우편물의 접수과정에서 수취인의 주소, 성명 및 우편번호 인식

김성원[0] 김형원 양윤모
고 려 대 학 교 전자 및 정보공학부
{swkim, hwkim}@hard.korea.ac.kr, ymyang@tiger.korea.ac.kr

Character Recognition of the Receiver' s Address, Name and Postal Code in Postal Reception Process

Seong-Won Kim Hyung-Won Kim Yun-Mo Yang
Dept. of Electronic & Information Engineering, Korea University

요 약

본 연구에서는 문자 인식의 실제 응용으로서 인쇄된 우편봉투의 주소를 인식한다. 스캐너로 입력된 우편봉투 영상으로부터 주소영역과 우편번호 영역을 분리한다. 분리된 각각의 영역에서 문자를 추출하고, 전처리로써 정규화, 특징추출 단계를 거쳐 우편번호와 주소를 각각 인식하였다. 이 때, 우편번호 인식에 의하여 알 수 있는 주소와 실제로 인식한 주소의 신뢰도를 계산하여, 주소 인식 결과를 보정하는 과정을 거쳐 우편봉투의 인식을 실행하였다.

1. 서론

문자 인식은 시각 정보를 통하여 문자를 인식하고 의미를 이해하는 사람의 능력을 컴퓨터로 실현하려는 시도로서 1928년 오스트리아의 G.Tauschec가 맹인을 위한 문자 인식의 원리를 개발한 이후 많은 연구가 진행되어왔다. 이러한 문자인식의 실제 응용으로서 인쇄된 우편봉투의 주소를 인식하는 연구를 진행하게 되었다. 우체국에서 다량의 인쇄된 우편물을 사람이 일일이 손으로 구분해야 하는 번거러움이 있고 많은 시간이 소모되므로 효율적인 분류와 우편물의 추적을 위해서 우편봉투의 주소인식은 필요하다. 이에 스캐너로 영상을 입력 받아 실시간으로 우편봉투의 주소를 인식하는 방법에 대해서 연구한다.

우편봉투의 인식과정은 그림 1과 같다. 스캐너로 입력 받은 우편봉투 영상을 2진화한 뒤 우편번호 영역과 주소 영역으로 나누어 한문자씩 분리한다. 분리된 각각의 문자를 정규화, 특징량 추출의 전처리 과정을 거친 후 미리 준비한 표준특징DB와 비교하여 인식한다. 이 때, 우편번호와 주소 영역이 각각 인식이 되고 후처리에 의한 통합인식 과정을 거쳐 보다 정확한 우편봉투의 주소를 인식하게 된다.

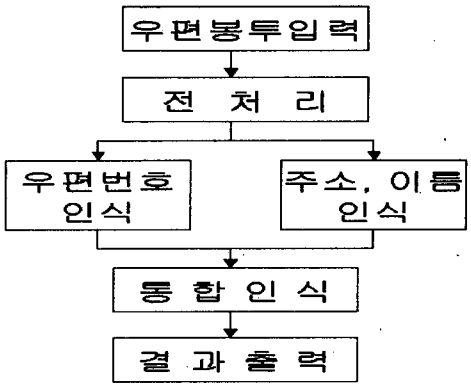


그림 1 우편봉투 인식 흐름도

인식대상 문자를 살펴보면, 우리나라에서 쓰이는 문자는 한글(완성형 2350자), 영문(52자), 숫자, 특수기호 등이 혼용되어 사용되고 있다. 그러나 실제로 우편번호에 의해서 주소로 쓰이는 문자(동, 리 단위까지)는 471자 이며, 영문 52자(대문자, 소문자), 숫자와 특수기호로 제한이 된다. 여기에 실제 생활에서 많이 쓰이는 상용문자 606문자 중 주소로 쓰이는 문자에 포함되지 않는 문자와 한자 3문자(貴, 下, 中), 세그먼트에 이용하기 위한 자음과 모음을 합하여 인식 대상문자는 715문자로 한정하였다.

