연결요소 분석에 기반한 인쇄체 한글 주소와 필기체 한글 주소의 구분

(Classification of Handwritten and Machine-printed Korean Address Image based on Connected Component Analysis)

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

요 약 본 논문에서는 우편봉투 상에 기입된 인쇄체 한글 주소와 필기체 한글 주소를 효과적으로 구분할 수 있는 방법을 제안한다. 문자인식 모듈을 포함하는 각종 응용 시스템에서 입력 영상이 인쇄체인지 필기체인지 구분하는 것은 매우 중요하다. 이는 대부분의 경우 인쇄체 영상과 필기체 영상이 갖는 특징이 상이하여, 각 영상에서의 문자 및 문자열 분리 방법, 문자 인식 방법 등이 매우 상이하게 개발되기 때문이다. 본 논문에서 제안한 구분 방법은 연결요소 추출 및 병합, 특징 추출, 영상 구분 순으로 수행된다. 연결요소 추출 및 병합 단계에서는 입력영상으로부터 연결요소를 추출한 후 일부 연결요소들에 대하여 병합을 시도하며, 특징 추출 단계에서는 병합결과 얻어진 연결요소들의 그룹들로부터 폭과 위치에 관련된 특징을 추출하고, 영상 구분 단계에서는 추출한 특징을 입력으로 제공받는 다층퍼셉트론을 사용하여 구분을 시도한다. 제안한 방법의 우수성을 증명하기 위해 실제 우편물로부터 추출된 3,147개의 한글 주소 영상을 사용하여 실험한 결과, 98.85%의 구분률을 보여주었다.

키워드 : 광학문자인식(OCR), 주소기입형식, 주소인식시스템, 다층퍼셉트론

Abstract In this paper, we propose an effective method for the distinction between machine-printed and handwritten Korean address images. It is important to know whether an input image is handwritten or machine-printed, because methods for handwritten image are quite different from those of machine-printed image in such applications as address reading, form processing, FAX routing, and so on. Our method consists of three blocks: valid connected components grouping, feature extraction, and classification. Features related to width and position of groups of valid connected components are used for the classification based on a neural network. The experiment done with live Korean address images has demonstrated the superiority of the proposed method. The correct classification rate for 3,147 testing images was about 98.85%.

Key words : Optical character recognition (OCR), Address type identification, Address reading system, Multi-layer perceptrons

1. 서 론

일일 처리를 요구하는 우편물량의 증가와 함께 신속하고 정확한 우편 서비스에 대한 고객의 요구가 증대됨에 따라 우리나라는 우편집중국을 건설하고 우편물 자동구분기를 도입하여 우편물의 자동처리 체계를 구축하고 있다[1]. 현재, 전국 우편집중국에 도입된 우편물 자동구분기는 우편번호 인식 시스템만을 탑재하여, 우편번호 인식결과에 따라 발송, 도착 구분의 자동화가 되어있다. 이러한 시스템은 우편번호가 기입되지 않았거나 기입되었다 하더라도 잘못된 정보를 담고 있는 경우 우편물 구분을 올바르게 수행하지 못하는 단점을 가지고 있으며, 시스템 처리율 저하의 주요 원인이 된다. 따라서 우편물 처리 자동화율을 높이기 위해서는 이러한 우편물에 기입된 우편번호뿐만 아니라 수취인 주소 인식이 반드시 필요하다[2-6].

† 정 회 원 : 한국전자통신연구원 우정기술연구센터 연구원
sijang@etri.re.kr
ktlim@etri.re.kr
†† 비 회 원 : 한국전자통신연구원 우정기술연구센터 연구원
sh-jeong@etri.re.kr
††† 비 회 원 : 한국전자통신연구원 우정기술연구센터 자동구분처리 연구팀장
ysnam@etri.re.kr
논문접수 : 2003년 3월 26일
심사완료 : 2003년 6월 19일

우편물에서 수취인 주소의 인식을 위해서는 수취인 주소 영역 추출, 주소 기입 형식 구분, 문자열 분리, 문자 분리, 문자 인식, 주소 해석 모듈 등이 필요하며, 각각의 모듈은 다음과 같은 순으로 진행된다. 먼저, 우편물에 기입된 주소 중 수취인 주소만이 주소 인식 시스템의 인식대상이기 때문에 수취인 주소 영역 추출이 수행된다. 다음으로, 우편물의 수취인 주소가 사람에 의해 필기된 주소인지 기계에 의해 인쇄된 주소인지 구분해 주는 주소 기입 형식 구분 모듈이 실행된다. 주소 기입 형식에 따라 인쇄체 또는 필기체 문자열 분리, 문자 분리, 문자 인식 모듈이 순차적으로 수행된 후, 주소 해석 모듈이 수행되어 최종 수취인 주소 인식 결과를 도출하게 된다. 이러한 일련의 과정 중에서 주소 기입 형식 구분 모듈의 역할은 매우 중요하다. 이는 대부분의 문자열 분리, 문자 분리, 문자 인식 모듈을 위한 알고리즘들이 입력 대상이 인쇄체 영상인지 필기체 영상인지에 따라 매우 상이한 방법으로 개발되기 때문이다. 이러한 이유로, 인쇄체 모듈에 필기체 영상이 입력될 경우 좋은 성능을 기대하기 어려우며, 그 반대의 경우도 마찬가지이다. 즉, 주소 기입 형식 구분 모듈의 성능에 따라 최적의 성능을 갖는 주소 인식 시스템의 개발이 가능하다.

