한글 문서 영상의 단어 검색 시스템

최윤성[0], 오일석
전북대학교 컴퓨터과학과
yschoi@cs.chonbuk.ac.kr, isoh@moak.chonbuk.ac.kr

A Keyword Spotting System of Korean Document Images

Yoon-Sung Choi, Il-Seok Oh
Department of Computer Science, Chonbuk National University

요 약

본 논문은 한글 문서 영상의 단어 검색 시스템과 그 성능을 제시한다. 두 단계 검색 방법은 검색 속도 증가를 목적으로 하며, 첫 번째 단계에서는 매우 빠른 속도로 거친 정합을 통하여 후보 단어들을 추출한다. 두 번째 단계는 후보 단어들 중에서 미세한 정합을 통한 단어 검색이 이루어진다. 시스템은 문서 영상 구조 분석 모듈과 단어 검색 모듈로 구성된다. 실험 자료를 통해 시스템의 유용성을 입증한다.

1. 서 론

전자 도서관에 방대한 양의 종이 문서가 스캔되어 데이터베이스 안에 저장되어있을 때, 사용자에게 제공되는 가장 중요한 서비스로 검색 기능을 들 수 있다. 스캔된 문서 영상의 검색 서비스는 전자 문서와 같거나 거의 동일한 수준을 제공해야한다[1]. 서구 언어로 된 문서영상에 대해서는 많은 연구들이 진행되고 있으며[2, 3], 또한 사용 가능한 시스템들도 개발 되어져있다[4].

한국에서는 NDL(National Digital Library) 프로젝트가 막대한 자금력을 동원하여 수행되고 있다. 이로 인하여 대용량의 문서 영상이 디지털 이미지 형식으로 데이터베이스에 저장되어 있으며, 현재 인터넷을 통해 서비스를 제공하고 있다[5]. 그러나 문서 영상에 대한 전문 검색(full-text retrieval)과 적합도 순위(relevance ranking)와 같은 진보된 서비스는 제공되지 않고 있다.

이 논문은 한글 문서 영상의 단어 검색 시스템을 기술한다. 제안한 2단계 검색 시스템은 거친 프로파일 특징(coarse profile feature)과 미세한 웨이블릿 특징(fine wavelet feature)을 사용하도록 설계되어있다. 실험 결과 두 단계에 의한 검색 구조는 기존의 한 단계의 검색 구조에 비해 대략 6배의 검색 속도 개선을 보였다.

2. 단어 검색 시스템

제안한 검색 시스템은 IBM PC에서 Visual C++ 언어를 사용하여 개발하였으며, 전남대학교에서 개발한 페이지 영상의 구조 분석 및 단어 단위 분할 모듈을 인스톨하였다[6]. 그림 1은 제안된 문서 영상 검색 시스템의 전반적인 흐름을 보여준다.

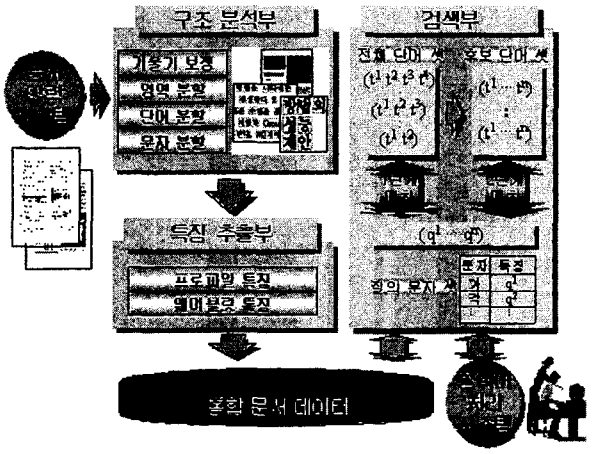
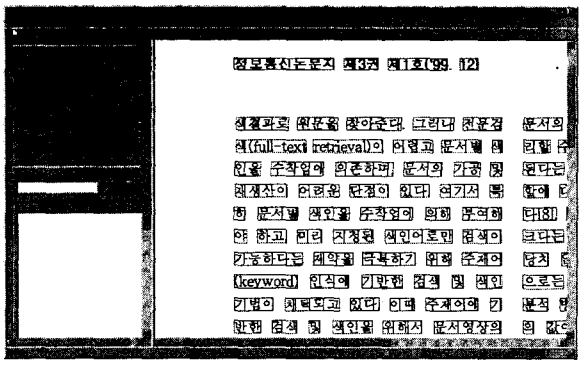