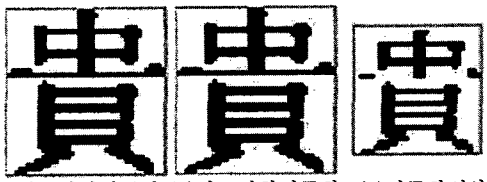
2. 전처리

2.1 히스토그램을 이용한 정규화

하나의 문자가 우편봉투 영상에서 분리되면 특징 추출 및 인식을 위하여 정규화 과정을 수행한다.

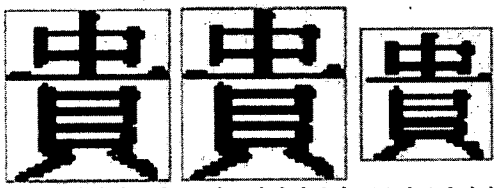
일반적인 정규화 알고리즘은 입력영상을 일정한 크기의 형상으로 선형 변형시키는 선형 정규화와 영상의 특징을 고려하는 비선형 정규화가 있다. 본 연구에서는 선형 정규화 방법을 수정, 보완하여 이용하였다. 이 때, 입력 영상이 정규화 영상의 크기보다 큰 경우, 중요한 정보를 갖는 단선성분이 제거되는 경우가 있어서 오인식의 원인이 되고있다. 이런 선형 정규화의 단점을 보완하기 위하여 문자영상의 히스토그램의 변화량을 이용하였다.

히스토그램의 변화량을 이용하는 방법은, 우선 입력 문자영상의 행, 열 방향의 히스토그램을 구하고, 앞 행(열)과 뒷 행(열)의 히스토그램과의 차이를 구하여 각 행(열)의 변화량을 구한다. 이 때에 선형 정규화에 의하여 삭제될 행(열)이 정하여 졌을 때, 삭제될 행(열)의 앞 행(열), 뒷 행(열)과의 변화량을 비교하여 변화량이 적은 행(열)을 삭제한다.



(a)세로방향정규화 (b)가로방향정규화 (c)정규화영상

그림 2 기존의 선형정규화



(a)세로방향정규화 (b)가로방향정규화 (c)정규화영상

그림 3 히스토그램을 이용한 선형 정규화

그림 2 는 임의의 입력영상(42×40)에 대한 일반적인 선형 정규화를 나타낸다(본 논문에서 정규화 영상의 크기는 32×32로 한다). 그림 2(a), (b) 는 각각 가로, 세로 방향의 정규화를 나타내고, 흐리게 표시된 선이 삭제가 될 행 또는 열이 되며 (c) 는 정규화된 영상을 나타낸다. 이때 그림 2 의 (c) 에서 영상의 중간부분의 단선 성분이 사라져 이후 특징량 구성 및 인식 단계에서 악영향을 끼치게 된다. 하지만, 그림 3의 (c) 는 히스토그램 변화량을 이용하여 중요 단선성분의 삭제를 방지한 것을 알 수 있다.

2.2 특징량 구성

문자영상의 정규화 이후 특징량 구성을 위하여 특징소를 추출한다. 문자의 윤곽선이 갖는 방향정보에 중점을 두어 4 가지 방향성분(0^0, 45^0, 90^0, 135^0)을 추출하게 된다.

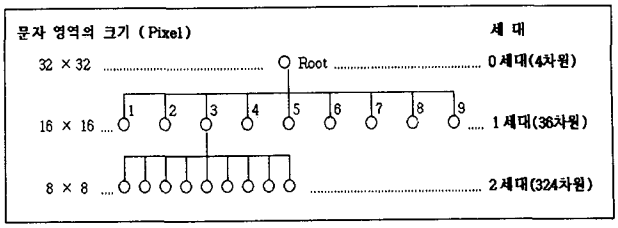


그림 4 9 진트리를 이용한 특징량 구성

이렇게 추출된 특징소는 그 위치에 따라 1/2 씩 중첩된 9 개의 소영역으로 분할하여 각기 서로 다른 위치별 가중치를 갖고 4 차원, 36 차원, 324 차원의 특징벡터로 구성이 된다(그림 4). 이러한 계층구조에 의하여 구성된 특징벡터에 의해서 문자 영상의 중앙부분을 강조하고, 어느 정도의 잡음에 대해서도 안정적인 인식이 가능하다.[1][2]

본 연구에서 특징량은 324 차원($9^2 \times 4$)으로 하였다.