본 논문에서는 주소 기입 형식 구분 방법을 제안한다. 제안한 방법에서는 연결요소로 표현된 입력 영상에서 최상위 문자열에 속해 있는 연결요소를 추출한 후, 추출된 연결요소들을 문자 단위의 그룹이 되도록 병합을 시도한다. 그 다음, 이전 단계의 결과로 얻어진 연결요소 그룹들로부터 특징을 추출하고, 추출된 특징을 한 개의 은닉층을 갖는 다층퍼셉트론의 입력으로 사용한다. 사용된 특징은 각 연결요소 그룹의 최소인접사각형으로부터 추출한 폭 및 위치 특징에 대한 히스토그램이다. 실험을 통하여 본 논문에서 제안한 방법이 유효함을 입증하였다.

본 논문의 나머지 구성은 다음과 같다. 먼저, 2장에서 관련 연구를 기술하고, 3장에서는 연결요소 추출 및 병합 방법, 특징 추출 방법, 분류 방법에 관하여 자세히 설명한다. 4장에서는 제안된 방법의 성능을 평가하기 위해 사용한 데이타 및 그 데이타를 이용한 실험 결과를 기술하며, 5장에서는 결론 및 향후 연구를 기술한다.

2. 관련 연구

문자 인식 모듈을 포함하는 응용 시스템은 입력 영상이 인쇄체인지 필기체인지에 따라 매우 다르게 개발된다. 또한, 인쇄체 시스템에 필기체 영상이 입력되거나, 필기체 시스템에 인쇄체 영상이 입력되면 시스템의 성능이 저하될 뿐만 아니라 의도된 결과를 얻을 수 없게 된다. 이와 같이 입력 영상의 종류 - 인쇄체 또는 필기체 - 의 구분이 매우 중요함에도 불구하고, 이에 대한 연구는 활발히 이루어지지 않고 있다[7-11].

Fan 등은 [7]에서 공간특징과 문자 블록 레이아웃 분산(character block layout variance)을 사용하여 문자열 블록 영상이 인쇄체인지 필기체인지를 판별하는 방법을 제안하였다. [7]에서 제안한 방법은 기울어짐이 없는 이진화된 문자열 블록 영상을 입력으로 가정하고, 문자열 및 문자 분리에 X-Y cut 알고리즘 [12]을 사용하였다. 다음으로, 문자의 크기 및 거리 히스토그램을 사용하여 문자열 블록을 몇 개의 문자열 부분 영상으로 나누었다. 여기서 크기 히스토그램은 문자의 높이에 대한 히스토그램이고 거리 히스토그램은 문자의 중심 간의 거리에 대한 히스토그램이다. 주어진 임계치보다 작은 문자가 있는 곳이나 주어진 임계치보다 큰 거리를 갖는 문자 사이에서는 하나의 문자열 영상이 다수의 문자열 부분 영상으로 나누었다. 인쇄체와 필기체를 구분하는 단위는 이렇게 얻어진 문자열 부분 영상이다. 문자열 부분 영상에 대하여 문자 블록 레이아웃 분산을 계산하여 임계치 보다 작으면 인쇄체로 분류하고, 그렇지 않으면 필기체로 분류한다. 문자 블록 레이아웃 분산은 문자 블록들이 얼마나 규칙적으로 배열되었는지를 나타내는 특징이다. 하지만, 실세계에서 마주치는 영상은 단순 X-Y cut 알고리즘으로 문자열 및 문자 분리가 매우 어렵다. 또한, X-Y cut 알고리즘으로 문자 단위로의 분리까지 완벽하게 이루어졌다 하더라도 사용된 특징이 유효하기 위해서는 서로 다른 문자들이 비슷한 크기 및 폭을 가지고 있어야하는 제약이 있다.

Pal과 Chaudhuri는 [8]에서 Devnagari와 Bangla로 쓰여진 문자열 영상을 인쇄체 또는 필기체로 구별해 주는 방법을 제안하고 있다. 입력은 이진화 과정을 거친 문자열 영상이다. 제안된 방법은 두 언어가 공통적으로 갖는 고유의 특성을 고려하여 개발된 구조적 특징과 [7]의 문자 블록 레이아웃 분산과 유사한 통계적 특징을 사용하였다. 먼저, 구조적 특징을 계산하여 임계치보다 크면 입력 영상을 인쇄체로 분류하며, 임계치보다 작은 영상은 분류를 보류한다. 그 다음, 통계적 특징을 계산하여 임계치보다 작으면 인쇄체 영상으로, 그렇지 않으면 필기체로 분류하였다. 그러나 이 방법은 Devnagari와 Bangla 언어에 특화된 구조적 특징을 사용함으로써, 다른 언어로 작성된 문자열 영상의 분류에 직접 적용되기 어렵다.