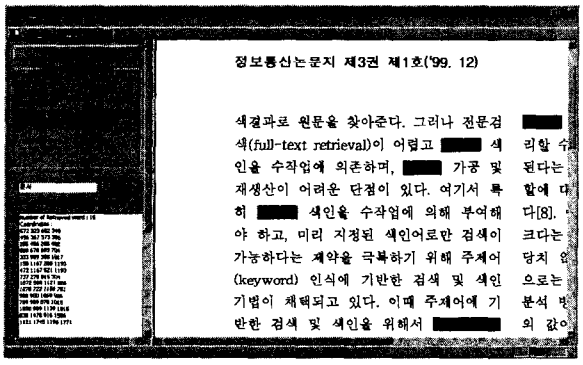
그림 1. 시스템 흐름도

문서 입력 시스템에 의해 문서 영상이 입력되면 구조 분석부에서 문서의 기울어짐 보정과 텍스트와 그림, 표 등의 영역을 분할하고, 텍스트 영역에서 단어 분할과 문자 분할을 하게된다. 특징 추출부에서는 분할된 각 문자에서 특징이 추출된 후 저장된다. 사용자가 입력한 질의어는 질의어 처리 시스템에 의해 처리된 질의어의 특징과 목적 단어 영상의 특징과의 두단계 매칭이 검색부에서 시행된다.

검색 시스템은 그림 2에서 볼 수 있으며, 인터페이스의 왼쪽에는 질의 입력 창과 문서 영상의 정보를 보여주고, 오른쪽에는 원 문서 영상과 검색된 단어영상을 표시하여 보여주고 있다. 그림 2(a)는 문서 영상을 분석하는 모듈의 결과이고, 그림 2(b)는 단어 검색 모듈에 대한 결과이다.



(a) 페이지 구조 분석



(b) 단어 검색

그림 2. 문서 영상 검색 시스템

3. 특징 추출과 단어 영상 매칭 알고리즘

단어 영상은 수직 투영 분석을 통해 문자 단위로 분할된다. 분할된 문자 영상은 32x32로 정규화 한 후, 2 종류(프로파일 특징, 웨이블릿 특징)의 특징을 추출한다.

첫 번째 단계에 사용할 특징을 얻기 위하여, 4개의 프로파일 특징은 위, 아래, 왼쪽, 오른쪽 방향의 투영을 통해 계산하며, 각 방향에 대해 32개의 값을 가지는 1차원 배열로 표현한다. 각 방향에 대해 32개의 값을 평균함으로써 얻는 4차원 거친 프로파일 특징 벡터를 획득한다.

두 번째 특징은 Harr 연산자를 사용한 웨이블릿 변환 기법에 의해 획득한다. 한글 단어 영상 검색에서 Harr 웨이블릿 특징들의 효율성은 [7]에 잘 기술되어있다. 웨이블릿 계수들을 특징 벡터로 사용하는데, 큰 값을 갖는 계수들은 원 영상을 대표하는 중요한 역할을 하기 때문에, 상위 k개의 계수만을 가지고 특징 벡터로 사용한다. 본 실험에서는 k=30으로 설정하였다.

본 논문은 거친 특징과 미세한 특징을 사용하여 2단계 검색 시스템을 구현한다. 그림 3은 방대한 양의 단어 영상을 보유하고 있는 전자도서관에서의 2단계 검색 시스템의 구조를 보여준다.

