인쇄 문서 영상의 단어 단위 속성 인식
(Recognition of Word-level Attributes in Machine-printed Document Images)

곽 희 규 † 김 수 형 ††
(Heekue Kwag) (Soohyung Kim)

요 약 본 논문은 문서 영상에 존재하는 개별 단어들에 대한 속성정보 추출 방법을 제안한다. 단어 단위의 속성 인식은 단어 영상 매칭의 정확도 및 속도 개선, OCR 시스템에서 인식률 향상, 문서의 재생산 등 다양한 응용 가치를 찾을 수 있으며, 메타정보(meta-information) 추출을 통해 영상 검색(image retrieval)이나 요약(summary) 생성 등에 활용할 수 있다. 제안하는 시스템에서 고려하는 단어 영상의 속성은 언어의 종류(한글, 영문), 스타일(볼드, 이탤릭, 보통, 밑줄), 문자 크기(10, 12, 14포인트), 문자 개수(한글: 2, 3, 4, 5, 영문: 4, 5, 6, 7, 8, 9, 10), 서체(명조, 고딕)의 다섯 가지 정보이다. 속성 인식을 위한 특징은, 언어 종류 인식에 2개, 스타일 인식에 3개, 문자 크기와 개수는 각각 1개, 한글 서체 인식은 1개, 영문 서체 인식은 2개를 사용한다. 분류기는 신경망, 2차형 판별함수(QDF), 선형 판별함수(LDF)를 계층적으로 구성한다. 다섯 가지 속성이 조합된 26,400개의 단어 영상을 사용한 실험을 통해, 제안된 방법이 소수의 특징만으로도 우수한 속성 인식 성능을 보임을 입증하였다.

Abstract In this paper, we propose a method to extract attribute information of individual words in machine-printed document images. The recognition of the word-level attributes helps in improving the accuracy and speed of the word image matching, the OCR performance, and the reproduction of the document. It is also useful in retrieving document image and creating summary, by an extraction of some meta-information from printed documents. The attributes of word images considered in the proposed system are language(Korean, English), style(bold, italic, regular, underline), size(10, 12, 14 points), the number of characters(Korean: 2, 3, 4, 5, English: 4, 5, 6, 7, 8, 9, 10), and typeface(Myung-jo, Gothic). Two features are used for the language classification, three for the style, one for the size, one for the number of characters, one for the Korean typeface, and two for the English typeface classification, respectively. The classifiers are constructed hierarchically using the neural network, quadratic discriminant function, and linear discriminant function. An experiment using 26,400 word images combining the five attributes demonstrates how the proposed method shows high performance of attribute recognition with only a small number of features.

1. 서 론

사람들은 흔히 접할 수 있는 신문, 잡지, 책, 계산서, 전표에서 보고서, 저널, 공문서에 이르기까지 수많은 인쇄문서(machine-printed document)로부터 정보를 얻고 있다. 최근 사무자동화에 의한 전자문서(electronic document)의 사용이 보편화되었음에도 불구하고, 인쇄문서에 대한 사람들의 여전한 선호 때문에 인쇄문서의 발생은 날로 증가하고 있다. 따라서 인쇄문서에 대한 자동 정보추출(automatic information extraction)에 대한 요구는 당연한 것이라 하겠다. 실제로, 전자도서관(digital library), 광파일(optical filing) 시스템 등과 같이 대규모 인쇄문서를 자동으로 저장, 가공, 검색, 재생산해야 하는 응용분야들이 늘고 있으며, 이에 대한 연구도 지속적으로 이루어지고 있다[1].

대용량 인쇄문서를 자동으로 검색하는 방법에는, 광학

* 본 연구는 한국과학재단 특정기초연구(98-0102-02-01-3) 지원으로 수행되었음
† 비 회 원 : 한국과학기술원 전산학과
hkkwag@ai.kaist.ac.kr
†† 종신회원 : 전남대학교 컴퓨터정보학부 교수
shkim@chonnam.ac.kr
논문접수 : 2000년 11월 1일
심사완료 : 2001년 4월 2일

문자인식(OCR: Optical Character Recognition) 기술을 적용하여 문서 영상을 컴퓨터 코드로 변환하는 방법과 문서 영상을 단어 단위 영상으로 분할한 후 임의의 검색어를 포함하는 영상을 탐색하는 주제어 검색(keyword spotting) 방법이 있다. OCR 기반 접근 방법은 인식 결과가 완전하지 못하고 색인 과정이 느리다는 제약이 있는 반면, 주제어 검색 방법은 검색 정확도가 높고 색인이 간편하지만 많은 검색 시간이 소요된다[2].

과거의 문자인식 및 주제어 검색에 관련된 연구들을 살펴보면, 문서에 존재하는 문자 속성의 인식에 대해서는 시스템의 성능에 별다른 영향을 미치지 않는 것으로 생각하여 연구 대상에서 제외되었다[3]. 특히, 문자인식 기술의 발전 추세를 보면, 문자의 속성에 무관한 문자인식 기술(omni-OCR)을 개발하는 방향으로 발전해 왔던 것이 사실이다. 최근에서야 비로소 문자가 가지고 있는 속성 정보의 유용성을 고려하여 문자의 속성을 인식(OFR: Optical Font Recognition)하는 연구가 활발히 진행되고 있고, OCR과의 결합을 통한 성능 향상을 꾀하고 있다[3-14]. 그러나 아직도 문서 영상 검색시스템(document image retrieval system) 등에 활용하기 위한 단어 단위의 속성 정보 추출에 대한 연구는 다양한 응용 가치를 찾을 수 있음에도 불구하고 실제적인 연구가 이루어지지 않고 있다.