Kuhnke 등은 [9]에서 이진화된 영문 문자 영상이 인쇄체인지 필기체인지를 구별하는 방법을 제안하였다. 제안된 방법은 전처리, 특징 추출, 분류 순으로 이루어진다. 전처리 단계에서는 문자 영상의 외곽선(contour)을 추출했다. 특징 추출 단계에서는 영문자가 대부분 직선

성분으로 이루어졌다는 특성을 고려하여 수평방향 직선 성분과 수직방향 직선성분을 추출하였으며, 직선이 존재하지 않는 영문자의 구별을 위하여 중심을 기준으로 화소들의 대칭성을 고려하여, 총 11개의 특징을 추출하였다. 추출된 특징은 두 개의 은닉층을 갖는 신경망을 이용해 구분하였다. 이 방법 또한 영문자에 특화된 특징을 사용함으로써, 다른 언어로 작성된 문자 영상의 분류에 적용하기 어렵다.

Imade 등은 [10]에서 하나의 문서를 사각형 모양의 여러 부분 영역으로 나누고, 각 부분 영역을 인쇄체 Kanji와 Kana 문자로 구성되는 영역, 필기체 Kanji와 Kana 문자로 구성되는 영역, 사진, 그림으로 분류하는 방법을 제안하였다. 부분 영역을 분류하는 방법은 다음과 같다. 먼저, 명도 레벨의 부분 영역 영상에서 32×32 블록을 무작위로 추출하여, 추출된 블록에서 분류를 위한 특징을 계산한다. 특징으로 32차원의 기울기 벡터(gradient vector) 특징과 24차원의 휘도 히스토그램 특징이 사용된다. 분류는 부분 영상에서 몇 개의 블록을 추출한 후, 하나의 은닉층을 갖는 신경망을 사용하여 각 블록의 분류를 수행하고 다수의 결정을 최종 결정으로 하여 분류하였다. 이 방법은 [7-9]에 비해 기울어진 영상이나 다른 언어로 작성된 문자열 및 문자열 블록의 분류에 적용이 용이하다. 그러나, 사용된 특징이 문자열 블록과 사진 또는 그림과의 구별에는 효과적이나, 인쇄체 문자열 블록과 필기체 문자열 블록의 분류에는 효과적이지 못하다.

마지막으로, Franke와 Oberlander는 [11]에서 형식 문서의 필드에 채워진 데이타가 인쇄체인지 또는 필기체인지 구별하기 위한 방법을 제안하였다. 이진화된 필드 데이타를 연결요소들의 최소인접사각형으로 표현하고, 그들로부터 네 종류의 기하학적 특징을 계산한 뒤, 서로 다른 특징에 특화된 네 개의 통계적 분류기를 사용하여 분류를 시도하였다. 각 분류기의 결과를 결합하기 위해서 Fisher의 선형 판별 함수[13]를 사용한 통계적 분류기를 이용하였다. 네 종류의 특징은 최소인접사각형들의 폭 히스토그램, 높이 히스토그램, 인접한 최소사각형 사이의 최소 거리 히스토그램 및 중심 거리 히스토그램이다. 이 방법은 서로 다른 문자의 폭 및 높이가 비슷하게 작성된 문자열 영상의 분류에 사용이 가능하다. 또한, 사용된 특징은 연결요소가 하나의 문자와 대응될 때 매우 유효하게 된다.

본 논문에서 제안한 방법은 기존 연구와 비교해 다음과 같은 차별성을 갖는다. 첫째, 제안된 방법의 입력은 우편물로부터 추출된 한글 주소 영상이다. 즉, 한글 문자열 블록 영상이다. 따라서, 문자열 영상의 분류에 사용된 방법[8, 11]을 활용하기 위해서는 문자열 블록을 문자열로 분리하는 단계가 요구된다. 그러나 제안 방법은 입력 영상에 기울어짐을 허용하기 때문에, [7]에서처럼 단순 X-Y cut 알고리즘으로 문자열 분리가 어렵다. 둘째, 문자열 분리 및 문자 분리가 이루어졌다 하더라도, 제안 방법은 한글 문자열 블록 영상을 구별하고자 하기 때문에, [8]과 [9]에서처럼 한글과 다른 특성을 갖는 언어에 특화되어 개발된 특징은 사용되기 어렵다. 제안 방법은 [11]의 방법에서 개발된 특징을 일부 사용하고 있으며, 기울어진 문자열 블록 영상도 구분할 수 있도록 새로운 특징을 개발하였다. 또한, 문자열 블록 영상 분류에 문자열 영상 분류에 사용되는 특징을 활용하기 위해서, 문자열 블록 영상으로부터 특징 추출을 위한 일부 영역 - 문자열 영상과 비슷한 - 을 효과적으로 추출하는 방법을 제안하였다.

