있다.

제안 방법은 기존 방법들[8-10,12,14,15]과 유사하게 연결요소를 기반으로 문자열을 생성한 후 이들을 결합하여 문자 영역을 추출하고자 하였다. 따라서 문자 영역 추출 후 수취인 주소 영역을 결정하는 방법의 부류로 분류될 수 있다. 그러나 수취인 주소 영역과 대응되는 하나의 영역을 찾으려고 했던 기존 방법들과는 달리 수취인 주소 영역이 여러 개의 영역으로 분할될 수 있다고 가정함으로써 기존 방법들의 한계를 극복하였다.

3. 수취인 주소 영역 추출

그림 3은 본 논문에서 제안된 수취인 주소 영역 추출 방법의 수행 순서를 보여준다. 먼저, 200dpi의 해상도에서 명도값으로 스캔된 우편영상에 대한 전처리를 수행한다. 전처리 단계는 영상의 이진화와 이진화된 우편영상으로부터의 연결요소 추출을 포함한다. 문자열 생성 단계에서는 수평 및 수직 거리를 기반으로 인접한 연결요소를 병합한다. 이때, 생성된 문자열은 우편영상에 기입된 문자열과 일대일 대응될 필요는 없다. 예를 들면 문자열 생성 과정에서 우편영상에 기입된 임의의 문자

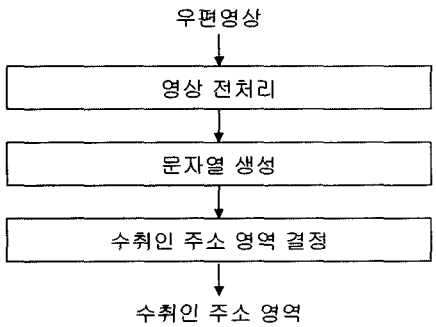


그림 3 수취인 주소 영역 추출 과정

열이 수직 또는 수평방향으로 분리되어 몇 개의 문자열로 생성될 수 있다. 반대로 두개의 문자열이 수직방향으로 겹쳐져 있는 경우 두 문자열이 하나의 문자열로 생성될 수도 있다. 다만 서로 다른 영역에 속하는 문자열이 하나의 문자열로 추출되지 않도록 주의해야한다. 수취인 주소 영역 결정 단계에서는 문자열들을 몇 개의 군집으로 군집화한다. 군집화 후 수취인 주소 영역과 대응되는 군집에 포함된 문자열을 병합함으로써 수취인 주소 영역이 결정된다.

제안 방법은 인쇄체 우편영상에서 수취인 주소 영역 추출의 오류를 최소화하면서 필기체 우편영상의 오류를 크게 발생시키지 않는데 주안점을 두고 개발되었다. 하루에 처리되는 우리나라 소형 통상 우편물 중 인쇄체 우편물이 90% 이상을 차지한다. 이 때문에 주소 인식 시스템의 전체 성능 향상을 위해서는 인쇄체 우편영상의 처리 성공률을 높여야 하기 때문이다.

3.1 영상 전처리

영상 전처리 단계에서는 영상의 이진화 및 축소, 연결요소 추출 및 잡영제거, 기업 바코드 제거 등이 수행된다. 영상의 이진화 및 축소 그리고 연결요소 추출은 인쇄체와 필기체 영상 모두에 동일하게 적용될 수 있는 방법이다. 잡영은 세 가지 유형 - salt and pepper 잡영, 영상의 외곽선에서 발생하는 잡영, 선 성분 잡영 - 으로 분류되어 제거되는데, 첫 번째와 두 번째 유형의 잡영은 인쇄체와 필기체 영상에서 비슷한 패턴으로 발생하는 잡영이므로 동일한 잡영제거 알고리즘이 적용가능하다. 세 번째 유형의 잡영은 창봉투나 주소레이블이 부착된 우편영상에서 발생되는 잡영으로 인쇄체 영상에서만 발생한다. 마찬가지로 기업 바코드 또한 인쇄체 영상에서만 발생한다. 따라서 세 번째 유형의 잡영제거 알고리즘과 기업 바코드 제거 알고리즘은 인쇄체 영상에서 수행될 때 효력이 발생하며 필기체 영상에서는 그 효력이 거의 없다.

영상의 이진화는 전역적 이진화 방법인 Otsu의 방법[16]을 사용하여 이루어졌다. Otsu의 전역적 이진화 방법은 이진화 임계치로 영상의 모든 화소를 흰 화소와 검은 화소로 구분하였을 때 평균 제곱 오차가 최대가 되도록 하는 명도값 k^*을 선택하는 방법으로, k^*는

$$k^* = \arg\max_k \xi^2(k) = \omega_0(k)\omega_1(k)[\mu_0(k)-\mu_1(k)]^2$$

와 같이 계산된다. 여기서, $\mu_0(k)$와 $\mu_1(k)$는 이진화 임계치로 k를 선택하였을 때의 흰 화소와 검은 화소 각각의 영역에서 얻어진 평균이며 $w_0(k)$와 $w_1(k)$는 분산이다.

