대용량 인쇄 한글 문서 검색을 위한 영상 기반 단어 매칭 방법

진영범° 오일석
전 북 대 학 교 컴 퓨 터 과 학 과
ybjin@cs.chonbuk.ac.kr isoh@moak.chonbuk.ac.kr

An Image-based Word Matching Method for Large volume Printed Hangul Document Retrieval

Young-Bum Jin° Il-Seok Oh
Department of Computer Science, Chonbuk National University

기계 인쇄된 문서 영상에서 주제어를 탐색하는 문제는 여러 응용 분야에 필수적인 핵심 기술이지만 수작업 또는 OCR 소프트웨어를 이용하여 텍스트로 변환하는 방법은 많은 비용 때문에 한계를 가지고 있다. 요즘 영상 형태로 원문을 저장하는 경우가 많으므로 본 논문은 영상-기반 매칭을 통한 검색 방법을 채택하였다. 문자 또는 단어 매칭에서 가장 중요한 요소가 특징인데 본 논문에서는 디지털도서관과 같이 매칭 대상 단어가 수천만~수십억에 달하는 대용량 한글 문서 검색에 이용될 수 있도록 비교적 간단히 추출할 수 있고 차원수 조절이 용이한 4방향 프로파일 특징을 이용하는 빠른 검색 방법을 제안한다. 실험결과 8-차원 정도의 간단한 특징으로도 의미 있는 검색 성능을 얻을 수 있음을 보였다.

1. 서론

기계 인쇄된 문서영상에서 주제어(keyword)를 탐색하는 문제는 여러 응용 분야에 필수적인 핵심 기술이다. 예를 들어, 디지털 도서관 구축에서는 하나의 문서(논문, 또는 기사 등)를 미리 색인(indexing)해 놓아야 하는데, 이런 작업을 위해서는 그 문서가 가지고 있는 주제어를 추출하는 작업이 선행되어야 한다[1]. 주제어 추출 작업에 있어서 이제까지는 일부 훈련된 전문인의 수작업으로 수행되어 왔고, 이러한 값비싼 입력방법은 자동 검색 시스템 구축의 걸림돌로 작용하고 있다. 또한 미리 주어진 주제어에 한정하여 문서 검색이 가능하다는 근본적인 한계를 가진다.

이러한 문제를 해결하기 위한 하나의 접근 방법은 입력 문자열 영상을 낱자단위로 분할한 다음 OCR 소프트웨어를 사용하여 개개의 낱자를 인식하여 후보 인식 결과를 만들어 낸 다음 문자열 매칭 알고리즘을 사용하여 검색하는 것이다. 이때 OCR 소프트웨어의 오인식 가능성을 대비하여 사전을 이용한 후처리 방법[2]을 사용하여 최종적인 단어의 인식 결과를 출력하기도 한다. 이 접근 방법은 OCR 소프트웨어의 오인식 문제 때문에 한계를 나타내고 있다. 기존 한글 인식기가 실용화에 충분할 정도의 성능을 보이지 못하고 있고, 자연어 처리에 기반한 후처리 모듈도[3] 인식 성공률을 높인다는 보장이 없다. 또한 사람이 오류를 수정하는 경우 엄청난 비용이 뒤따른다. 현재 OCR 소프트웨어의 성능은 처음부터 사람이 인식하는 비용과 OCR 소프트웨어로 인식하고 사람이 오류를 수정하는 비용이 엇비슷한 상태로 보인다[4]. 이러한 문제를 완화하기 위해 OCR 소프트웨어의 오류 경향을 표로 만들어 이를 이용하여 오류-교정 검색을 시도한 논문도 발표되어 있다[5].

또 다른 접근방법은 영상-기반 매칭 방법이다. 영상-기반 매칭 방법이란 문서를 스캔하여 영상형태로 저장한 상황하에서 질의단어가 주어지면 이에 대한 영상특징을 추출하여 문서 영상 내에 존재하는 단어 영상과 매칭을 시도하여 일치하는 단어를 검색하는 기법이다.