3. 주소 기입 형식 구분

그림 1은 본 논문에서 제안한 주소 기입 형식 구분 방법의 순서를 보여준다. 먼저, 입력된 한글 주소 영상으로부터 최상위 문자열에 속하는 연결요소를 추출한 후, 추출된 연결요소들에 대하여 대략적인 문자 단위가 되도록 병합을 시도한다. 문자 단위로 병합된 연결요소들의 그룹들로부터 분류를 위한 특징을 계산하며, 사용된 특징은 그룹들의 폭 및 위치 정보가 누적된 히스토그램이다. 마지막으로, 추출된 특징을 사용하여 주소 영상을 인쇄체 또는 필기체로 분류하기 위하여 다층퍼셉트론을 채택하였다.

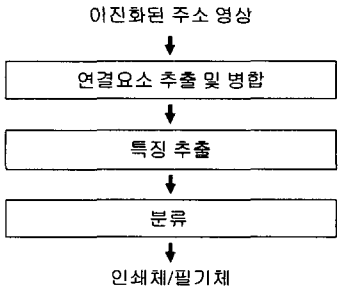


그림 1 주소 기입 형식 구분 방법

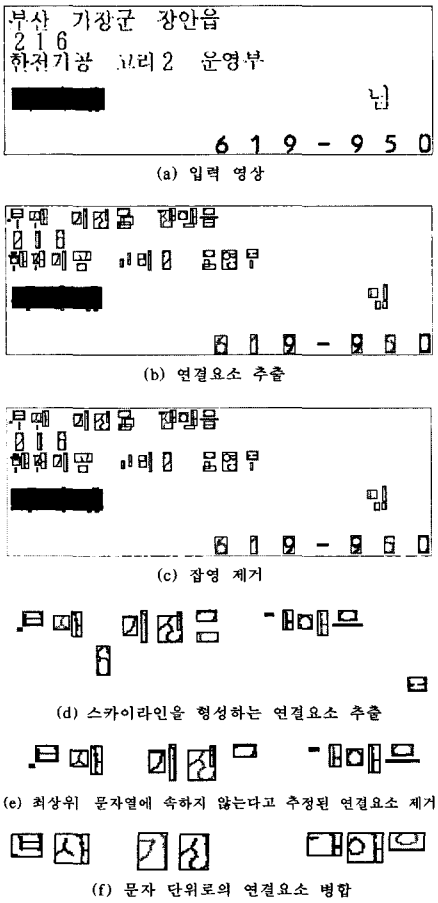
3.1 연결요소 추출 및 병합

제안 방법의 입력 주소 영상은 200dpi(dots per inch)의 해상도로 스캔된 후, Otsu의 전역적 방법 [14]을 사용하여 이진화된 우편영상으로부터 추출되었기 때문에, 그림 2의 (a)와 같이 잡영이 존재하거나 문자의 획 일부가 손실된 경우가 많다. 이러한 저품질 영상으로부터 연결요소를 추출하여 대략적일지라도 문자 단위로 병합하기 위해서는 일부 연결요소를 잡영으로 분류하여 제

거하는 과정과 하나의 문자가 여러 개의 연결요소로 표현된 경우 병합하는 과정이 필요하다.

특징 추출을 위한 연결요소의 추출 및 병합 과정은 총 다섯 단계로 구성된다. 그림 2의 (b)에서 (f)까지는 각 단계의 결과 영상을 보여주고 있다. 총 다섯 단계 중 첫 번째 단계와 두 번째 단계는 특징 추출을 위한 연결요소 추출의 전처리 단계에 해당된다. 첫 번째 단계에서는 주어진 입력 영상으로부터 8방향 연결성을 가지는 연결요소를 추출한다. 두 번째 단계에서는 잡영으로 추정되는 연결요소들을 제거한다. 이때, 제거 대상은 연결요소의 화소수가 전체 연결요소의 평균 화소수의 5%이하이거나, 연결요소의 최소인접사각형의 면적이 화소 기준 102이하인 연결요소이다. 그림 2의 (c)영상은 잡영 제거를 수행한 후 얻어진 결과이다.

세 번째 단계부터는 실제로 특징 추출을 위한 단위를 얻기 위하여 해당 연결요소를 추출하고 병합하는 단계이다. 병합 결과 얻어진 연결요소들의 최소인접사각형들