우편영상을 이진화한 후 다음 단계는 영상 축소단계이다. 우편영상에서 대략적인 문자열을 생성하기 위해서

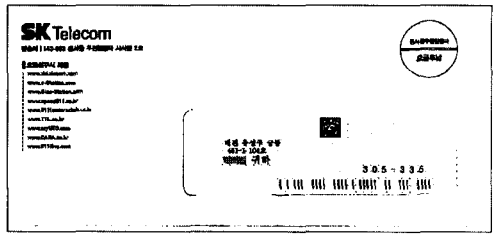
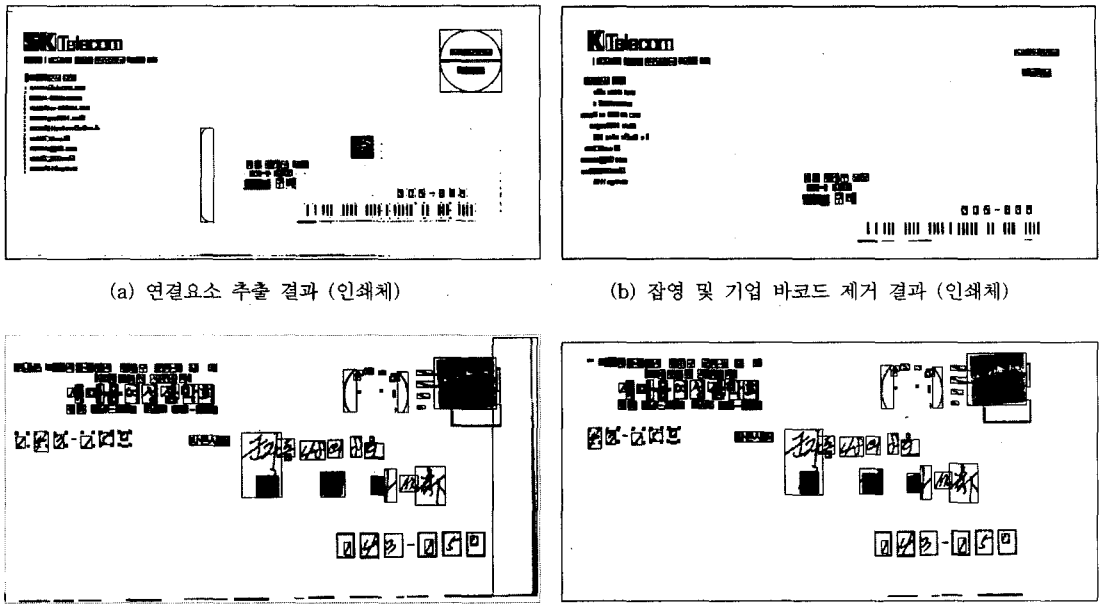


그림 4 이진화 및 1/4로 축소된 영상의 예

200dpi 해상도의 영상이 필요하지는 않다. 제안 방법에서는 처리속도 향상을 위하여 우편영상을 1/4로 축소하였다. 영상 축소 방법은 영상의 왼쪽 상단부터 2×2 윈도우를 영상에 씌우고 윈도우 내에 검은 화소의 개수가 하나라도 있으면 4개의 화소를 하나의 검은 화소로 변환하는 방법이다. 그림 4는 이진화 및 영상의 축소가 이루어진 영상의 예를 보여주고 있다.

영상 축소 후 다음 단계에서는 축소된 영상으로부터 8 방향 연결성을 가지는 연결요소를 추출한다. 그리고 그 다음 단계에서는 연결요소의 최소 인접 사각형의 크기 및 밀도 정보를 이용하여 잡영이라고 간주되는 연결요소들을 휴리스틱 규칙 기반으로 제거한다. 제안 방법에서는 잡영을 3가지 유형으로 나누어 제거하였다. 첫 번째 유형의 잡영은 salt and pepper 잡영으로 간주되는 잡영이다. 10개 화소 이하로 구성된 연결요소를 제거하였다. 두 번째 유형의 잡영은 그림 4에서 볼 수 있듯이 영상의 외곽선에서 발생하는 잡영이다. 수취인 주소 영역이 우편영상의 오른쪽 하단에 존재함을 고려하여 우편영상의 좌측 끝 5% 그리고 우편영상의 상하우측 끝 2.5% 이내에 존재하는 연결요소를 제거하였다. 세 번째 유형의 잡영은 선으로 분류되는 잡영이다. 우편영상에서 발생하는 선 성분은 대부분 창이나 주소레이블의 경계에서 발생한다. 선 성분을 나타내는 연결요소들은 크기에 비해 연결요소를 구성하는 화소의 수가 적다. 이 특성을 이용하여 연결요소 중에서 폭이나 높이가 100 보다 크면서 동시에 화소의 개수가 (폭×4+높이×4) 보다 작은 연결요소를 제거하였다.

잡영제거 단계 후 바코드가 제거된다. 우편영상에 존재하는 바코드는 기업 바코드와 우편번호 바코드로 나누어볼 수 있다. 기업 바코드는 우편물 발송업체에서 배달정보와 관련없는 정보를 바코드로 표시한 것이며 우편번호 바코드는 우편번호를 바코드로 표시한 것이다. 따라서 주소 인식 시스템 관점에서 기업 바코드는 제거의 대상이며 우편번호 바코드는 인식의 대상이다. 이 단계에서는 기업 바코드를 찾아서 제거한다. 기업 바코드는 2 차원 바코드와 폭 변조 바코드 등으로 표현되는 1 차원 바코드가 있다. 2 차원 바코드의 경우 폭과 높이의



(a) 연결요소 추출 결과 (인쇄체) (b) 잡영 및 기업 바코드 제거 결과 (인쇄체)

(c) 연결요소 추출 결과 (필기체) (d) 잡영 제거 결과 (필기체)

그림 5 영상 전처리 수행 결과의 예

차이가 2 화소 이하이며 크기가 28×28에서 40×40 사이를 가지는 연결요소를 찾은 후 연결요소의 밀도가 0.45 이상이며 왼쪽 경계와 하단 경계의 화소의 수가 (폭+높이)×0.8보다 많은 경우 2 차원 바코드로 간주하고 제거하였다. 1 차원 바코드의 경우 연결요소의 수평 히스토그램을 활용하여 바코드의 개수를 계산한다. 바코드는 25 화소 이상 바코드가 없는 부분은 20 화소 이하로 하여 몇 번의 교차가 있는지 계산하였다. 다음으로 연결요소의 밀도를 계산하고 교차 회수와 밀도에 따라 제거 여부를 결정한다. 교차 20회 이상이고 밀도가 0.4 이상이거나 교차가 25회 이상이고 밀도가 0.35 이상이거나 교차가 40회 이상이고 밀도가 0.25 이상이면 바코드로 간주하고 제거하였다.