앞에서 언급했듯이 OCR 소프트웨어의 인식률 한계와 사람이 입력하는경우 막대한 비용 때문에 문서를 영상형태로 저장하는 경우가 많이 있다. 예를 들어 KORDIC의 경우 각종 저널들을 스캔하여 영상형태로 저장하고 이를 기반으로 원문 검색을 해주고 있다. 이러한 상황에서 수천만 페이지에 달하는 방대한 분량의 문서에서 주제어를 탐색하기 위해서는 고속으로 작동하고 검색 성능이 높은 방법을 사용해야 한다.

이러한 접근 방법이 한글에 적용된 사례는 웨이블릿(wavelet) 특징을 이용한 김혜금[6] 논문을 제외하고는 거의 없으나 영문, 일본어, 중국어 등에 대해서는 상당한 연구사례가 발표되어 있다. 그 예로서 Zhu 등[7]은 인쇄체 중국어 문서를 대상으로 영상-기반 주제어 인식을 시도했다. 이 연구의 주된 내용은 낱자별 인식에서 탈피하여 주제어 사전을 이용한 단어 단위 인식이다. Kuo와 Agazzi[8]는 인쇄체 영문 문서를 대상으로 질의를 통해 문서 내에서 주제어 위치를(keyword spotting) 찾아내는 방법을 제안하였다.

본 논문에서는 인쇄 한글 문서 검색을 위한 영상-기반 주제어 매칭 방법을 제안하여 디지털 도서관등에 응용할 수 있는 계기를 마련하고자 한다. 본 논문은 한 페이지의 문서 영상이 전처리 단계를 거쳐 단어 단위로 분할되었다는 가정 하에서 단어 영상 매칭 알고리즘을 제안한다 데이터베이스로는 전남대에서 구축한 인쇄 한글 단어 영상을 사용하였다. 조사 문제와 장평 문제등을 고려하여 우선 단어 영상을 낱자영상으로 분할한다. 분할된 낱자영상에서 특징벡터를 추출하고 유클리디언 거리를 계산하여 매칭정도를 계산한다. 고속 수행을 위하여 특징은 저차원을 사용한다. 실험에서는 저차원부터 고차원까지 검색성능과 속도를 측정하였다. 실험결과 저차원.특징으로도 의미있는 검색성능을 얻을 수 있음을 알 수 있었다.

2장에서는 본 논문이 사용한 검색 방법을, 3장에서는 실험에 사용된 DB구축 방법과 표준 폰트를 사용한 문자모델 구축 방법을, 4장에서는 실험결과를 설명하였다.

2. 영상-기반 검색 방법

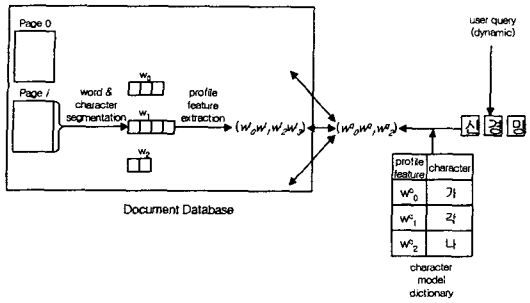


그림 1. 영상-기반 검색 방법

그림 1은 우리가 구축한 시스템을 보여주고 있다. 모든 문서 페이지 영상들은 미리 줄, 단어, 문자들로 분할되어 있고 모든 필요한 특징들도 추출되어 데이터베이스에 저장되어 있다. 이 방법에서 사용자 질의(user query)는 동적(dynamic)으로 주어진다. 사용자로부터 질의 단어가 입력됐을 때, 단어를 구성하는 각각의 문자들에 대한 특징 벡터(feature vector)를 구성하기 위해 문자 모델 사전(character model dictionary)을 참조한다. 읽어온 문자모델을 이용하여 검색하고자 하는 이미지상의 단어들과 문자단위로 매칭하고 이 결과를 조합하여 단어의 매칭여부를 최종 결정한다.