(a) 입력 영상

(b) 연결요소 추출

(c) 잡영 제거

(d) 스카이라인을 형성하는 연결요소 추출

(e) 최상위 문자열에 속하지 않는다고 추정된 연결요소 제거

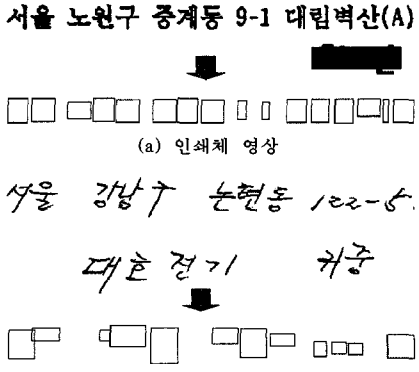
(f) 문자 단위로의 연결요소 병합

그림 2 연결요소 추출 및 병합 과정

이 특징 추출을 위한 단위가 된다. 제안한 방법에서는 특징 추출을 위한 연결요소들을 최상위 문자열에서 추출하고자 하였다. 최상위 문자열에 속하는 연결요소를 추출함으로써, 문자열간의 접촉에 영향을 덜 받으면서 동일한 문자열에 속하는 연결요소를 비교적 용이하게 추출할 수 있다는 장점을 갖는다. 세 번째 단계에서는 최상위 문자열에 속하는 연결요소를 추출하기 위하여 스카이라인을 형성하는 연결요소를 추출한다. 그러나, 그림 2의 (d)에서 보여주는 것처럼 단어와 단어사이에 공백이 존재하거나, 그 다음 문자열이 최상위 문자열보다 긴 경우 최상위 문자열 아래의 문자열들에 속하는 연결요소들이 추출될 수 있다. 네 번째 단계는 이러한 연결요소들을 제거하는 단계이다. 이때, 세 번째 단계까지 제거되지 않은 각 연결요소의 최상위 y값의 평균보다 아래에 존재하는 연결요소가 제거된다. 다섯 번째 단계에서는 대략적인 문자 단위로의 병합을 시도한다. 주어진 임계치 이내의 거리에 있는 작은 연결요소를 이웃한 연결요소와 병합하거나 임계치 이상으로 서로 겹침이 발생한 연결요소들을 병합한다.

3.2 특징 추출

특징은 대략적인 문자 단위로 병합된 연결요소 그룹의 최소인접사각형들로부터 추출된다. 인쇄체 문자들은 기계에 의해 인쇄되기 때문에 폭이 일정하고, 동일한 문자열에서 추출된 문자들은 일정한 기울기를 가진다. 반면, 사람이 직접 필기한 문자 영상에서는 위와 같은 규칙성을 보장할 수 없다. 또한, 연결요소 추출 및 병합을 위해 개발된 알고리즘은 인쇄체의 요구사항이 만족되도록 개발되어졌기 때문에, 필기체 영상에 알고리즘을 적용하는 경우 인쇄체 영상에서와는 다른 결과를 출력한다 - 그림 3 참조. 이러한 차이를 고려하여 인쇄체와 필기체를 분류하기 위해서 제안 방법은 세 종류의 특징을 사용하였으며, 이들은 각각 최소인접사각형의 폭의



(a) 인쇄체 영상

(b) 필기체 영상

그림 3 특징추출을 위한 연결요소 추출 및 병합 결과의 예

크기에 대한 분포 특징, 폭의 분산에 대한 분포 특징, 위치에 대한 분포 특징이다.

3.2.1 폭 크기 히스토그램

주어진 입력 영상으로부터 N개의 최소인접사각형이 추출되었다고 하자. 이를 왼쪽에서 오른쪽 순으로 B_1, …, B_N이라 표기하자. 이때, i 번째 최소인접사각형 B_i의 폭 크기 특징 w_i는 아래와 같이 계산된다. 여기서, x_i^{max}는 B_i의 최대 x 좌표 값이며, x_i^{min}은 B_i의 최소 x 좌표 값이다.

$$w_i = x_i^{max} - x_i^{min} + 1, \quad i = 1, \ldots, N$$

이렇게 계산된 N개의 폭 크기 특징은 히스토그램의 해당 계급구간에 누적되어 분류 특징으로 사용된다. 훈련 데이타를 기반으로 10개의 계급구간을 갖는 히스토그램이 사용되었으며 각 계급에 속하는 값은 총합이 1이 되도록 정규화 된다. 정규화 과정은 폭 분산 히스토그램과 위치 히스토그램에도 동일하게 적용된다.

3.2.2 폭 분산 히스토그램

i 번째 최소인접사각형 B_i의 폭 분산 특징은 아래처럼 계산된다. 여기서, $\overline{w}$는 w_i들의 평균이다.

$$|w_i - \overline{w}|, \quad i = 1, \ldots, N$$

폭 크기 히스토그램처럼 10개의 계급구간을 갖는 히스토그램이 사용되었다. 폭 크기 히스토그램은 폭의 크기와 분산을 모두 포함하고 있다. 마찬가지로, 폭 분산 히스토그램 역시 폭의 크기와 분산 정보를 가지고 있다. 차이점은 폭 분산 히스토그램에 있는 분산에 대한 정보가 크기에 대한 정보보다 더 많다는 것이다. 이 때문에, 폭 분산 히스토그램은 문자의 크기에 상대적으로 영향을 받지 않게 된다.

3.2.3 위치 히스토그램

위치 특징은 동일 문자열에 속하는 문자들이 일직선상에 놓여있는 정도를 측정한 것이다. 이를 측정하기 위하여 각 최소인접사각형의 최상위 y좌표의 상대적 위치를 사용한다. 최상위 y좌표의 상대적 위치를 사용한 이유는 영상의 기울어짐에 영향을 받지 않기 위해서이다. 제안 방법에서는 상대적 위치 특징을 얻기 위하여 그림 4와 같이 양끝 연결요소를 잇는 가상의 직선을 도입하였다.