그림 5는 영상 전처리 수행 결과를 보여주고 있다. 그림 5의 (a)와 (b)는 그림 4의 인쇄체 영상에서 연결요소 추출 결과와 잡영 및 2 차원 기업 바코드를 제거한 영상의 예이며, 그림 5의 (c)와 (d)는 필기체 영상에서 연결요소 추출 결과와 잡영제거 결과의 예이다. 앞서 언급한 바와 같이 필기체 영상의 경우 기업 바코드가 존재하지 않으며 창봉투의 창의 경계나 주소레이블의 경계에서 발생하는 긴 선 성분도 거의 존재하지 않는다. 마지막으로 영상 전처리 과정에서 사용된 각종 상수는 200dpi의 영상에서 실험적으로 좋은 성능을 보이는 값으로 결정되었다. 이러한 상수들은 입력으로 들어오는 우편영상의 해상도에 따라 달리 결정될 수 있다.

3.2 문자열 생성

문자열 생성 단계에서는 인접한 연결요소들을 결합하여 문자열을 생성한다. 앞서 언급하였듯이 문자열 생성 단계는 우편영상에 기입된 문자열과 일대일 대응되는 문자열을 추출하는 단계가 아니다. 단순히 그 결과가 문자열과 비슷한 연결요소의 결합을 시도하는 단계이다. 연결요소의 결합 결과는 하나의 문자열과 대응될 수도 있지만 하나의 문자열이 여러 개의 문자열로 분할된 결과일 수도 있고 반대로 상하 인접한 문자열이 하나로 결합된 결과일 수도 있다. 이 단계에서 충족해야할 요구사항은 서로 다른 영역에 속하는 연결요소들이 결합되지 않도록 하는 것이다. 수취인 주소 영역에 속하는 문자열과 다른 영역에 속하는 문자열이 결합되지 않아야 한다는 단순한 조건 때문에 문자열 생성 알고리즘은 인쇄체와 필기체 우편영상 모두에 동일하게 적용될 수 있다. 단지 필기체 영상의 수취인 주소 문자열들의 위치 변화가 인쇄체 영상에 비해 심하므로 동일한 성능을 기대하기는 어렵다. 그러나 4장에서 기술하는 실험결과 - 인쇄체 영상의 문자열 생성 오류는 0%이며, 필기체 영상의 문자열 생성 오류는 0.9%임 - 로 제안된 문자열 생성 알고리즘이 인쇄체와 필기체 영상 모두에 적용될 수 있음을 알 수 있다.

연결요소의 결합은 이웃한 연결요소와의 겹침 정도와 거리 정보를 기반으로 이루어진다. 이웃한 연결요소가 y축 방향으로 50% 이상 중첩되고, x축 방향으로 50 화소 이하의 거리에 있는 연결요소들을 결합하였다. 그림 6에서 보는바와 같이 y축으로 a/b가 0.5(=50%)이상이면서, x축 방향으로 50화소 이내에 위치할 경우 두 연결요소를 결합하게 된다. 연결요소의 결합은 더 이상 결합이 이루어지지 않을 때까지 재귀적으로 이루어진다. 영상 전처리와 마찬가지로 문자열 생성 단계에서는 사용된 각종 상수는 200dpi의 영상에서 실험적으로 좋은 성능을 보이는 값으로 결정되었다.

그림 7은 문자열 생성의 예를 보여준다. 그림 7의 (a)는 그림 5의 (b) 영상에서 연결요소의 결합을 시도한 결과이며, 그림 7의 (b)는 필기체 영상에서 문자열 생성의 예이다. 그림에서 볼 수 있듯이 인쇄체 영상의 경우 연결요소의 결합 결과가 하나의 문자열과 대응되는 경우가 많지만, 필기체 영상의 경우 두 개의 문자열이 결합되어 하나의 문자열로 추출되거나 또는 하나의 문자열이 여러 개의 문자열로 분리되어 추출되는 경우가 종종 발생한다.

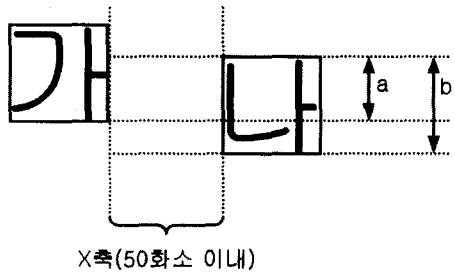
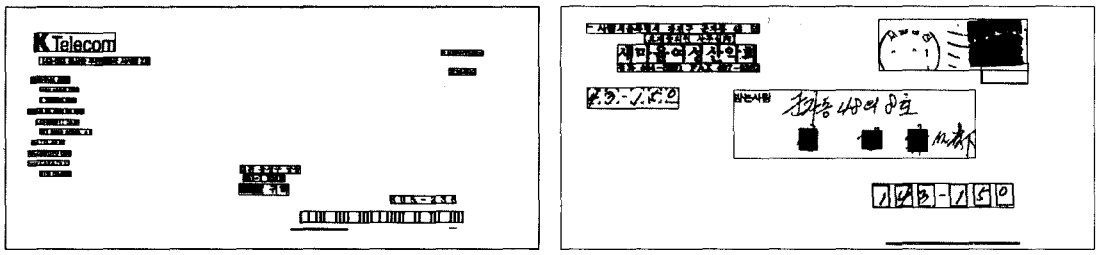