본 논문에서는 광학 문자인식 시스템 및 주제어 검색 시스템의 성능 향상에 활용하기 위한 단어 단위의 속성 정보 인식 시스템을 제안한다. 즉, 문서 영상을 단어 단위로 분할한 후, 분할된 단어 영상들이 가지고 있는 다양한 속성 정보들을 추출하고자 한다. 단어 단위의 속성 정보 추출이 갖는 유용성은 다음과 같은 것이 있다[3-12]. 첫째, 단어 단위 영상 매칭의 정확도와 속도를 높일 수 있다. 이것은 속성 정보를 인식 과정에 활용하고, 속성 인식을 위해 추출한 특징들을 인식 과정에서 사용함으로써 가능하다. 둘째, OCR 시스템에의 활용에서 문자의 인식률 향상에 기여할 수 있다. 문자의 속성 정보를 이용하여 속성별로 간소화된 인식기(mono-font OCR)를 구축할 수 있다. 셋째, 문서 영상의 재생산(reproduction)의 요구에 대해 본래의 문서와 동일한 속성을 가진 문서를 생성할 수 있다. 넷째, 문서 영상으로부터 주제어를 검색할 때, 탐색 공간(search space)을 현저하게 줄일 수 있다. 일반적으로 문서에서 특정 부분을 강조하거나 중요한 부분은 서로 다른 속성을 사용한다. 이것은 문서의 제목 등에 나타나는 볼드(bold), 이탤릭(italic) 스타일을 예로 생각할 수 있다. 따라서 주제어 검색에 대해 유사한 속성을 지닌 단어 영상만을 탐색할 수 있다. 다섯째, 문서 영상의 내용에 대한 요약(summary)을 생성할 때 유용하다. 보통의 텍스트보다는 특정 속성을 가진 단어들의 리스트를 유지하여 쉽게 요약을 생성할 수 있다.

제안하는 시스템에서 고려하는 단어의 속성 정보는, 언어(language)의 종류, 스타일(style), 문자의 크기(size), 문자의 개수, 서체(typeface)의 다섯 가지이다. 이러한 속성들은 훈글 97 환경에서 생성하여 출력과 스캐닝을 통해 문서 영상을 얻은 후, 단어 분할 과정에서 개별적인 단어 영상을 획득하였다. 단어 영상의 속성 정보 분류를 위한 분류기는 신경망(neural network), 2차형 판별함수(QDF: Quadratic Discriminant Function), 선형 판별함수(LDF: Linear Discriminant Function)를 계층적으로 구성한다. 먼저 입력 단어 영상으로부터 언어의 종류를 분류하기 위한 특징을 추출하고, 훈련된 신경망 분류기를 통해 한글 또는 영문의 두 부류 중 하나로 분류한다. 그리고 각 부류별로 서로 다른 특징 추출 과정을 통해 특징을 추출하고, 훈련된 QDF를 통해 문자의 크기, 스타일을 분류하며, 문자의 개수를 분류한다. 서체 분류는 언어의 종류, 스타일별로 개별적인 4개의 서체 분류 LDF를 구축하여 수행한다. 계층적 분류 모델에서 분류기는 각 특징에 대해 가장 높은 정확도를 형성하는 분류기를 채택하였다.

본 논문의 구성은 다음과 같다. 2장에서는 텍스트의 속성 정보 추출에 대한 관련 연구들에 대해 설명한다. 3장에서는 속성 인식시스템의 계층적 모델과 특징 추출에 대해 자세히 기술한다. 4장에서는 실험 결과 및 분석에 대해, 마지막으로 5장에서는 결론 및 향후 연구 과제에 대해 언급한다.

2. 관련 연구

앞서 언급한 바와 같은 속성 정보의 유용성을 고려하여 문자의 속성을 인식하는 연구(OFR: Optical Font Recognition)[3]나 다양한 속성 정보로부터 문서 영상의 메타정보를 추출하는 연구[12]가 발표되고 있다. 이러한 메타정보는 문서 영상으로부터 중요한 단어나 요약 등의 검색에 대해 탐색공간을 매우 감소시키는 효과를 가진다.

Zramdini[3]는 텍스트의 폰트 인식이 문서 구조 분석 및 이해, OCR 성능 개선과 문서 영상의 재생산(reprinting) 등에 매우 유용하다는 것을 강조하여 문서 영상에 존재하는 다양한 폰트 인식에 대한 통계적 접근 방법을 제안하였다. 제안한 *ApOFIS(A priori Optical Font Identification System)*에서는 사전 폰트 인식(*a*

priori font recognition) 방법을 채택하는데, 먼저 폰트 인식 방법(omni-char OFR)을 텍스트 영상에 적용하여 폰트를 식별하고, 각 폰트별로 문자 인식기(mono-font OCR)를 구축하는 모델을 제시하고 있다. 폰트 인식을 위한 전역적(global) 특징은 텍스트의 가중치(weight), 크기(size), 방향(orientation)과 간격(spacing) 등으로부터 추출하여 사용하고, 클래스 분류에는 Bayesian 분류기를 사용한다.

박문호 등[4]은 한글 문서를 대상으로 문서를 구성하는 문자의 서체와 문자의 크기 및 기울기를 인식하는 방법을 제안하였다. 서체를 인식하기 위하여 일정한 크기의 블록을 추출하여 주파수 분석을 수행하였고, 단어의 외접 사각형의 수직거리를 이용하여 문자의 크기를, 수직 방향의 투영 프로파일을 이용하여 문자의 기울기를 인식하였다. 서체 인식을 위한 인식기는 오류 역전파 알고리즘으로 학습된 MLP(Multi-layer Perceptron)를 사용하였고, 문자의 크기와 기울기를 분류하기 위하여 Mahalanobis 거리를 이용하였다. 실험에서 10개의 서체에 대하여 95.19%, 5가지 문자 크기에 대하여 97.34%, 문자의 기울기는 평균 89.09%의 인식률을 얻었다.

Garain과 Chaudhuri[12]는 OCR 접근 방법을 사용하지 않고 문서 영상으로부터 메타 정보들을 추출하는 연구를 제안하였다. 이 논문에서는 문서에서 중요한 단어들이 이탤릭(italic), 볼드(bold), 대문자(capital) 스타일을 가진다는 것을 통계적으로 검증하였고, 이러한 스타일 검출이 중요한 단어들을 포함하는 문장들뿐만 아니라 제목, 저자, 부제목, 참고문헌 등을 포함하는 텍스트 라인의 자동 추출에 매우 유용하다는 것을 실험을 통해 입증하였다. 그리고 스타일을 가진 단어들을 포함하는 문장들을 이용하여 문서의 요약을 자동으로 추출하고, 단어 인식의 과다한 계산 비용을 절감할 수 있음을 보였다.