직선의 기울기는 아래와 같이 계산된다. 여기서, x_i는 B_i의 중앙 x 값이고 y_i는 B_i의 최상위 y좌표의 값이다.

$$\theta = \tan^{-1}\left(\frac{y_N - y_1}{x_N - x_1}\right)$$

이때, i 번째 최소인접사각형 B_i의 위치 특징 p_i는 x_i와 직선과의 최소 거리로 계산된다. 거리는 아래의 식을 사용하여 계산될 수 있다. 위치 히스토그램에서는 14

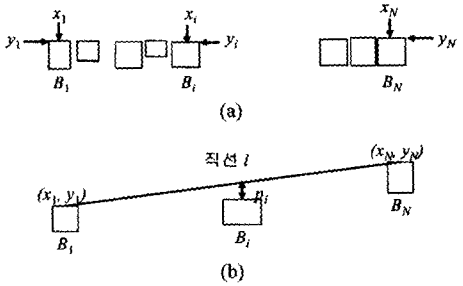


그림 4 위치 특징 p_i와 직선

개의 계급구간을 사용하였다.

$$p_i = (x_i - x_1)\tan\theta, \quad i = 1, \ldots, N$$

3.3 분류

위에서 계산된 총 34차원의 특징벡터를 기반으로 인쇄체와 필기체 영상을 구분하기 위하여 채택된 분류기는 그림 5와 같이 한 개의 은닉층을 갖는 다층퍼셉트론이다. 입력층은 특징벡터의 차원과 동일한 34개의 노드, 은닉층은 20개의 노드, 출력층은 2개의 노드로 구성된다. 출력층의 각 노드는 인쇄체와 필기체를 나타낸다.

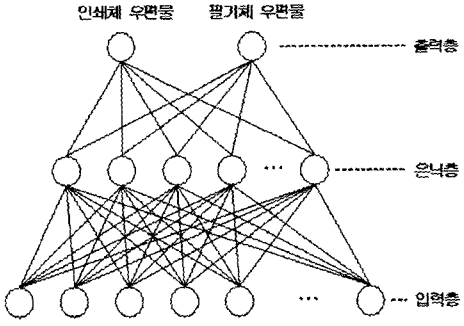


그림 5 다층퍼셉트론의 구조

또한, 입력층과 은닉층에는 편기항(bias term) 노드를 하나씩 추가하였으며, 은닉층과 출력층의 활성화 함수로는 시그모이드(sigmoid) 함수를 사용하였다. 각 노드 사이의 연결강도(weight)를 훈련하기 위해 목표값과 출력값의 차이에 대한 평균 제곱을 최소화하도록 연결강도를 조정하였으며, 이때 평균 제곱에 대한 최소값을 찾기 위하여 오류역전파 학습 알고리즘(error back-propagation algorithm)을 사용하였다. 이때, 사용된 훈련 파라메터 α(학습률)와 η(모멘텀)의 값으로 각각 0.1과 0.7을 사용하였다.

4. 실험 및 결과

4.1 실험 데이타

제안 방법의 성능을 평가하기 위하여 실제 우편 영상에서 수작업으로 추출된 수취인 주소 영상을 사용하였다. 앞에서 언급했듯이, 200dpi의 저해상도 영상이며 이진화 과정을 거친 영상이다. 표 1은 실험에 사용된 데이타를 요약하고 있다.

표 1 실험 데이타

	훈련 영상의 개수	테스트 영상의 개수
인쇄체	4,775	2,372
필기체	4,253	775
총합	9,028	3,147

다층퍼셉트론의 훈련을 위하여 총 9,028개의 주소 영상을 사용하였으며, 성능 테스트를 위하여 총 3,147개의 영상을 사용하였다. 테스트 영상 집합을 구성할 때 인쇄체 영상과 필기체 영상의 비율을 대략 3:1로 하였다. 이 비율은 실제 하루 처리되는 우편물량에서 인쇄체 우편물량과 필기체 우편물량의 비율을 반영한 것이다.

4.2 성능 평가

폭 크기 히스토그램, 폭 분산 히스토그램, 위치 히스토그램 특징 각각에 대한 성능 평가를 실시하였다. 이와 더불어, 특징을 결합하였을 때 분류 성능의 변화를 조사하였다. 그림 6은 그 결과를 보여주고 있다.

개개의 특징을 살펴보면, 폭 크기 히스토그램 특징과 위치 히스토그램 특징이 96% 정도의 비슷한 구분 성공률을 보여준다. 폭 분산 히스토그램의 특징이 폭 크기 히스토그램 특징보다 낮은 성능을 보이고 있다. 이는 크기에 대한 정보가 분산에 대한 정보보다 더 중요함을 의미한다. 그 이유는 주소 영상에 기입된 인쇄체 문자들의 크기가 필기체 문자들의 크기보다 상대적으로 작기 때문인 것으로 조사되었다. 두 개의 특징을 결합한 경우의 성능을 살펴보면, 폭과 위치 정보가 모두 이용될 때 그 성능이 우수함을 알 수 있었다. 마지막으로, 모든 특징이 사용될 때 가장 높은 구분 성공률을 얻을 수 있었다. 이때, 구분 성공률은 98.95%이다.