그림 6 문자열 생성 기준

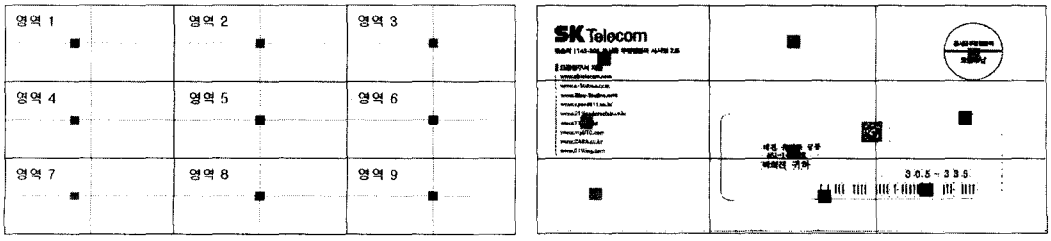
3.3 수취인 주소 영역 결정

문자열 생성 단계가 끝나면 문자열들에 대하여 군집화를 수행하고 생성된 군집들 중 수취인 주소 영역과 대응되는 군집들을 선택한다. 제안 방법에서는 우리나라 우편봉투에 기입되는 정보의 유형별 기입 위치 패턴에 따라 그림 8의 (a)처럼 우편영상을 총 9개의 영역으로 균등 분할하고 각 영역의 중심을 각 군집의 초기 중심값으로 설정하였다. 영역 1은 발신인 주소 영역과 대응될 수 있으며, 영역 3은 우표 및 소인 영역, 영역 7은 광고성 로고 및 문구 영역, 영역 5, 영역 6, 영역 8, 영역 9는 수취인 주소 영역과 대응될 수 있다. 영역 2는 발신인 주소 영역 또는 우표 및 소인 영역에 대응 될 수 있으며 영역 4는 발신인 주소 또는 광고성 로고 및 문구 영역과 대응될 수 있다. 수취인 주소 영역을 4개의 영역으로 분할한 이유는 하나의 문자열이 여러 개의 문자열로 나뉘어지는 경우 문자열 분류의 오류를 줄이기 위해서이며, 또한 군집화 과정에서 문자열의 위치 변화



(a) 인쇄체 (b) 필기체

그림 7 문자열 생성의 예



(a) 아홉 개의 영역 분할 방법 (b) 아홉 개의 영역으로 분할된 우편영상의 예

그림 8 아홉개의 군집 초기값 설정

에 덜 민감하기 위해서이다. 하나의 문자열이 여러 개의 문자열로 나뉘거나 문자열의 위치 변화는 인쇄체 영상보다는 필기체 영상에서 비교적 빈번히 발생하므로 수취인 주소 영역을 4개의 영역으로 분할하여 군집화를 수행함은 필기체 영상에서 보다 더 효과적이다.

총 9개의 군집 중 6개의 군집은 영역의 정 중앙 좌표를 초기값으로 설정하였으나, 영역 1, 영역 4, 영역 7에 해당하는 군집의 초기 중심 x좌표의 위치를 중심에서 왼쪽으로 이동시켰다. 이는 수취인 주소가 우편영상의 왼쪽에 치우쳐서 기입되었을 때 발생하는 군집화 오류를 줄이기 위해서이다.

군집화를 위해 문자열로부터 추출된 특징은 문자열의 중심좌표이다. i번째 문자열 l_i의 중심좌표를 (x_i, y_i)이라 할 때, x_i는 문자열의 최소 인접 사각형의 최대 x좌표와 최소 x좌표의 평균이며 y_i는 최대 y좌표와 최소 y좌표의 평균으로 계산된다. 군집화 방법으로 분할적 군집화 방법 중 하나인 k-Means 방법을 사용하였다. 앞서 언급했듯이 k는 9이며 각 문자열은 군집의 중심과 가장 가까운 군집으로 분류된다. 모든 문자열이 각 군집에 분류된 후 j번째 군집의 중심 $C_j(x, y)$은 다음의 식과 같이 변경된다.

$$C_j(x,y) = (C_j(x), C_j(y)) = \left(\frac{1}{n_j}\sum_{i\in C_j}^{n_j} x_i,\ \frac{1}{n_j}\sum_{i\in C_j}^{n_j} y_i\right)$$

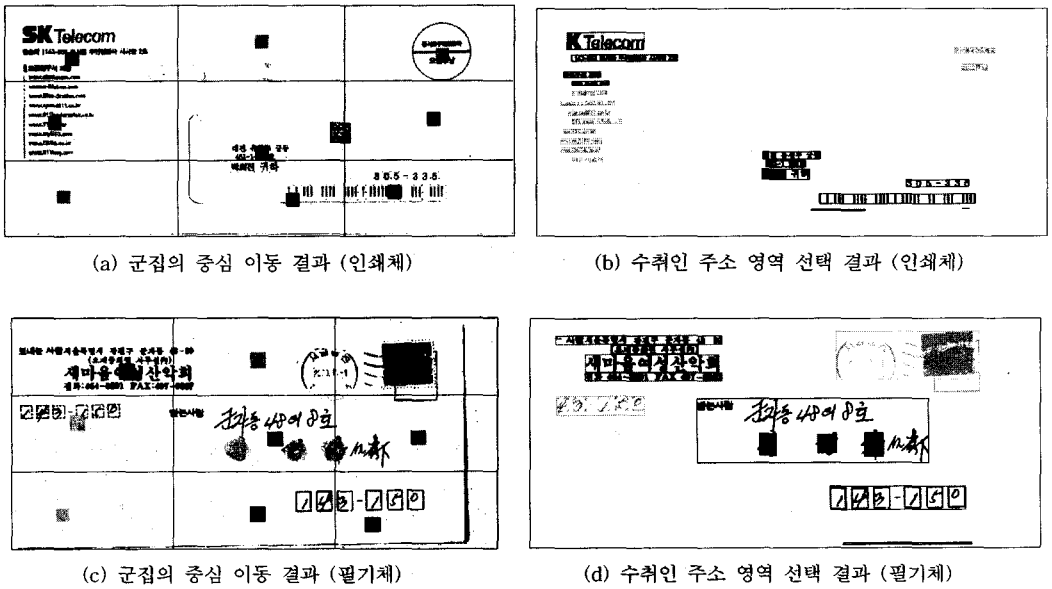
위 식에서 n_j는 j번째 군집에 분류된 문자열의 개수이다. 즉 군집의 중심은 해당 군집에 분류된 문자열들의 중심좌표의 평균으로 변경된다. i번째 문자열과 j번째 군집과의 거리는 다음의 수정된 맨하탄 거리 (Manhattan distance) $D(l_i, C_j)$를 사용하여 계산되었다.