3. 단어 단위 속성 인식 시스템

3.1 시스템 개요

단어 단위 속성 인식 시스템은 광학 문자인식 및 주제어 검색 시스템의 성능 향상에 활용하기 위한 것으로, 문서 영상을 단어 단위로 분할한 후, 분할된 단어 영상들을 입력으로 다양한 속성 정보들을 추출한다. 지금까지의 연구들이 적용 대상을 한글 또는 영문 문서만으로 제한하였다면, 제안하는 방법은 한글과 영문이라는 언어의 종류에 무관한 특징을 가진다. 단어 단위 속성 인식 과정에서는 문서 영상에 존재하는 단어들의 다양한 속성을 분류하기 위해 특징들을 설계하고, 이러한 특징들로 훈련된 분류기를 계층적으로 구성한다.

표 1 단어 영상의 다섯 가지 속성 조합

언어 종류	문자 크기	서체	스타일	문자 개수	계
한글	3클래스 (10,12,14 pts)	2클래스 (명조, 고딕)	4클래스 (볼드,이텔릭 보통,밑줄)	4클래스 (2,3,4,5개)	3×2×4×4 = 96클래스
영문	3클래스 (10,12,14 pts)	2클래스 (명조, 고딕)	4클래스 (볼드,이텔릭 보통,밑줄)	7클래스 (4,5,6,7,8 ,9,10개)	3×2×4×7 = 168클래스
계					264클래스

문서에 존재하는 단어들은 다양한 속성들을 가지고 있고, 이러한 속성들은 다양한 조합 형태로 나타난다. 본 연구에서는 표 1에 나타낸 것처럼 다섯 가지의 속성을 조합한 총 264 클래스(한글: 96클래스, 영문: 168클래스)를 인식 대상으로 고려한다. 따라서 임의의 단어 영상은 표 1에서 고려한 264클래스 중 하나의 클래스에 해당된다고 가정한다. 다섯 가지 속성을 가진 단어 영상의 예가 그림 1에 제시되어 있다.

단어 영상	언어 종류	문자 크기	스타일	문자 개수	서체
상류사회의	한글	10	이텔릭	5	명조
antipathic	영문	10	볼드	10	고딕

그림 1 영문과 한글 단어 영상이 가진 속성

단어 영상의 속성 인식을 위해 먼저 특징을 추출하고, 그 특징으로 훈련된 분류기를 구축한다. 따라서 속성 분류를 위해 사용하는 특징과 분류기의 종류를 결정하는 것이 본 연구의 주안점이다. 제안방법에서는 단어의 내용에 무관한 전역적(global) 특징을 추출하는데, 이것은 단어 영상에서 나타나는 문자들의 동일한 성질을 구함으로써 가능하다. 따라서 전역적 특징은 단어의 길이, 즉 문자의 개수에 영향을 받기도 하는데 단어가 많은 문자를 포함할수록 속성의 변별력이 크기 때문이다[3].

속성 분류를 위한 분류기는 신경망(neural network), 2차형 판별함수(QDF: Quadratic Discriminant Func-

tion), 선형 판별함수(LDF: Linear Discriminant Function)를 사용하는데, 다양한 실험을 통해 해당 속성에 최적의 인식률을 보이는 것을 채택하였다. 제안방법에서는 이러한 분류기들을 계층적으로 구성하는데, 이것은 분할 해결법(divide-and-conquer)을 의미한다. 즉, 입력 단어 영상의 속성을 264클래스 중 하나로 분류하는 문제를 언어의 종류를 분류하는 2클래스 문제, 스타일을 분류하는 4클래스 문제 등의 하부문제(sub-problem)들로 나누어 해결하고, 결과를 합쳐서 단어 영상의 속성을 분류하는 방법이다. 264클래스 분류 문제는 많은 특징을 필요로 하고 복잡한 분류기가 필요한 반면, 계층적 모델에서는 보다 적은 수의 특징과 간소화된 분류기를 채택할 수 있다. 각 속성 분류를 위한 분류기의 계층적 모델은 그림 2에 나타나 있다.

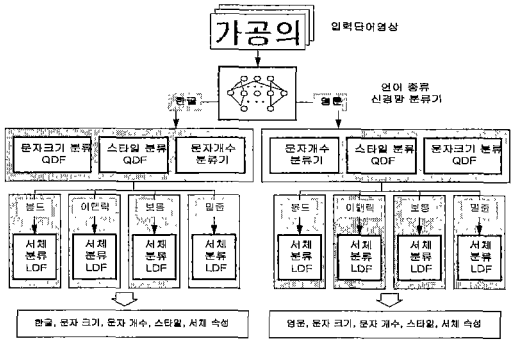

그림 2 속성 분류를 위한 분류기의 계층적 모델

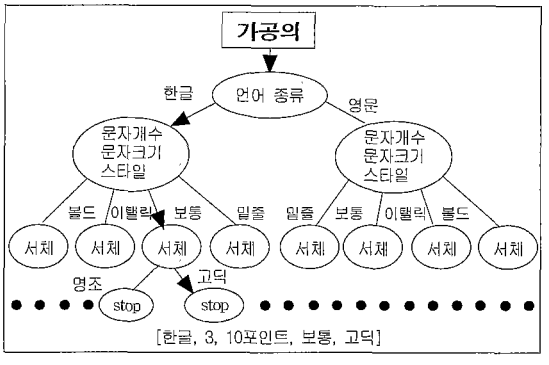

그림 3 단어 영상 속성 분류를 위한 결정 트리

그림 3에서 보듯이, 그림 2의 계층적 모델은 결정 트리(decision tree) 형태를 취하고 있다. 먼저, 단어 영상에 대한 언어 종류를 분류하는 특징을 추출하고 훈련된 신경망 분류기에 의해 한글 또는 영문으로 분류한다. 다음으로, 한글 또는 영문에 해당하는 문자 개수, 스타일, 문자 크기 분류를 위한 특징을 추출하고, 언어의 종류에 종속적인 분류기에 의해 이들 세 가지 속성들을 인식한다. 마지막으로, 한글 또는 영문 속성에 따라 서체 분류를 위한 특징을 추출하여 문자 크기와 개수에 무관하고 스타일 종류에 따라 구축된 분류기에 의해 해당 서체를 인식한다. 다음 각 절에서 제안방법의 속성 분류를 위한 특징 및 분류기를 자세히 설명한다.