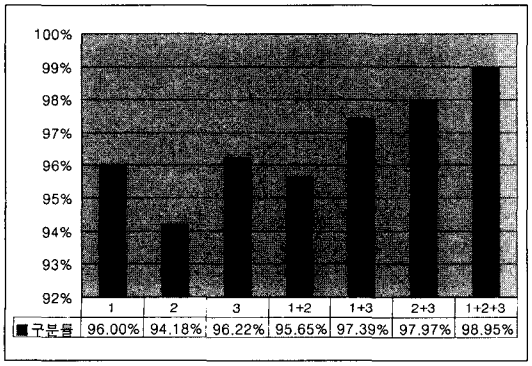


그림 6 특징에 따른 주소 기입 형식 구분 방법의 성능 변화

4.3 오류 분석

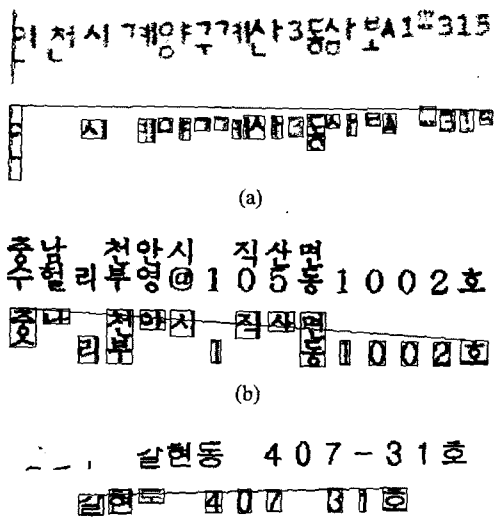
총 3,147개의 테스트 영상 중 33개의 주소 영상에 대하여 구분을 실패하였다. 실패의 형태는 두 종류로 나뉠 수 있다. 첫 번째 형태는 입력 영상이 인쇄체일 때 필기체로 분류되는 경우이며, 이를 Type I 오류로 정의하였다. 두 번째 형태는, 입력 영상이 필기체일 때 인쇄체로 분류되는 경우이며, 이를 Type II 오류로 정의하였다. 표 2는 제안 방법의 오류를 분석한 결과이다. 주소 인식 시스템 차원에서는 후자의 오류가 전자의 오류보다 더 중요하므로 후자의 오류를 낮게 출력하는 주소 기입 형식 구분 방법이 더 우수하다고 평가하게 된다. 그 이유는 인쇄체 영상을 대상으로 개발된 문자열 및 문자 분리 그리고 문자 인식 알고리즘에 필기체 영상이 들어오는 경우, 반대의 경우에 비해 상대적으로 성공을 기대하기 어렵기 때문이다. 본 실험에서 Type II 오류로 분류된 영상은 12개의 영상으로 단순 수치로 보면 작으나, 테스트 집합의 인쇄체와 필기체 영상의 비율을 고려하면 Type I 오류와 거의 비슷함을 알 수 있다. 향후 연구에서는 Type II 오류를 줄이기 위하여 분류에 앞서 수행될 기각방법을 연구하고자 한다.

표 2 Type I 오류와 Type II 오류

오류의 형태	오류 영상의 개수
Type I 오류	21
Type II 오류	12
총합	33

4.3.1 Type I 오류 유형

총 21개의 Type I 오류 영상을 분석한 결과, 크게 세 종류의 오류 형태로 요약할 수 있었다. 그림 7은 각 오류 형태의 대표적인 영상을 보여주고 있다. 그림 7의 (a)와 같은 오류는 창봉투의 경계에서 발생하는 긴 수직획의 잡영이 최상위 문자열의 첫 번째 문자와 겹치면서 발생한다. 이는 창봉투의 경계에서 발생하는 긴 수직획 잡영을 제거함으로써 해결될 수 있다. 그림 7의 (b)와 같은 오류는 특징 추출을 위한 연결요소 추출과정에서 최상위 문자열에 속하지 않는 연결요소들이 접촉이나 기울어짐에 의해 제거되지 않았을 때 발생한다. 이는 특징 추출을 위한 연결요소 추출 알고리즘의 개선에 의해 해결될 수 있다. 그림 7의 (c)와 같은 오류는 영상이 훼손된 경우 발생한다.