3. 인식방법 및 실험

3.1 문자인식 실험

우편봉투의 실제 인식에 앞서, 715 문자에 대한 예비 실험을 실행 하였다. 715 문자의 15 개 세트를 만들어 그 중 8개 세트로 표준 특징 벡터를 만들고, 나머지 7 세트에 대해서 인식 실험을 하였다(표 1). 거리계산은 'City Block 거리 계산 방법'을 사용하였으며, 715 개의 표준 특징벡터와의 거리를 계산한 후, 거리가 가장 작은 문자를 1 위 후보로 하였다.

표1 문자 인식 실험결과

특징벡터 (324차원)	1위 인식율	2위 인식율	3위 인식율	10위 인식율
7set 평균	96.11	96.83	97.06	99.93

참고로 학습에 이용한 DB 를 인식한 실험(8 세트)에서는 1위 인식율의 평균이 99.1% 를 나타내었다.

3.2 우편번호 인식

현재 전국의 우편번호는 총 8354 개이며, 단지 숫자 6 개를 인식하여 주소의 동, 리 단위까지의 정보를 얻을 수 있고, 그 정보를 이용하여 주소인식의 정확성을 높일 수 있다.

또한 인식대상 문자가 숫자 0~9 까지 10 개에 불과하지만 실제 우편봉투에서 우편번호를 추출할 때 두 숫자사이가 너무 좁아 두 숫자가 같이 분리(segmentation)되는 문제가 많이 발생한다. (그림 5 참조)

(a) 원 영상 (5 0 0 □ 0 4 1)

(b)실제 분리된 영상(5 00 04 1) (c)붙은숫자의 정규화영상

그림 5 두 숫자가 붙어서 분리된 예

이런 경우 정규화 이전의 원영상의 가로,세로 길이의 비율(가로/세로 길이)에 의해서 붙은 것을 판별하였다. 인식에서는 2 개의 숫자가 붙은 것을 다시 분리하지 않고, 10개의 숫자가 각각 붙은 모든 경우를 하나의 문자로 간주하여 인식하였다(그림5(c)참조).

인식된 우편번호를 이용하여 우편번호주소DB를 검색, 후처리에 이용될' 주소문자열' 을 확보한다.

3.3 주소 인식

주소 인식과정에서 인식율에 크게 영향을 미치는 원인은 분리(segmentation) 과정이다. 우편번호 인식과는 반대로 한 개의 문자가 따로 떨어져서 두개의 문자로 세그먼트되는 경우가 많이 발생하게 된다.(그림6참조)

(a) 원영상 (특 별 시)

(b) 세그먼트 영상 (특 별 ㅅ ㅣ)

그림 6 한 문자가 분리되어 세그먼트된 예

이처럼 하나의 문자가 두개로 분리된 경우, 분리된 각각의 문자 크기가 반각이 된다. 따라서 현재 인식할 문자가 반각인 경우 반각문자 표준DB와 비교하여 분리된 자음인지 여부를 알 수 있으며 feedback과정을 거쳐 정확한 세그먼트를 하였다.

4. 통합인식

주소영역에서 한 문자의 인식이 이루어지면 우편번호 인식(3.2)에 의하여 저장된 ' 주소문자열' 을 토대로 후처리를 실행한다. 인식하고자 하는 문자의 대부분은 후보문자 10위안에 포함된다(표1참조). 따라서 10위 후보문자까지 ' 주소문자열' 과 비교하여 인식의 정확도를 높일 수 있다.