3.2 언어 종류 분류

언어 종류 인식(language identification)은 한글과 영문을 병행 사용하는 국내 환경에서는 매우 중요한 문제이다. 즉, 언어 종류 인식은 단어 영상에 대한 적합한 문자인식 알고리즘 선택과 색인(indexing) 등의 관점에서 그 필요성을 찾을 수 있다.

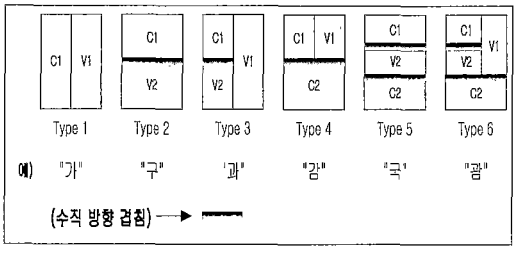

그림 4 한글 초성, 중성, 종성의 결합 형태

제안하는 시스템에서는 단어 영상의 한글 또는 영문 속성 인식을 수행하고, 분류를 위한 특징은 두 가지를 사용한다. 특징 추출을 위해 먼저 단어 영상에 대한 연결요소(connected component) 분석을 수행한다. 첫 번째 특징은 단어 영상의 폭(width) 정보에 대한 연결요소 개수의 비율이다. 대부분의 영문 문자는 하나의 연결요소('i', 'j' 제외)로 표현된다. 그러나 한글의 한 문자는 초성, 중성, 종성이 결합된 여섯 가지 형태중 하나이다(그림 4). 이것은 한 문자가 두 개 이상의 연결요소로 이루어진다는 것을 의미한다. 따라서 영문 단어 영상은 1에 가까운 값을, 한글 단어 영상은 2에 가까운 값을 가진다.

두 번째 특징은 단어 영상의 폭 정보에 대한 연결요소들의 수직 방향 겹침 비율이다. 영문은 겹침이 발생하지 않지만, 한글은 그림 4(Type 2~6)의 결합 형태에서 보듯이 수직 방향으로 일정량의 겹침이 발생한다. 이것은 연결요소의 최 외곽사각형(BB: Bounding Box)을 구하여 BB의 겹침 정도로 구할 수 있다. 그림 5에서 한글 및 영문 단어 영상에 대한 두 특징 값을 나타내고 있다.

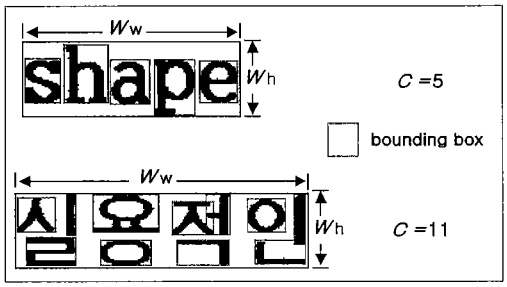

그림 5 한글 및 영문의 연결요소와 BB

▫ 특징 ① : 연결요소의 비율 = $C / (W_w / W_h)$

(C : 연결요소의 개수, W_w : 단어의 폭, W_h : 단어의 높이)

▫ 특징 ② : 연결요소의 수직 방향 겹침 비율 = $C_{overlapped} / (W_w / W_h)$

$$C_{overlapped} = \sum_i \sum_j \text{right} - \text{left} \quad (i \neq j)$$

$$\text{right} = MIN\{BB_i.x_{max}, BB_j.x_{max}\},$$

$$\text{left} = MAX\{BB_i.x_{min}, BB_j.x_{min}\},$$

BB_i : i 번째 연결요소의 BB

언어 종류 분류는 오류 역전파 알고리즘(back-propagation algorithm)에 의해 학습된 다층 퍼셉트론(MLP: Multi-layer Perceptron)을 사용하는데, 입력층은 2개, 은익층은 4개, 출력층은 2개의 뉴런으로 구성한다.

3.3 스타일 분류

흔히 문서에서 제목, 저자, 부제목, 참고문헌 등 강조하고 싶은 특정 부분에 스타일 속성이 포함된다. 제안시스템에서는 단어 영상의 스타일 속성으로 볼드(bold), 이탤릭(italic), 보통(regular), 밑줄(underlined) 속성의 인식을 목표로 한다. 속성 분류를 위한 세 가지 특징은 다음과 같이 추출한다.

첫 번째 특징은 볼드 속성을 분류하기 위한 특징으로, 수직 획(stroke)의 평균 폭 정보를 추출한다. 수직 획의 폭 정보는 단어 영상에 존재하는 수평 런 길이(run length)중 빈도수가 가장 높은 런 길이로부터 추출한다. 수평 런 길이에 대한 분포를 조사하면, 최대 빈도 값을 갖는 수평 런 길이 HRL_{max_freq}을 구할 수 있다. 그리고 특징 HRL_{avg}은 다음과 같이 수평 런 길이 HRL_{max_freq}, $HRL_{max_freq} \pm 1$와 해당 런 길이의 빈도 값을 곱하여 평균을 취함으로써 구한다.