$$D(l_i, C_j) = |x_i - C_j(x)| + |y_i - C_j(y)| \times r, \text{ 여기서}$$

$$r = \frac{\text{우편영상의 폭}}{\text{우편영상의높이}}$$

수정된 맨하탄 거리에서는 y축 방향으로의 거리에 가중치를 두어 주어진 거리보다 더 크게 계산되도록 하였다. 우편영상의 폭은 높이보다 약 2배 정도 크다. 이 때문에 문자열의 x 좌표의 분산은 y좌표의 분산보다 크다. 이는 x축 방향에서 계산된 거리와 y축 방향에서 계산된 거리가 동일할 때 y좌표 방향에서 계산된 거리가 실제적으로 더 큰 거리임을 의미한다. 이를 반영하기 위하여 제안 방법에서는 문자열과 군집간의 거리를 계산할 때 y축 방향의 거리에 가중치를 둔 수정된 맨하탄 거리를 사용하였다. 마지막으로 제안 방법에서는 군집의 중심들이 변화하지 않을 때까지 군집화 과정을 반복 수행하였다.

군집화 과정이 끝난 후 수취인 주소 영역에 해당하는 군집을 선택한다. 수취인 주소의 문자열들이 네 영역에 골고루 분포되어 있는 경우 군집 5, 군집 6, 군집 8 그리고 군집 9에 해당하는 모든 영역들을 수취인 주소 영역으로 결정함이 타당하나, 수취인 주소 문자열들이 치우쳐서 분포하는 경우 수취인 주소 영역에 해당되는 군집의 선택이 필요하다. 우편영상을 분석하여 보면 그림



(a) 군집의 중심 이동 결과 (인쇄체)

(b) 수취인 주소 영역 선택 결과 (인쇄체)

(c) 군집의 중심 이동 결과 (필기체)

(d) 수취인 주소 영역 선택 결과 (필기체)

그림 9 수취인 주소 영역 추출의 예

9에서 볼 수 있듯이 필기체 영상의 수취인 주소 영역이 인쇄체보다는 넓게 분포한다. 이 때문에 네 개의 군집 중 수취인 주소 영역의 선택은 인쇄체 영상에서는 효과적이나 필기체 영상에서는 일부 영역이 누락되어 수취인 주소 영역 추출의 실패를 가져올 수도 있다. 4장의 실험결과에서 이에 대한 분석이 기술될 것이다.

제안 방법에서는 군집 중심간의 거리 정보를 기반으로 수취인 주소 영역에 해당하는 군집을 선택하였다. 수취인 주소 영역의 선택 대상 군집은 군집 5, 군집 6, 군집 8 그리고 군집 9이다. 편의상 군집 5, 군집 6, 군집 8 그리고 군집 9를 수취인 주소 영역 후보 군집이라고 하고 다른 군집들을 비 후보 군집이라고 하자. j번째 군집과 후보내의 군집과의 최소 거리와 비 후보내의 군집들과의 최소 거리의 비율이 임계치 이상 작으면 j번째 군집은 수취인 주소 영역에 포함된다. 거리 계산을 위하여 유클리디안 (Euclidean) 거리 함수를 사용하였다. 임계치가 커지면 수취인 주소 영역의 일부가 제거될 확률이 커지며 임계치가 작아지면 수취인 주소 영역에 다른 영역의 정보가 포함될 확률이 커지게 된다. 임계치는 2.5로 하였다. 그림 9는 수취인 주소 영역 추출 결과를 보여준다. 그림 9의 (a)와 (c)의 각 영역의 점은 이동된 군집의 중심을 나타내며 (b)와 (d)에서 가장 진한 검은색 문자열들이 수취인 주소를 나타낸다.

4. 실험 및 결과

4.1 실험 데이타

제안 방법의 성능 평가를 위하여 200dpi 해상도로 스캔된 실제 우편물 영상을 사용하였다. 수취인 주소를 기준으로 프린터나 타자기 등에 의해서 인쇄된 주소를 갖는 영상을 인쇄체 우편영상이라 하고 사람이 손으로 직접 필기한 주소를 갖는 영상을 필기체 우편영상이라 할 때, 총 1,000개의 인쇄체 우편영상과 988개의 필기체 우편영상을 실험에 사용하였다.

4.2 성능 평가

표 1에 총 1,988개의 영상에 대한 제안 방법의 성능을 제시하였다. 인쇄체 우편영상의 경우 96%의 성공률을 얻을 수 있었으며 필기체 우편영상의 경우는 주소 기입 패턴의 다양성 때문에 인쇄체 영상에 비해 비교적 낮은 91.09%의 성공률을 얻을 수 있었다. 총 성공률은 93.56%이었다. 그러나 실제 우편물에서 인쇄체가 차지하는 비율이 90% 이상인 점을 고려할 경우 이보다 더 높은 수취인 주소 영역 추출 성공률을 얻을 수 있다.