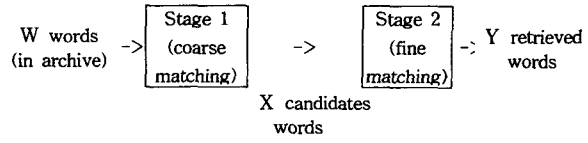


그림 3. 2단계 검색

1단계에서는 4차원 프로파일 특징 벡터를 이용하여 거친 정합을 하게된다. 첫 번째 단계의 효과는 매우 빠른 속도로 후보 단어들을 추려내는 것이다.

매칭은 질의어(Q)와 목적(T) 단어 영상 사이의 유클리드 거리 계산법을 이용하여 수행한다. 질의어의 특징으로 사용되는 모델 문자 셋은 2,350자로서 노이즈가 없는 깨끗한 대표 문자 영상에서 추출했다.

다음의 수식에서 q^i와 t^i는 하나의 문자에 대한 4차원 프로파일 특징 벡터들을 나타낸다. p_1과 p_2는 매칭 규칙에서 사용하는 임계값으로서 이 값을 조절함으로써 검색 성능(정확률과 재현률)을 조절할 수 있다.

1 단계:

질의 단어 : $Q = (q^1, q^2, \cdots, q^n)$, $q^i = (x_1^i, x_2^i, x_3^i, x_4^i)$

목적 단어 : $T = (t^1, t^2, \cdots, t^n)$, $t^i = (y_1^i, y_2^i, y_3^i, y_4^i)$

문자간 거리 : $d^i = (\sum_{k=1,4}(x_k^i - y_k^i)^2)^{1/2}$

Q 와 T의 거리 : $d^{qwe} = \sum_{k=1,n} d^k/4$

매칭 규칙 : $d^i < p_1$, $1 \le i \le n$, and $d^{qwe} < p_2$

유사한 방법으로 두 번째 단계의 매칭을 수행한다. 다른 점은 4차원 프로파일 특징 벡터 대신에 30차원 웨이블릿 특징 벡터를 사용한다는 점과, 매칭 조건의 임계값으로 p_3과 p_4를 사용한다는 것이다.

좋은 성능을 위해서는 적절한 임계값 사용이 중요하다. p_1 과 p_2 는 낮은 정확률을 허용하면서 높은 재현률을 유도해내야 하며, p_3과 p_4는 높은 정확률과 재현률을 유도해야 한다. 임계값들은 실험에 의해 적절한 값으로 설정했다.

4. 성능 측정과 오류 분석

검색 성능 측정을 위해 아래와 같이 정의되는 재현률(recall)과 정확률(precision)을 사용하였다.

재현률 = (검색된 적합 단어의 수) / (적합 단어의 총 수)

정확률 = (검색된 적합 단어의 수) / (검색된 단어의 총 수)

4.1 문자 단위 성능 측정

단어 검색의 성능 측정을 위한 실험으로 문자별 검색 성능을 알아볼 필요가 있다. 성능에 영향을 주는 것으로는 문서 영상의 품질, 해상도, 폰트 등을 들 수 있다.

실험에 사용된 데이터는 국문으로 이루어진 기술저널로서 총 238페이지 분량이며, 해상도는 300dpi로 스캔받은 문서 영상이다.

표 1은 프로파일 특징만으로 문자를 검색했을 때와 웨이블릿 특징만으로 문자를 검색했을 경우, 2단계를 통해 문자를 검색했을 경우 구해진 각 단계별 정확률과 재현률을 보여준다.

표 1. 각 단계별 성능

	프로파일	웨이블릿	2단계 검색
재현률	99.33	89.33	87.33
정확률	14.78	33.33	57.71

표 1에서 보듯이 프로파일 특징을 사용하여 문자를 검색했을 경우 낮은 정확률과 높은 재현률을 볼 수 있다.

그림 4는 유사한 형태의 문자들로서 프로파일 특징이나 웨이블릿 특징으로는 문자의 정보를 추출하기에 난해한 형태의 문자들이다.

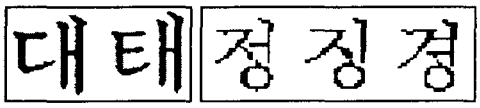


그림 4. 유사한 형태의 문자

중앙 부분에 의해 문자들이 구분되는 경우는 외곽 쪽의 삐침이나 획에 의해 프로파일 투영으로는 분별하기 어렵다. 따라서 복잡한 문자의 중앙 부분에 획들이 모여 있는 경우 문자의 정보를 추출하기 어려운 원인이 되기도 한다.