▫ 특징 ① : $HRL_{avg} = \frac{1}{N} \sum_{i=HRL_{max\_freq}-1}^{HRL_{max\_freq}+1} H[i] \times i$

$H[i]$ (수평 런 길이 i의 빈도 값)

$$N = \sum_{i=HRL_{max\_freq}-1}^{HRL_{max\_freq}+1} H[i] \quad \text{(수평 런의 개수)}$$

두 번째 특징은 밑줄 속성을 분류하기 위한 특징으로, 단어 영상의 하단 부분에 존재하는 최대 수평 런 길이를 구하여 단어의 폭 정보로 나눈 값으로 결정한다. 영상의 하단 부분에 존재하는 수평 런 길이를 $HRL_{,,}$ 단어의 폭 정보를 W_w라고 하면, 특징 HRL_{max}은 다음과 같이 구한다.

▫ 특징 ② : $HRL_{max} = MAX\{ HRL_i \} / W_w$

세 번째 특징은 이탤릭 속성을 분류하기 위한 특징으로, 먼저 단어 영상에 대한 수직 투영 프로파일을 구한 후, 프로파일 값을 1차 미분한다. 미분 값은 이웃하는 투영 프로파일 값들 간의 차이(difference)로 구하고, 특징 값은 1차 미분 최대 값으로 결정한다. 이탤릭의 속성을 가진 경우, DVP_{max} 값은 매우 낮은 값을 가진다. 그림 6에는 단어 영상에 대한 수직 투영 프로파일과 1차 미분한 결과를 보여준다.

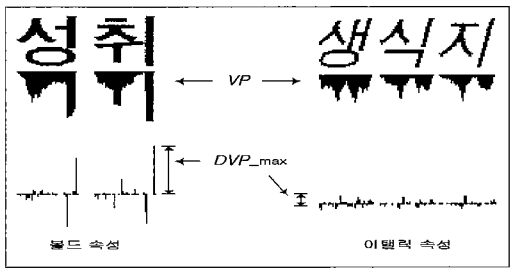

그림 6 단어 영상의 수직 투영 프로파일과 1차 미분

▫ 특징 ③ : $DVP_{max} = MAX\{ VP[i] - VP[i+1] \}$

$VP[i]$: i 번째 수직 프로젝션 프로파일

스타일 분류를 위한 분류기는 언어의 종류에 따라 개별적으로 학습된 2차형 판별함수를 구축하여 사후확률을 최대로 하는 스타일 부류를 결정한다.

3.4 문자 크기 분류

단어 영상의 크기를 인식하기 위한 특징으로 영상의 수직 거리를 이용하는데, 언어의 종류에 따라 추출하는 방법이 다르다. 한글의 경우, 단어를 구성하는 문자는 여섯 가지 형태(그림 4) 중 하나이기 때문에 개별 문자를 고려하기보다는, 단어 영상을 외접하는 사각형의 수직 거리를 특징으로 이용한다. 영문의 경우, 텍스트라인은 3등분되어 상위(upper), 하위(lower), 중심지역(middle zone)으로 나뉘지만, 중심지역에 집중적으로 분

포한다. 따라서 단어 영상으로부터 중심지역을 추출하여 수직거리를 특징 값으로 사용한다. 그림 7은 한글 및 영문 단어 영상에 대한 수직 거리 특징 값을 나타내고 있다. 영문에서 중심지역은 단어 영상의 수평 방향 투영 프로파일(horizontal projection profile)을 구한 후, 두 개의 최대 peak가 나타나는 지점에서 산출한다.

그림 7 한글 및 영문 단어 영상에 대한 수직 거리

제안 시스템에서는 한글, 영문 모두 10, 12, 14 포인트 크기를 고려하며, 언어의 종류에 따라 개별적으로 학습된 2차형 판별함수를 구축하여 사후확률을 최대로 하는 크기 부류에 속성을 분류한다.

3.5 문자 개수 분류

단어 영상을 구성하는 문자의 개수를 인식하기 위해 하나의 특징을 사용하며, 한글은 문자 개수 2, 3, 4, 5의 4클래스를, 영문은 4, 5, 6, 7, 8, 9, 10의 7클래스를 고려한다. 분류를 위한 특징 값은, 한글의 경우 문자의 종횡비가 1:1의 비율을 가지기 때문에 단어 영상의 높이에 대한 폭의 비율로 산출한다. 반면, 영문은 하나의 연결요소가 한 문자를 나타내기 때문에 연결요소의 개수로 산출한다. 물론 영문의 'i'와 'j'는 연결요소의 크기 분석을 통해 하나의 연결요소로 간주한다.

3.6 서체 분류

문서에서 서체의 변화는 흔히 발견할 수 있는데, 일반적으로 특정 부분을 강조하고 싶을 때 서로 다른 서체를 사용한다. 훈글 97의 환경에서는 일반적으로 명조와 고딕 계열의 서체를 가장 많이 사용하기 때문에 본 논문의 서체 분류는 단어 영상이 가지고 있는 명조와 고딕 속성을 인식하는 것이다. 따라서 영문 서체에 대한 명조와 고딕은 훈글 97의 폰트를 기준으로 가정한 것이다. 단어 영상의 서체를 인식하기 위한 특징은 언어의 종류에 따라 다른 특징을 사용한다. 서체 인식을 위한 분류기는, 분류기의 계층적 모델(그림 2)에서 제시한 것처럼 언어의 종류, 스타일에 따라 학습된 개별적인 8개의 선형 판별함수를 구축하여 사후확률을 최대로 하는 서체 부류를 결정한다.

서체 변화의 두드러진 특징은 세리프(serif)의 변화를 통해 감지할 수 있다. 한글의 경우, 명조는 문서에서 가장 많이 쓰이는 것으로 획의 끝 부분이 세리프로 장식되어 있고, 획의 폭이 균일하지 않은 것이 특징이다. 반면, 고딕은 특정 부분을 강조하는 경우에 사용하고, 세리프가 존재하지 않으며 획의 폭이 균일한 것이 특징이다. 영문의 경우, 명조와 고딕 모두 세리프가 존재하며 모양이 서로 다르다. 따라서 세리프의 형태로부터 추출한 특징을 사용해야 한다. 명조는 세리프의 존재가 불분명할 정도로 폭이 좁고 영역이 작은 반면, 고딕은 세리프의 형태가 두드러져서 폭이 넓고 영역이 크다. 그림 8에는 한글과 영문 문자에 대한 세리프 영역이 예시되어 있다.