제안 방법에서 군집 5, 군집 6, 군집 8 그리고 군집 9 모두를 수취인 주소 영역으로 선택할 때 수취인 주소 영역 추출 성능과 그들 중 일부를 선택하는 경우의 수취인 주소 영역 추출 성능을 비교하여 보았다. 인쇄체

표 1 수취인 주소 영역 추출 방법의 성능

	성공 영상 개수	실패 영상 개수	성공률
인쇄체	960	40	96.00%
필기체	900	88	91.09%
합	1,860	128	93.56%

영상의 경우 전자의 방법을 사용할 때 935개의 영상을 성공함으로써 후자의 방법이 더 효과적임을 알 수 있었으며 필기체 영상의 경우 전자의 방법을 사용할 때 910개의 영상을 성공함으로써 전자의 방법이 더 효과적임을 알 수 있었다. 주소 인식 시스템의 전체 성능 관점에서 보면 추후 처리과정에서 처리 성공률이 높은 인쇄체 영상의 성공률이 높아야 하므로 수취인 주소 영역을 결정할 때 후자의 방법을 선택함이 타당하다. 그림 10은 수취인 주소 영역 추출에 성공한 우편영상의 예를 보여준다. 가장 진한 검은색 문자열들이 수취인 주소를 나타낸다.

4.3 오류 분석

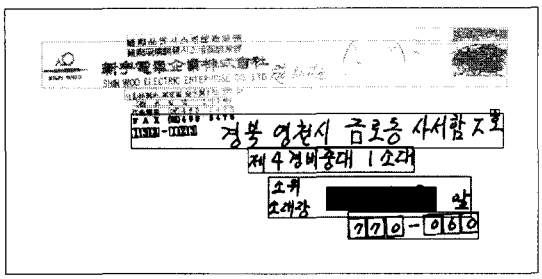
수취인 주소 영역 추출 오류는 크게 두 유형으로 나누어 볼 수 있다. 유형 I의 오류는 수취인 주소 영역에 다른 영역의 정보가 추가된 경우이며 유형 II의 오류는 수취인 주소의 일부가 손실된 경우이다. 유형 II의 오류에서 최종 배달점 코드를 획득하기 위해서 필요한 수취인 주소의 핵심 요소가 손실된 경우에는 주소 인식의 실패를 직접 유도하므로 주소 인식 시스템 관점에서 치명적인 오류에 속한다. 그러나 수취인 주소 영역에 다른 영역의 정보가 일부 추가된다하더라도 주소 해석에서 올바른 주소 인식 결과를 도출할 수 있는 가능성이 있으므로 유형 I의 오류는 주소 인식 시스템 관점에서 치명적인 오류는 아니다. 표 2는 세분화된 수취인 주소 영역 추출 오류 유형을 요약 제시하고 있다.

그림 11은 유형 I의 오류로 분류되는 수취인 주소 영역 추출에 실패한 영상의 예를 보여주고 있다. 가장 진한 검은색 문자열이 수취인 주소의 문자열이다. 오류는 문자열 생성 과정에서 발생하는 오류와 군집화 과정에서 발생하는 오류로 나누어 볼 수 있다. 그림 11의 (a)는 수취인 주소의 문자열이 발신인 주소의 문자열과 결합되어 발생한 오류를 보여주고 있다. 그림 11의 (b)와 (c)는 군집화 과정에서 발생한 오류를 보여주고 있다. (b)는 수취인 주소 영역 위쪽에 위치하는 발신인 주소의 일부가 수취인 주소에 추가된 경우의 예이며 (c)는 수취인 주소 영역 아래에 기입된 광고성 문구가 수취인 주소에 추가된 예이다. 그림 11의 (a)와 (b)에 해당하는 오류는 제안된 방법에서 문자열 생성 방법이나 군집화 방법을 개선함으로써 해결할 수 있으나 그림 11의 (c)에 해당하는 오류는 인식을 수행하기 전에는 해결하기 어

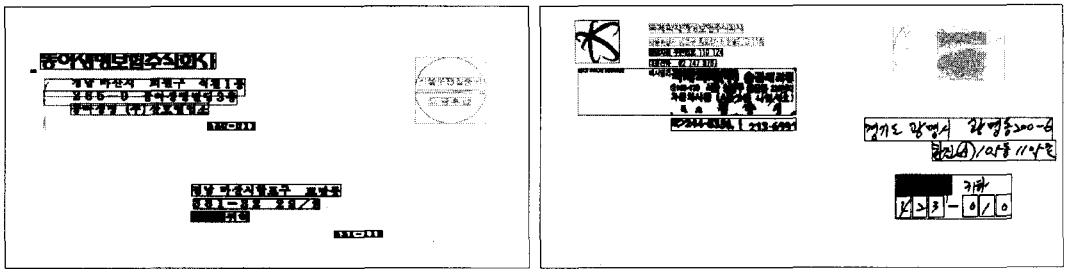
그림 10 수취인 주소 영역 추출 성공 영상의 예

표 2 오류 유형에 따른 영상의 통계

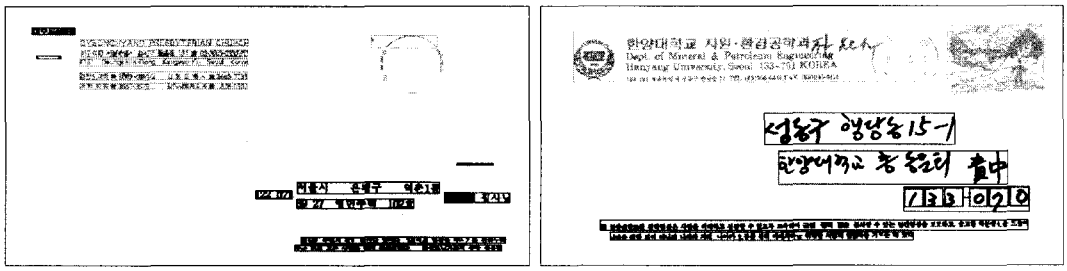
오류 유형			인쇄체	필기체
정보 추가 (유형 I)	문자열 생성 오류		-	5
	군집화 오류	위 영역의 정보 추가	3	24
		아래 영역의 정보 추가	18	6
정보 손실 (유형 II)	영상전처리 오류		17	3
	문자열 생성 오류		-	4
	군집화 오류		2	44
기타			-	2
합			40	88