2.1 문자모델

사용자 질의에 대한 특징을 추출하는 방법에는 표준 폰트를 사용하는 방법과, 문서에서 대표 문자 패턴을 골라 사용하는 방법이 있다. 본 논문에서는 표준 폰트를 사용하고 있다.

문자 모델 사전은 Font Xplorer(그림 2) 라는 프로그램을 이용하여 구축하였다. 이 프로그램은 윈도우에서 사용 가능한 모든 폰트를 제공하며, 원하는 글자를 임의의 크기 이미지 형태(BMP)로 잡음과 변형이 전혀 없는 폰트 고유의 상태로 획득할 수 있는 프로그램이다.

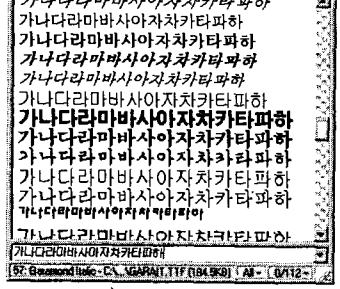
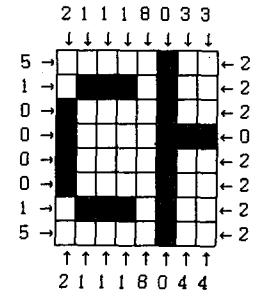


그림 2. Font Xplorer

2.2 특징 추출

문자 또는 단어 매칭에서는 분별력 높은 특징을 선택하여 사용하는 것이 중요하다[9]. 이때 같이 고려할 사항이 특징에 관련한 비용이다. 이 비용은 크게 두 가지 측면이 있는데 하나는 추출하는데 드는 비용이고 다른 하나는 실제 매칭과정에서의 계산량이다. 우리 문제의 경우는 미리 특징을 추출하여 문서 데이터베이스와 문자모델 사전(그림 1 참조)에 저장하기 때문에 첫 번째 비용은 문제가 없다. 하지만 두 번째 비용은 검색속도와 비례하고 디지털 도서관같은 전형적인 응용에서 매칭 대상 단어가 수천만~수십억 단어이므로 매우 중요하다. 이 논문은 비교적 간단히 추출할 수 있고 차원수 조절이 용이한 특징을 사용한다. 실험에 사용된 특징은 4방향(top, bottom, left, right) 프로파일 특징이다. 그림 3이 추출과정을 설명한다. 입력 단어 영상에 대해 문자별로 분할후 각 문자에 대해 이진화, 크기 정규화 과정을 거친다. 그런 다음 그림 3에서와 같이 각 4변 시작점에서의 white run 수를 특징값으로 사용된다. 실제 실험에서는 문자영상을 64× 64로 정규화하여 추출하였다.



$\overline{F_{32}}$ = (2,1,1,1,8,0,3,3,2,1,1,1,8,0,4,4, 5,1,0,0,0,0,1,5,2,2,2,0,2,2,2,2)
$\overline{F_{16}}$ = (2,1,8,3,2,1,8,4,5,0,0,1,2,2,2,2)
$\overline{F_{8}}$ = (2,3,2,4,5,5,2,2)

그림 3. 4방향 프로파일 특징 벡터

2.3 매칭 방법

매칭은 사용자가 제시한 질의단어(query word)를 문서내에 있는 단어와 하나씩 수행해 나간다. 현재 매칭 대상이 되는 문서내 단어를 목적단어(target word)라 하자. 질의단어와 목적단어를 각각 Q와 T로 표기하자. Q와 T에 K개의 문자가 있다면 다음과 같이 표기할 수 있다.(Q와 T의 문자 개수가 조사등의 이유로 다를 수 있는데 Q를 T 내에서 이동하면서 매칭을 시도하므로 동일한 개수가 있다고 가정하고 알고리즘을 기술한다.)

$$Q = (\overline{F^q_1}, \overline{F^q_2}, \ldots, \overline{F^q_k})$$
$$T = (\overline{F^t_1}, \overline{F^t_2}, \ldots, \overline{F^t_k})$$