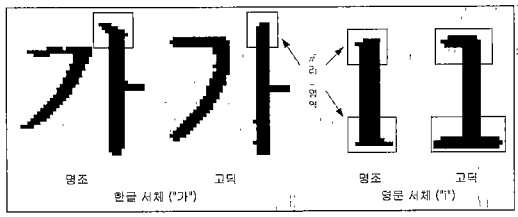

그림 8 한글 및 영문의 세리프 영역

한글의 서체 인식을 위한 특징은, 단어 영상에 존재하는 세리프 영역을 추출하고 각 영역에 존재하는 수평 런(run)들의 평균 방향 벡터를 구한 후, 가리키는 방향을 36사분면 중 하나의 정수 값으로 결정한다. 수평 런의 방향 벡터는 가장 상위 런의 시작점과 하위 런들의 시작점을 연결한 벡터이고, 이러한 벡터들의 합으로 계산된다.

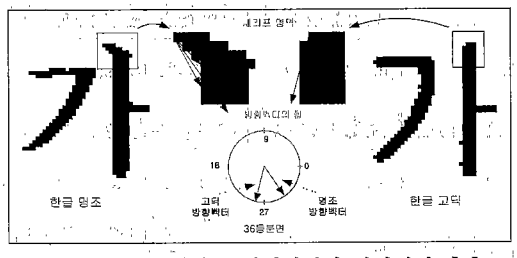

그림 9 한글 세리프 영역에서의 방향벡터 추출

그림 9는 한글의 세리프 영역에서 추출한 고딕과 명조에 대한 방향 벡터와 36등분면 상의 위치를 나타내고 있다. 대개, 고딕은 27이하의 평균값을, 명조는 27보다 큰 평균값을 갖는다. 영문의 서체 인식을 위한 특징은 두 가지를 사용하는데, 첫 번째는 세리프 영역에 존재하는 수평 런들의 평균 길이이고, 두 번째는 블록(block)들의 평균 면적이다. 수평 런의 길이는 런 길이를 가장

높은 빈도의 수평 런 길이 HRL_{max_freq}로 나누어 문자의 크기에 무관한 값을 갖도록 한다.

$$HRL_{avg} = \frac{1}{M}\sum_{}\frac{1}{N}\sum_{i} rl_i$$

N : 세리프 영역의 런의 개수, rl_i = i번째 런의 길이 / HRL_{max_freq}

M : 세리프 영역의 개수

블록은 시작점과 끝점이 같은 런들의 집합으로, 블록의 폭은 역시 문자의 크기에 무관한 특징을 갖도록 HRL_{max_freq} 값으로 나눈다.

4. 실험 결과 및 분석

4.1 단어 영상 데이타베이스

단어 영상의 속성은 다섯 가지 속성을 조합하여 총 264클래스를 고려하는데, 실험을 위해 각 클래스별로 서로 다른 100개의 단어 영상(학습용: 50, 테스트용: 50)을 구축하였다. 따라서 실험에 사용되는 단어 영상은 총 26,400개로, 한글 영상이 9,600개, 영문 영상이 16,800개로 구성된다. 단어 영상은 흔글 97 워드프로세서 상에서 해당 속성들을 생성한 후, SHARP ScanJX 스캐너를 사용하여 300dpi 해상도로 스캔하였다. 그림 10은 서체 및 스타일 속성을 가진 한글 및 영문 단어 영상의 예를 보여준다.

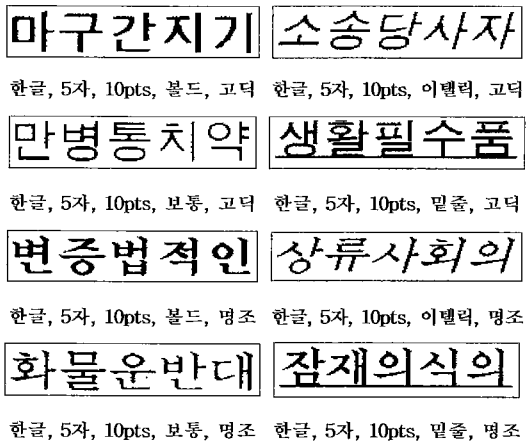

그림 10 속성 인식을 위한 단어 영상 DB의 예

4.2 실험 결과 및 분석

단어 영상에 대한 속성 인식 과정은 13,200개(속성별 50개)의 단어 영상에 의한 분류기 학습 과정과 13,200개의 테스트 데이타에 대한 인식율 측정 과정으로 나눈다. 특징 추출 과정은 3단계로 나누어, 언어 종류 분류를 위한 특징 추출, 문자 크기, 문자 개수, 스타일 분류를 위한 특징 추출, 서체 분류를 위한 특징 추출 단계로 구성한다. 분류기의 계층적 모델에 의한 속성 분류 결과는 그림 11에 나타나 있다.

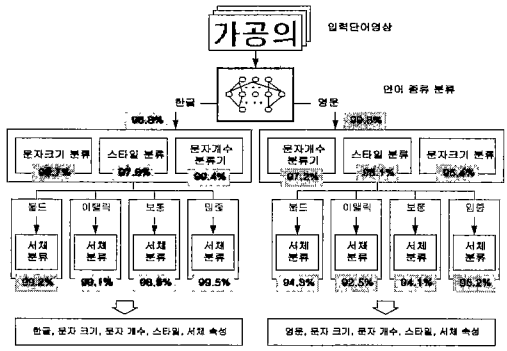

그림 11 분류기의 계층적 모델에 의한 분류 결과

(1) 언어 종류 분류 결과

단어 영상의 언어 종류 분류는 13,200개(한글: 4,800, 영문: 8,400)의 단어 영상을 사용하여 MLP를 학습하고, 13,200개의 영상에 대한 인식률을 측정하였다. 인식 결과는 98.6%인데, 한글 단어 영상의 경우 96.8%, 영문의

경우 99.8%이다. 언어의 종류에 대한 인식 결과, 한글의 인식률이 낮게 나타나는 이유는 한글 단어 영상에서 초성, 중성, 종성간의 접촉이 발생하여 하나의 연결요소로 나타나는 경우이다. 이런 현상에서는 한글의 한 문자가 두 개 이상의 연결요소로 이루어진다는 특징을 나타내지 못하거나 연결요소간의 수직 방향 겹침 정도도 0으로 나타나기 때문에 오류를 유발하였다.