(a) 문자열 생성 오류



(b) 군집화 오류 : 발신인 주소의 정보 추가



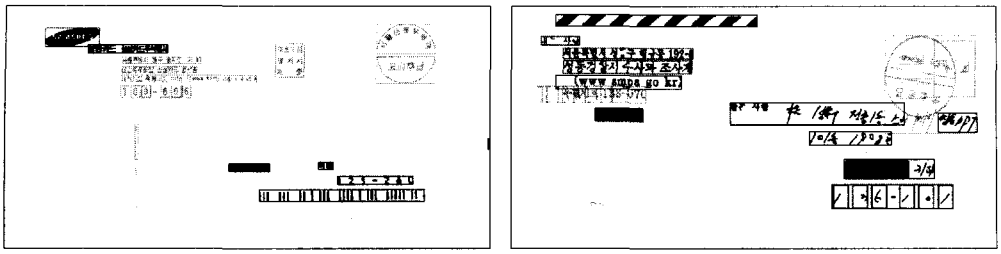
(c) 군집화 오류 : 광고 문구의 정보 추가

그림 11 정보의 추가로 인한 수취인 주소 영역 추출에 실패한 영상의 예

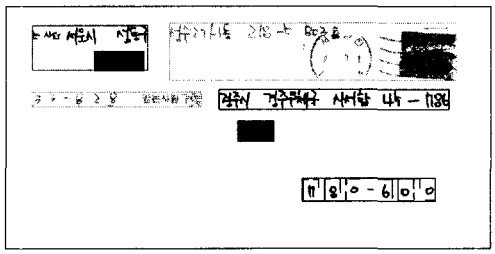
렵다. 인쇄체 영상의 오류 대부분이 아래의 영역의 정보가 수취인 주소로 추가됨으로써 발생하였다.

그림 12는 유형 II의 오류로 분류되는 수취인 주소 영역 추출에 실패한 영상의 예를 보여주고 있다. 가장 진한 검은색 문자열이 수취인 주소의 문자열이다. 오류는 영상 전처리 과정에서 발생하는 오류와 문자열 생성 과정에서 발생하는 오류 그리고 군집화 오류로 나누어 볼 수 있다. 영상 전처리 오류에는 그림 12의 (a)에서 보여 주듯이 연결요소 제거 과정에서 수취인 주소의 핵심 요소가 손실된 경우와 수취인 주소의 일부와 다른 영역의 정보가 하나의 연결요소로 추출됨으로써 발생하는 오류가 있다. 수취인 문자열이 문자열 생성 과정에서의 발생하는 오류는 앞에서 설명했듯이 수취인 주소의 문자열과 다른 영역의 문자열이 결합되는 경우이다. 그림 12의 (b)는 문자열 생성 과정에서 발생한 오류의 예를 보여준다. 그림 12의 (c)에서 보여주듯이 군집화 오류는 문자열의 군집화 과정에서 수취인 주소의 문자열 일부가 다른 영역으로 군집화됨으로써 발생하는 오류이다. 문자열 생성 오류와 군집화 오류 그리고 일부 영상 전처리 오류는 각 방법의 개선으로 해결할 수 있으나 그림 12의 (a)의 오른쪽 영상에서처럼 각기 다른 영역에 속하는 정보가 하나의 연결요소로 추출되는 오류는 연결요소 기반인 제안된 방법에서 해결하기 어려운 오류이다.

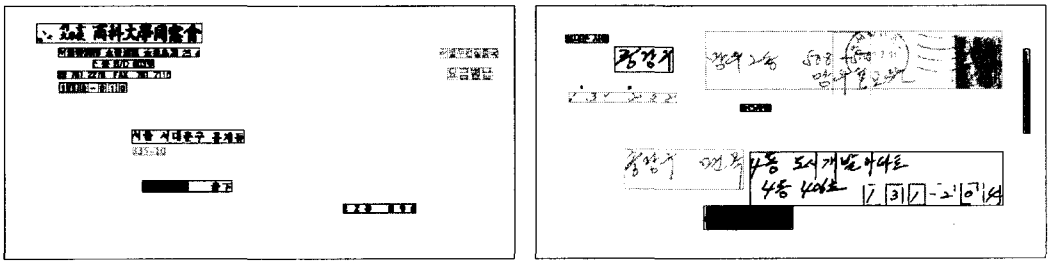
기타로 분류된 오류는 발신인 주소와 수취인 주소의 위치를 바꾸어 쓴 경우이다. 발신인 주소와 수취인 주소는 주소라는 공통된 특성을 가지므로 인식을 수행한다 하더라고 발신인 주소와 수취인 주소를 구별하기 어렵



(a) 영상 전처리 오류



(b) 문자열 생성 오류



(c) 군집화 오류

그림 12 정보의 손실로 인한 수취인 주소 영역 추출에 실패한 영상의 예

다. 따라서 이와 같이 발신인 주소와 수취인 주소의 위치가 바뀐 우편영상에서 수취인 주소 영역을 올바르게 찾는 것은 불가능하다.

5. 결론 및 향후연구

본 논문에서는 우리나라 인쇄체 우편영상과 필기체 우편영상 모두에 효과적으로 적용될 수 있는 수취인 주소 영역 추출 방법을 제안하였다. 제안 방법에서는 문자 영역 추출 후 수취인 주소 영역과 대응되는 하나의 영역을 찾으려고 했던 기존 방법들과는 달리 수취인 주소 영역이 여러 개의 영역으로 분할될 수 있다고 가정하였다. 즉, 우리나라 우편봉투에 기입된 정보의 유형별 기입 위치 패턴을 분석하여 우편영상의 영역을 분할하고 각 영역으로 군집화되는 문자열 중 오른쪽 하단에 위치하는 영역을 나타내는 군집들에 분류된 문자열들을 모아서 수취인 주소 영역을 결정하였다. 1,988개의 실제 우편영상을 가지고 제안된 방법의 성능을 평가한 결과 약 93.56%의 성능을 얻을 수 있었다. 향후 연구에서는 영상 전처리와 문자열 생성 방법을 개선함으로써 각 과정에서 발생하는 오류를 좀 더 줄이고자 하며 또한 문자열 각각에 대하여 군집화 결과의 타당성을 검증하는 방법을 개발함으로써 군집화 과정에서 발생하는 오류를 줄이고자 한다.