여기서 $\overline{F_i}$는 i번째 문자에서 추출한 특징 벡터로서 다음과 같이 표기할 수 있다. n은 특징벡터의 크기이다.

$$\overline{F_i} = (f_{i,1}, f_{i,2}, \ldots, f_{i,n})$$

두 개의 문자간 매칭 거리는 아래공식으로 계산한다.(계산속도 향상을 위해서는 제곱근 대신 절대값을 사용할 수도 있다.)

$$Distance(\overline{F^q_i}, \overline{F^t_i}) = \sum_{j=1}^{n} \sqrt{(f^q_{i,j} - f^t_{i,j})^2}$$

이제 K개의 문자간 매칭거리를 이용하여 두 개의 단어 Q와 T간의 최종 매칭여부를 결정해야 한다. 두 가지 규칙을 생각할 수 있다.

규칙 1 : $\left(\sum_{i=0}^{k} Distance(\overline{f_i^q}, \overline{f_i^t}) \right) / K < T$ 이면 성공적인 매칭

규칙 2 : 모든 i에 대해 $Distance(\overline{f_i^q}, \overline{f_i^t}) < T$ 이면 성공적인 매칭

규칙 1은 문자의 매칭거리의 평균을 이용한 것으로서 4장에 기술한 실험에서 이 규칙을 사용하였다.

3. 데이터베이스

제안한 알고리즘의 성능을 객관적으로 평가하기 위해서는 우수한 데이터베이스가 필요하다. 우리 실험에서는 전남대에서 구축한 한글 단어영상 데이터베이스를 사용하였다. 이 데이터베이스 내용을 표 1이 설명하고 있고 그림 4는 몇 가지 예제를 제시한다.

표 1. 전남대 단어 영상 데이터베이스

폰트	고딕		명조	
속성	bold	regular	bold	regular
단어당 문자수	2 ~ 5 문자			
크기	10, 12, 14 point			
단어수	1175	1075	1181	1086

가지 경력 반암 망상
강제의 나중에 느리게 붙잡다
저주하다 희게하다 우중충한 저주받은
필사적으로 새로이하다 특별열람실 홍역에걸린

고딕 bold 고딕 regular 명조 bold 명조 regular

그림 4. 실험에 사용한 한글 단어 영상 예

2장에 기술한 단어 매칭 알고리즘은 문자 영상을 필요로 하므로 단어 영상을 낱자로 분할하는 작업을 수행하였다. 우선 단어영상의 수직투영(vertical projection)을 구한 후, 얻어진 히스토그램과 문자의 표준 폭 정보를 이용하여 분할하였다. 그림 5는 단어영상의 수직투영 결과를 보이고 있다. 이렇게 분할된 문자에 대해 2.2절에 기술한 프로파일 특징을 추출하여 그림 1의 문서 데이터베이스와 문자 모델 사전에 저장하였다.

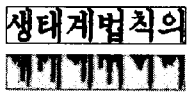


그림 5. 수직 투영

4. 실험

문자 모델 사전은 그림 2의 Font Explorer 소프트웨어를 사용하여 문자 패턴을 생성하고 여기에서 특징을 추출하여 저장하였다. 특징의 차원수가 검색 속도와 검색 성능에 미치는 영향을 관찰하기 위해 특징은 저차원에서 고차원까지 추출하여(8-D, 20-D, 36-D, 256-D) 사용하였다.

검색 성능은 아래식으로 정의되는 재현율과 정확률로 측정하였다.