(2) 스타일 분류 결과

단어 영상의 스타일 분류는 언어의 종류에 따라 학습된 2차형 판별함수(QDF)를 유지하여 사후확률이 가장 큰 스타일로 분류한다. 한글 단어 영상의 스타일 분류 결과는 97.8%, 영문 단어 영상의 스타일 분류 결과는 98.1%이며, 각 클래스별 사용 데이타 및 분류 결과는 다음 표 2와 같다.

표 2 각 언어별 스타일 분류 결과 (%)

언어	한글				영문			
스타일	볼드	이탤릭	보통	밑줄	볼드	이탤릭	보통	밑줄
데이타	1,200	1,200	1,200	1,200	2,100	2,100	2,100	2,100
분류율	96.9	100	94.4	99.9	98.1	96.2	98.1	100
계	97.8				98.1			

한글 단어 영상의 경우 다른 스타일에 비해 볼드와 보통의 인식률이 낮게 나타났고, 영문의 경우 이탤릭의 속성 인식이 낮게 나타났다. 스타일 속성 분류의 오류를 살펴보기 위한 한글과 영문에 대한 혼동행렬(confusion matrix)은 표 3, 4와 같다.

표 3 한글 단어 영상의 스타일 분류에 대한 혼동행렬

	볼드	이탤릭	보통	밑줄	계	정확율
볼드	1163	0	37	0	1200	96.9%
이탤릭	0	1200	0	0	1200	100%
보통	60	5	1133	2	1200	94.4%
밑줄	0	0	1	1199	1200	99.9%
계					4800	97.8%

표 4 영문 단어 영상의 스타일 분류에 대한 혼동 행렬

	볼드	이탤릭	보통	밑줄	계	정확율
볼드	2060	0	39	1	2100	98.1%
이탤릭	0	2020	78	2	2100	96.2%
보통	13	26	2061	0	2100	98.1%
밑줄	0	0	0	2100	2100	100%
계					8400	98.1%

한글과 영문 단어 영상에서 특징으로 사용하는 획의 평균 폭 정보(HRL_{avg})가 볼드와 보통 클래스를 분류하는데 변별력이 낮아서 볼드 속성이 보통으로, 보통 속성이 볼드로 분류되는 경우가 있었다. 이런 현상은 문서의 출력이나 스캐닝 과정에서 주로 발생한다. 영문 단어 영상에서는 수직 투영 프로파일의 1차 미분 값(DVP_{max})이 이탤릭과 보통 클래스를 분류하는데 변별력이 낮았다. 이탤릭 속성이 보통으로 분류되는 경우는 'l'이 포함된 단어 영상들에서 주로 발생하였고, 보통 속성이 이탤릭으로 분류되는 경우는 영문의 중심지역(middle zone)에만 위치하는 문자('a', 'c', 'e', 'o', 's' 등)로 구성된 단어 영상에서 나타났다.

(3) 문자 크기 분류 결과

단어 영상의 문자 크기 분류는 언어의 종류별로 학습된 2차형 판별함수를 유지하여 사후확률이 가장 큰 문자 크기로 분류한다. 한글 단어 영상의 문자 크기 분류는 99.7%, 영문의 문자 크기 분류는 96.4%이며, 각 클래스별 사용 데이타 및 분류 결과는 다음 표 5와 같다.

표 5 각 언어별 문자 크기 분류 결과 (%)

언어	한글			영문		
문자크기	10pts	12pts	14pts	10pts	12pts	14pts
데이타	1,600	1,600	1,600	2,800	2,800	2,800
결과	100	99.5	99.5	99.9	94.9	94.5
계	99.7			96.4		

영문의 경우 수직거리 특징을 추출하기 위한 영상의 중심지역을 계산하는데, 단어 영상의 수평 방향 투영 프로파일의 형태가 복잡하여 중심지역의 추출에서 오류가

발생했다. 이런 경우는 단어 영상이 소수의 문자로 구성되어 있을 때 주로 발생하였다.

(4) 문자 개수 분류 결과

한글 단어 영상의 문자 개수 분류는 단어 영상의 높이에 대한 폭의 비율을 사용하는데, 99.4%의 분류 결과를 얻었다. 영문의 경우는 연결요소의 개수를 사용하는데, 97.2%의 결과를 얻었다. 연결요소의 개수 특징은 문자가 조각으로 부서지는 경우에 대해 오류를 유발하였다.

(5) 서체 분류 결과

단어 영상의 서체 분류는 언어의 종류와 스타일별로 개별적인 LDF 분류기를 구축하는데, 한글은 하나의 특징을, 영문은 두 개의 특징을 사용한다. 한글의 서체 분류 결과는 99.2%, 영문은 94%이고, 각 클래스별 분류 결과는 표 6과 같다.

표 6 한글 및 영문 단어 영상의 서체 분류 결과 (%)

	스타일	볼드		이탤릭		보통		밑줄	
	서체	명조	고딕	명조	고딕	명조	고딕	명조	고딕
한글	결과	98.3	100	99.1	99.1	97.7	100	98.9	100
	소계	99.2		99.1		98.9		99.5	
영문	결과	99.2	89.4	97.5	87.4	98.8	89.3	98.8	91.6
	소계	94.3		92.5		94.1		95.2	

영문의 서체 분류에서 인식률이 낮은 이유는 수평 런 길이의 평균과 블록의 평균 면적으로 얻어지는 특징이 명조에 비해 고딕의 특성을 잘 표현하지 못했기 때문이다. 고딕의 속성을 가진 단어 영상의 특징 값들의 분포는 분산이 매우 크게 나타났다. 즉, 추출한 특징 값들이 영문의 중심지역에만 위치하는 문자로 구성된 단어 영상에서는 명조와 특별한 차이를 보이지 않았기 때문에 변별력이 낮았다.