참 고 문 헌

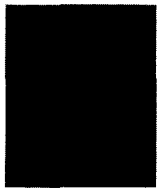
[1] 순로구분 자동처리 시스템 개발 - 최종 연구개발보고서, 정보통신부, 2001.

[2] 이성환, 김은순, "주소 및 성명에서의 한글인식을 위한 효율적인 오인식 교정 알고리즘", 한국정보과학회 논문지, 제20권, 제5호, pp. 729~738, 1993.

[3] 원유헌, 함경수 등, "필기 한글인식을 위한 오류 후처리 기법", 한국정보과학회 춘계 학술발표 논문집, pp. 829~836, 1993.

[4] 권진욱, 이일병 등, "한글주소인식 시스템", 한국정보

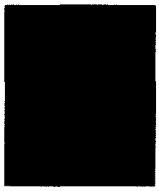
과학회 춘계 학술발표 논문집, pp. 529~532, 1997.
[5] 이관용, 권진욱, 이일병, "단어 수준의 음절 공기 확률을 이용한 한글 주소 인식", 한국정보과학회 논문지, 제25권, 제12호, pp. 1758~1768, 1998.
[6] 김수형, "최소거리분류 및 사전기반 후처리의 강결합에 의한 필기한글 주소열의 인식," 한국정보과학회 논문지, 제25권, 제8호, pp. 1195~1205, 1998.
[7] S.H. Jeong, K.T. Lim and Y.S. Nam, "A Combination Method of Two Classifiers Based on the Information of Confusion," Proceedings of the 8th International Workshop on Frontiers in Handwriting Recognition, pp. 519~523, Niagara-on-the Lake Ontario, Canada, 2002.
[8] A.K. Jain and S.K. Bhattachrjee, "Address Block Location on Envelopes Using Gabor Filters: Supervised Method," Proceedings of the 11th International Conference on Pattern Recognition, Vol. 2, pp. 264~267, 1992.
[9] S.W. Lee and K.C. Kim, "Locating Destination Address Block on Handwritten Korean Envelopes," Proceedings of the 12th International Conference on Pattern Recognition, pp. 619~621, Jerusalem, Israel, 1994.
[10] N. Nakajiman, T. Tsuchiya, T. Kamimura and K. Yamada, "Analysis of Address Layout on Japanese Handwritten Mail - A Hierarchical Process of Hypothesis Verification," Proceedings of the 13th International Conference on Pattern Recognition, pp. 726~731, Vienna, Austria, 1996.
[11] A.P. Whichello and H. Yan, "Locating Address Block and Postcodes in Mail-piece Images," Proceedings of the 13th International Conference on Pattern Recognition, pp. 716~720, Vienna, Austria, 1996.
[12] B. Yu, A.K. Jain and M. Mohiuddin, "Address Block Location on Complex Mail Pieces," Proceedings of the 4th International Conference on Document Analysis and Recognition, pp. 897~901, Ulm, Germany, 1997.
[13] M. Wolf, H. Nieman and W. Schmidt, "Fast Address Block Location on Handwritten and Machine Printed Mail Piece Images," Proceedings of the 4th International Conference on Document Analysis and Recognition, pp. 753~757, Ulm, Germany, 1997.
[14] J. Xue, X. Ding, C. Liu, S. Pan and H. Kong, "Destination Address Block Location on Handwritten Chinese Envelope," Proceedings of the 5th International Conference on Document Analysis and Recognition, pp. 737~740, 1999.
[15] W. Chanpongsaeng, P. Kumhom and K. Chamnongthai, "Locating Destination Address Block on Tahi Envelopes," Proceedings of the 5th Symposium on National Language Processing, Hua Hin, Thailand, 2002.
[16] N. Otsu, "A Threshold Selection Method from Gray-level Histogram," IEEE Transactions on Systems, Man, and Cybernetics, Vol. 9, pp. 62-66, 1979.



정 선 화
1996년 전남대학교 통계학과(이학사). 1998년 전남대학교 전산통계학과(이학석사). 2001년 전남대학교 전산통계학과 (이학박사). 2001년~현재 한국전자통신연구원 우정기술연구센터 선임연구원. 관심분야는 패턴인식, 문자인식, 영상처리, 컴퓨터비젼, 신경망 등

장 승 익
2000년 연세대학교 전산학과(이학사). 2002년 한국과학기술원 전산학과(공학석사). 2002년~현재 한국전자통신연구원 우정기술연구센터 연구원. 관심분야는 패턴인식, 문자인식, 영상처리, 컴퓨터비젼, 신경망 등

임 길 택
1993년 경북대학교 전자공학과(공학사). 1995년 경북대학교 전자공학과(공학석사). 1999년 경북대학교 전자공학과(공학박사) 1999년~현재 한국전자통신연구원 우정기술연구센터 선임연구원. 관심분야는 패턴인식, 문자인식, 영상처리, 컴퓨터비젼, 신경망 등



남 윤 석
1984년 아주대학교 산업공학과(학사) 1989년 Polytechnic Univ.(New York), Dept. of the Industrial Engineering (공학석사). 1992년 Polytechnic Univ. (New York), Dept. of the Industrial Engineering(공학박사). 1993년~현재 한국전자통신연구원 우정기술연구센터 자동구분처리연구팀장 관심분야는 소프트웨어 공학, 패턴인식 등