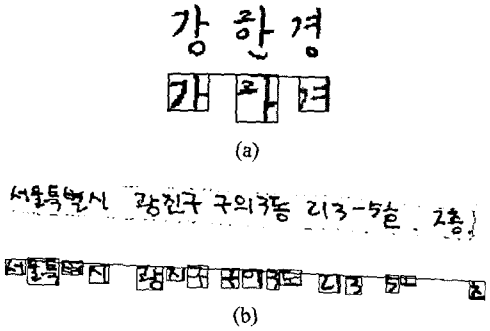


(a) (b) (c)

그림 7 인쇄체 오류 영상

4.3.2 Type II 오류 유형

총 12개의 Type II 오류 영상을 분석한 결과, 두 종류의 오류 형태로 요약할 수 있었다. 그림 8은 오류 형태별로 대표적인 영상을 보여주고 있다. 그림 8의 (a)와 같은 오류는 특징 추출을 위한 정보가 절대적으로 부족한 경우에 발생한다. 이를 개선하기 위해서 최상위 문자열이 아닌, 주소 영상에서 구분 정보가 가장 많은 문자열을 선택하는 방법이 필요하다. 그러나, 이러한 문자열을 추출하는 것은 문자열 분리와 같은 문제로 필기체 영상에서는 매우 어려운 문제 중 하나이다. 그림 8의 (b)와 같은 오류는 최소인접사각형의 폭과 위치가 인쇄체 영상과 비슷할 때 발생한다. 이를 개선하기 위해서 최소인접사각형 간의 수직/수평 성분들의 관계를 반영하는 특징의 추가가 필요하다.



(a) (b)

그림 8 필기체 오류 영상

5. 결론 및 향후 연구

본 논문에서는 한글 주소 영상이 인쇄체인지 필기체인지를 구별해 주는 방법을 제안하였다. 제안된 방법은 인쇄체 문자의 폭 및 위치의 규칙성을 필기체와 구분되는 특징으로 사용하였고, 인쇄체와 필기체를 구분하기 위하여 신경망을 사용하였다. 200dpi의 해상도로 스캔된 총 3,147개의 실제 우편 영상에서 수취인 주소 영상을 수작업을 통해 추출한 뒤 이를 이용하여 성능을 테스트한 결과 98.95%의 구분률을 얻을 수 있었다. 향후 연구에서는 입력 영상이 인쇄체일 때 필기체로 구분되는 오류(Type II)를 줄이기 위하여 기각 알고리즘 및 수직/수평의 관계에 관한 특징을 개발할 예정이며 특징 추출을 위한 연결요소 추출 방법을 개선하여 전체 오류를 줄이고자 한다.

참 고 문 헌

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

[2] U. Mahadevan and S.N. Srihari, "Parsing and Recognition of City, State, and ZIP Codes in Handwritten Addresses," Proceedings of 5th International Conference on Document Analysis and Recognition, pp. 325 -328, Bangalore, India, 1999.

[3] G. Dzuba, A. Filatov and A. Volgunin, "Handwritten ZIP Code Recognition," Proceedings of 4th International Conference on Document Analysis and Recognition, pp. 766-770, Ulm, Germany, 1997.

[4] A. Brakensiek, J. Rottland, G. Rigall, "Handwritten Address Recognition with Open Vocabulary Using Character N-grams," Proceedings of 8th International Workshop on Frontiers in Handwriting Recognition, pp. 357-362, Niagara-on-the-Lake, Canada, 2002.

[5] N. Kata, K. Todumoto and Y. Nemoto, "A Large Scale Japanese Handwritten Address Recognition System Using Rough and Fine Classification," Proceedings of 5th International Conference on Signal Processing, Vol. 3, pp. 1423-1426, 2000.

[6] F. Kimura and M. Shridhar, "Handwritten Address Interpretation Using Extended Lexicon Word Matching," Proceedings of 5th International Workshop on Frontiers in Handwriting Recognition, pp. 369-372, Essex, England, 1996.

[7] K. Fan, L. Wang and Y. Tu, "Classification of Machine-Printed and Hand-Written Texts Using Character Block Layout Variance," Pattern Recognition, Vol. 31, No. 9, pp. 1275-1284, 1998.

[8] U. Pal and B. Chaudhuri, "Automatic Separation of Machine-Printed and Hand-Written Text Lines," Proceedings of 5th International Conference on Document Analysis and Recognition, pp. 645-

648, Bangalore, India, 1999.
[9] K. Kuhnke, L. Simonicini and Z. Kovacs-V, "A System for Machine-Written and Hand-Written Character Distinction," Proceedings of 3rd International Conference on Document Analysis and Recognition, pp. 811-814, 1995.
[10] S. Image, S. Tatsuta and T. Wada, "Segmentation and Classification for Mixed Text/Image Documents Using Neural Network," Proceedings of 2nd International Conference on Document Analysis and Recognition, pp. 930-934, 1993.
[11] J. Franke and M. Oberlandwr, "Writing Style Detection by Statistical Combination of Classifiers in Form Reader Applications," Proceedings of 2nd International Conference on Document Analysis and Recognition, pp. 581-584, 1993.
[12] S. Tsujimoto and H. Asada, "Major Components of a Complete Text Reading System," Proceeding IEEE, Vol. 80, No. 7, pp. 1133-1149, 1992.
[13] S. Jeong, S. Kim and W. Cho, "Performance Comparison of Statistical and Neural Network Classifiers in Handwritten Digits Recognition," Proceedings of 6th International Workshop on Frontiers in Handwriting Recognition, pp. 419-428, Taejon, Korea, 1998.
[14] N. Otsu, "A threshold selection method from gray-level histogram," IEEE Transactions on Systems, Man, and Cybernetics, Vol. 9, pp. 62-66, 1979.

장 승 익

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

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

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