재현율 = (검색된 적합 단어의 수) / (적합 단어의 총 수)
정확률 = (검색된 적합 단어의 수) / (검색된 단어의 총 수)

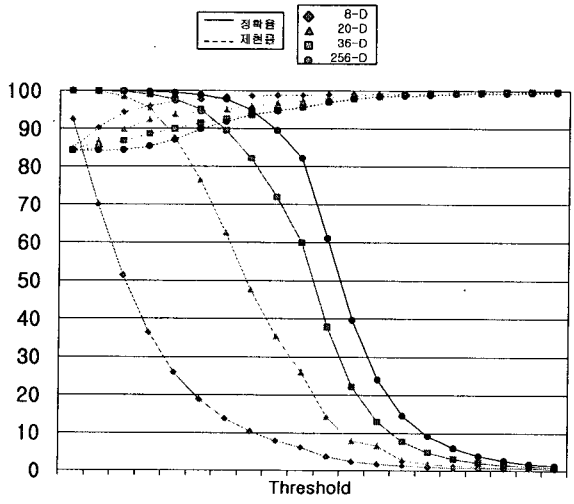


그림 6. 실험 결과

그림 6은 실험결과를 보여준다. 전체적인 성능은 많은 차원의 특징을 이용해 검색할 경우 정확률과 재현율이 만나는 지점이 높게 형성된다. 하지만 재현율만을 놓고 볼 때 저차원의 경우가 훨씬 높게 형성됐다는 사실을 알 수 있다.

실험은 셀러론 500MHz, 64MB 메인메모리의 PC에서 수행하였고, 검색 속도를 표 2가 보이고 있다.

표 2. 검색 속도

	8-D	20-D	36-D	256-D
문자/초	185만	55만	35만	68930
단어/초 (5문자단어가정)	35만	11만	7만	13786
쪽/초 (쪽당 5백단어 가정)	700	220	140	27.5

5. 결론 및 향후 연구 과제

이와 같이 영상-기반 주제어 검색에 대한 방법을 설명하고 그 가능성에 대해서 알아보았다. 본 논문의 방법이 검색 속도는 월등히 빠르나 성능은 떨어지는 단점을 보완하기 위해서 앞으로는 현재 두 종류 폰트에 대해 774문자에 대해 구축돼 있는 문자 모델 사전을 한글 2,350자에 대해 구축할 예정이며, 인식성능을 향상시킬 수 있도록 첫 번째 단계에서는 적은수의 재현률이 높은 특징을 사용하여 매칭하고 추출된 단어들에 대해 정확률이 높은 특징을 사용하는 두 단계 매칭 방법을 연구해 봄직하다.

6. 참고문헌

[1] 김태수, 유양근, 정준민, 최석주, "디지털 도서관," 사이텍미디어, 2000
[2] 김윤호, 이종국, 김항준, 이상조, "형태소 분석을 이용한 문자인식 에러의 검출," 제4회 한글 및 한국어 정보처리 학술 발표회 논문집, pp.545-553, 청주, 1992.
[3] 홍남희, 이원일, 이종혁, 이근배, "어절 정보와 문자열 정보를 이용한 문자 이닉에서의 오인식 수정 기법에 관한 연구," 제1회 문자인식 워크샵 발표논문집, pp.109-113, 청주, 1993.
[4] 한선화, 이충식, 이준호, 김진형, "문자 인식 기술을 이용한 데이터베이스 구축," 정보처리학회 논문지, 제6권 제7호(99.7), pp.1713-1722, 1999.
[5] 안재철, 홍기천, 오일석, "OCR 소프트웨어의 오인식 특성을 반영한 한글 단어 검색 방법," 한국 정보과학회 호남제주지부 춘계 학술발표 논문집, Vol.12, No.1, pp.187-195, 동신대학교, 2000년 8월.
[6] 김혜금, 최윤근, 오일석, "웨이블렛 다중 규모 공간에서 한글단어의 매칭성능 분석," 한국 정보과학회 춘계 학술발표논문집 26(1), 목포대학교, pp.558-560, 1999년 4월.
[7] Jason Zhu, Tao Hong and Jonathan J. Hull, "Image-based keyword recognition in oriental language document images," Pattern Recognition, pp.1293-1300. 1997.
[8] Shyh-shiaw Kuo and Oscar E. Agazzi, "Keyword spotting in poorly printed documents using pseudo 2-D Hidden Markov Models," IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol.16, No.8, Aug. 1994.
[9] Menahem Friedman, Abraham Kandel, "Introduction To Pattern Recognition," World Scientific, 1999.