또한 1위 후보 문자의 인식 거리가 문턱값 이상인 경우와 1위 후보문자와 2위 후보문자간의 거리 비율(2위/1위)이 1.05 이하인 경우 인식불가 판정(reject)을 시행하여 인식의 신뢰도를 높였다.

실제 우편봉투 인식 실험에서는 256 Gray level의 300 dpi 해상도로 스캔된 우편봉투 35 장, 200 dpi해상도로 스캔된 우편봉투 5장으로 총 40장의 우편봉투를 대상으로 하였다. 이중 레이져 프린터로 인쇄된 것이 24, 잉크젯 프린터로 인쇄된 것이 16장이며, 폰트별로는 명조체가 24, 고딕체가 15, 기타 1 장이다.

표 2 우편봉투 인식 실험결과

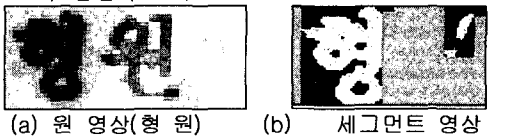
해상도	총봉투수	인식봉투수	인식율(%)
300dpi	35	31	88.6
200dpi	5	0	0

위의 표2에 우편봉투 40장의 인식 실험 결과를 나타내었다. 300dpi로 스캔된 봉투 35장 중 31장의 주소를 정확히 인식하였지만, 200dpi의 저해상도로 스캔된 봉투 5장은 주소인식에 모두 실패하였다. 이 때, 300dpi로 스캔된 35장의 봉투에 쓰인 문자의 총 개수는 884개이며 reject된 문자는 192문자 이다. 이때 오인식율을 살펴보면 reject된 문자를 제외한 692문자 중 18문자가 오인식을 하여 97.4%의 문자 인식율을 나타내었다(표3참조).

표 3 우편봉투 문자인식 결과 (300dpi)

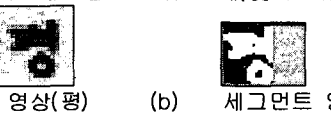
	총 문자 수	1 위 인식이 안된 총 문자 수		전체 Reject 된 문자 개수		1위 인식 문자	오인식 문자수
		우편번호 인식 부분	번지 이름 부분	1위 Reject	그 외 Reject	인식 문자 수	오인식 문자 수
개 수	884	78	131	58	134	674	18
전 체		209		192			

오인식 및 Reject 문자의 원인을 살펴보면, 가장 많은부분은 세그먼트 오류에 있다. 우편봉투의 주소부분에 비닐포장이 되어있는 경우, 영상 입력시에 빛의 간섭이 발생하여 문자 추출 단계에서 문장영역의 정보 손실이 발행한다(그림 7(b)참조). 또한 인쇄된 문자의 크기가 너무 작거나 입력 해상도가 낮은 경우(200dpi)에는 문자가 찌그러져 노이즈가 발생하게 된다(그림 (b) 참조). 이런 경우 오인식 및 Reject의 원인이 된다.



(a) 원 영상(형 원) (b) 세그먼트 영상

그림 7 세그먼트 오류의 예(빛의간섭)



(a) 원 영상(평) (b) 세그먼트 영상

그림 8 문자에 노이즈가 발생한 예(저해상도 입력)

향후 발전방향으로 저품질 문자(노이즈, 저해상도) 인식에 관한 연구가 필요하다.

5. 참고문헌

[1] 강선미, 이기용, 황승욱, 양윤모, 김덕진," 고속문자인식을 위한 특징량 추출에 관한 연구" , 전자공학회 논문지 29B, Vol. 11, pp.1047~1056, 1992.

[2] 송효섭, 장세진, 신병주, 양윤모," 손의 형상과 움직임 방향 정보를 이용한 수화인식" , 정보과학회 논문지(B), Vol. 26, No. 6, pp,804~810 1999.