5. 결론 및 향후 연구과제

본 논문에서는 문서 인식의 효율을 위해 문서구조 분석 후 분할된 단어 영상들이 가지고 있는 다양한 속성 정보들을 추출하는 방법을 제안하였다. 제안방법에서는 언어의 종류, 스타일, 문자 크기, 문자 개수, 서체의 다섯 가지 속성 정보를 복합적으로 포함하고 있는 단어 영상들로부터 해당 속성들을 분류하였다. 특히, 본 연구는 한글 및 영문 속성에 대한 제약이 없고, 적은 수의 특징(한글: 8, 영문: 9)을 사용하며, 신경망, 2차형 판별함수, 선형 판별함수를 계층적으로 구성하였다. 단어 영상 26,400개를 사용한 실험을 통해, 적은 차원의 특징을 사용하는 계층적 분류 모델의 성능을 관찰하였다. 이러한 속성 인식의 성능은 개별적 단어에서 문장, 문단 등의 전역적 측면으로 확장하여 개선할 수 있다. 단어 단위의 속성 인식이 문서 영상 검색(Document Image Retrieval)이나 OCR 시스템에 활용할 때 가질 수 있는 기대 효과는 단어 단위 영상 매칭의 정확도와 속도 개선, 단어 인식을 위한 특징 추출 과정의 간소화, 주제어 검색에 대한 탐색공간 축소 등을 들 수 있으며, 해당 속성에 전문화 된 인식기를 구축함으로써 문자 인식률의 향상 등을 도모할 수 있다. 따라서 본 연구의 향후 과제는 전자도서관(digital library) 상에서의 문서영상 검색 시스템, 광파일(optical file) 시스템 등의 실제 응용 분야에서 기여도를 검증하는 것이다.

참 고 문 헌

[1] AIIM'96 Conference Handbooks, Association for Imaging and Information Methodologies, 1996.

[2] D. Doermann, "The Indexing and Retrieval of Document Images: A Survey," *Computer Vision and Image Understanding*, Vol. 70, No. 3, pp. 287-298, 1998.

[3] A. Zramdini, "Study of Optical Font Recognition Based on Global Typographical Features," Ph. D thesis, University of Fribourg, 1995.

[4] 박문호, 손영우, 김석태, 남궁재찬, "인쇄된 한글 문서의 '폰트 인식," 한국정보처리학회논문지, 제 4권, 제 8호, pp. 2017-2024, 1997.

[5] S. Kahan, T. Pavlidis and H.S. Baird, "On the Recognition of Printed Characters of Any Font and Size," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 9, No. 2, pp. 274-288, 1987.

[6] M.C. Jung, Y.C. Shin and S.N. Srihari, "Multifont Classification Using Typographical Attributes," *Proc. 5th Int. Conf. Document Analysis and Recognition*, Bangalore, pp. 353-356, 1999.

[7] B.B. Chaudhuri and U. Garain, "Automatic Detection of Italic, Bold and All-Capital Words in Document Images," *Proc. 14th Int. Conf. Pattern Recognition*, Brisbane, pp. 610-612, 1998.

[8] T.K. Ho, J.J. Hull and S.N. Srihari, "A Computational Model for Recognition of Multi-Font Images," *Machine Vision and Applications*, Vol. 5, No. 1, pp. 157-168, 1992.

[9] S. Zhao and S.N. Srihari, "A Word Recognition

Algorithm for Machine-Printed Word Images of Multiple Fonts and Varying Qualities," *Proc. 3rd Int. Conf. Document Analysis and Recognition*, Montreal, pp. 351-354, 1995.
[10] T.K. Ho, "Font Identification of Stop Words for Font Learning and Keyword Spotting", *Proc. 5th Int. Conf. Document Analysis and Recognition*, Bangalore, pp. 333-336, 1999.
[11] Z. Lu, R. Schwartz, P. natarajan, I. Bazzi and J. Makhoul, "Advances in the BBN BYBLOS OCR System," *Proc. 5th Int. Conf. Document Analysis and Recognition*, Bangalore, pp. 337-340, 1999.
[12] U. Garain and B.B. Chaudhuri, "Extraction of Type Style Based Meta-Information from Imaged Documents," *Proc. 5th Int. Conf. Document Analysis and Recognition*, Bangalore, pp. 341-344, 1999.
[13] D. Xi, S. Lee and Y. Tang, "A Novel Method for Discrimination between Oriental and European Languages by Fractal Features," *Proc. 5th Int. Conf. Document Analysis and Recognition*, Bangalore, pp. 345-348, 1999.
[14] Y. Zhu, T. Tan and Y. Wang, "Font Recognition Based on Global Texture Analysis," *Proc. 5th Int. Conf. Document Analysis and Recognition*, Bangalore, pp. 349-352, 1999.

곽 희 규
1996년 전남대학교 전산학과 졸업(학사). 1998년 전남대학교 대학원 전산통계학과 졸업(이학석사). 2001년 전남대학교 대학원 전산통계학과 졸업(이학박사). 2001년 ~ 현재 한국과학기술원 박사후 연구원. 관심분야는 패턴인식, 영상처리, 컴퓨터비젼.

김 수 형
1986년 서울대학교 컴퓨터공학과 졸업(학사). 1988년 한국과학기술원 전산학과 졸업(공학석사). 1993년 한국과학기술원 전산학과 졸업(공학박사). 1990년 ~ 1996년 삼성전자 멀티미디어연구소 선임연구원. 1997년 ~ 현재 전남대학교 컴퓨터정보학부 조교수. 관심분야는 패턴인식, 영상처리, 컴퓨터비젼, 신경망학습.