웨이블릿 특징에서의 검색 오류는 획의 위치나 두께, 길이 등이 유사한 경우에 발생할 수 있는데 이는 획의 정보를 잘 표현하는 웨이블릿 계수의 근사한 값 때문이다. 이런 경우 역시 정확률이 낮아지게 된다. 또한 프로파일 특징에 비해 웨이블릿 특징은 획의 수가 적고 간단한 경우 문자의 특징 표현이 약하기 때문에 검색 효율을 떨어뜨린다.

또 다른 원인으로는 거의 드물게 나타나지만 정확하지 못한 단어나 문자의 분할로 문자의 정보가 변형이 되는 경우와, 문서 영상의 노이즈, 정보의 손실 등이 검색 성능에 영향을 준다. 정보 손실의 경우는 스캔된 문서 영상에서는 폰트나 획의 두께가 다양하기 때문에 높은 재현률을 요구하는 프로파일 특징으로는 문자를 추출해 내지 못하는 원인이 될 수 있다.

4.2 단어 단위 성능 측정

300dpi 해상도의 문서 영상에서의 실험 결과는 초당 199,918 단어의 검색 속도를 보이고, 성능은 99.6% 재현률과 97.5% 정확률을 얻었다. 30개의 웨이블릿 특징 벡터를 가지고 1 단계 처리에 의한 검색만 할 경우, 유사한 수준의 재현률과 정확률은 얻을 수 있으나, 초당 33,063단어의 저하된 검색 속도를 보였다. 따라서 2단계 검색 시스템은 기존의 성능을 유지하며 대략 6배 검색 모듈에 의해 속도의 증가를 얻어낼 수 있었다.

저 품질의 문서 영상에서도 단어 검색 성능을 평가하기 위해 한국정보과학회 저널에서 얻은 200dpi로 스캔받은 문서 영상에서 검색을 실험하였다. 재현률과 정확률은 각각 91.0%, 88.0%을 얻었다.

따라서 문서의 인쇄 품질과 스캔시의 해상도가 검색 성능에 많은 영향을 미치는 것을 볼 수 있으며, 이러한 저 품질의 문서 영상에서 좋은 성능을 발휘할 수 있는 알고리즘의 개발이 요구되어진다.

5. 결론

한글 문서 영상의 단어 검색 시스템을 개발하였으며, 예비 실험 결과를 제시하였다. 이 시스템은 정보 검색을 위한 방대한 데이터베이스에서 온라인 단어 검색에 유용하게 사용될 수 있음을 실험결과에 나타냈다. 향후 연구로 저 품질의 문서 영상에서의 성능 개선을 위한 특징 추출 알고리즘 개발이 이루어져야 한다. 또한 문자의 폰트나 속성 정보를 추출하는 알고리즘을 개발하여, 그런 정보들을 문서 영상 검색에 적용할 수 있다면 보다 나은 검색 시스템을 구현할 수 있을 것이다.

참고문헌

[1] 오일석, 김수형, 유태웅, 곽희규, "문서 영상 처리 기술과 디지털 라이브러리," *정보과학회지*, 8 2002.

[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] M. Mitra and B.B. Chaudhuri, "Information retrieval from documents: a survey," *Information Retrieval*, pp.141-163, 2000.

[4] "SCRIBBLE: SRI's keyword spotting system," http://www.erg.sri.com/ projects/scribble.

[5] National Digital Library, http://www.dlibrary.go.kr/.

[6] 곽희규, *문서 영상의 단어 단위 분할 및 단어 영상의 속성 추출에 관한 연구*, 전남대학교 박사학위논문 2001.

[7] 김혜금, 양진호, 이진선, 오일석, "웨이브렛을 이용한 영상기반 인쇄 한글 단어 검색," *한국정보과학회 논문지*, 제28권, 제2호, pp.91-103, 2 2001.