문자 별 특징 모델을 이용한 한글 문서 영상에서 키워드 검색

박 상 철[†] · 김 수 형[††] · 최 덕 재[†††]

요 약

본 논문에서는 저 품질의 한글 문서 영상에서 OCR 기반 검색 시스템의 대안으로 키워드 검출 시스템(Keyword Spotting)을 제안하고 OCR 기반 문서 검색 시스템과 비교한다. 제안 시스템은 문자 분할, 키워드 특징 추출 그리고 단어 매칭으로 구성된다. 문자 분할 단계에서는 인접한 두 문자간의 연결을 효과적으로 분리하면서 문자 넓이 값의 분산이 최소가 되도록 하는 문자 분할 방법을 제안한다. 키워드 특징은 서체별 문자 모델의 결합으로 구성한다. 단어 매칭 단계에서는 문자 매칭에 기반한 단어 대 단어 매칭 방법을 적용한다. 본 논문에서 제안한 키워드 검출 시스템의 성능을 평가하기 위해 한글 문서 영상을 대상으로 OCR 기반 문서 검색 시스템과 비교하였다. 그 결과 한글 글자 크기가 작고 문서의 상태가 좋지 않은 경우 제안한 키워드 검출 시스템에 의한 검색 성능이 OCR 기반 검색 시스템 보다 우수함을 입증하였다.

키워드 : 문서 영상 검색, 키워드 검출, 광학문자인식

Keyword Spotting on Hangul Document Images Using Character Feature Models

Sang Cheol Park[†] · Soo Hyung Kim[††] · Deok Jai Choi[†††]

ABSTRACT

In this paper, we propose a keyword spotting system as an alternative to searching system for poor quality Korean document images and compare the proposed system with an OCR-based document retrieval system. The system is composed of character segmentation, feature extraction for the query keyword, and word-to-word matching. In the character segmentation step, we propose an effective method to remove the connectivity between adjacent characters and a character segmentation method by making the variance of character widths minimum. In the query creation step, feature vector for the query is constructed by a combination of a character model by typeface. In the matching step, word-to-word matching is applied base on a character-to-character matching. We demonstrated that the proposed keyword spotting system is more efficient than the OCR-based one to search a keyword on the Korean document images, especially when the quality of documents is quite poor and point size is small.

Key Words : Document Image Retrieval, Keyword Spotting, OCR

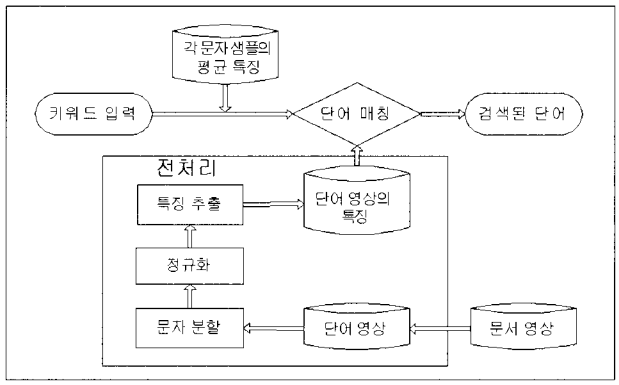
1. 서 론

종이 문서로부터 스캔되어 디지털 도서관에 저장되어 있는 문서는 제목, 요약, 키워드만 텍스트로 저장되어 있고 전문(full-text)은 비트맵 영상 형태로 저장되어 있다. 따라서 극히 제한된 키워드로의 검색은 가능하나 전문을 대상으로 하는 검색은 불가능하다는 한계를 안고 있다[1]. 문서 영상에서 전문 검색이 가능하게 하는 방법으로는 문자 인식(OCR)을 이용하는 방법과 영상-기반 방법이 있다. 이는 다른 표현으로 키워드 검출(Keyword Spotting)이라고도 한다.

OCR을 이용하는 방법은 문서 영상의 문자를 인식하여 기계가 판독할 수 있는 텍스트 형태로 변형을 한 후 텍스트 매칭을 이용하여 검색한다. 이 방법의 문제점은 검색 문자가 오인식될 경우 검색이 불가능하다는 점이다. 특히 문서의 훼손이 심한 경우 검색 성능이 급격히 떨어지는 단점이 있다. 이런 문제를 해결하기 위해 후처리 기법은 필수적이지만[2, 3], 많은 시간과 고비용 연산이 요구되어 새로운 접근이 필요하다[4].

영상-기반 키워드 검색은 OCR을 이용한 단어 검색 방법의 대안이 될 수 있다. 이 방법은 키워드 영상을 문서 영상의 일부분과 매칭하여 유사정도를 기준으로 검색하는데, 크게 전처리 단계와 검색 단계로 나뉘어 진다. 전처리 단계는 문서 영상을 단어 단위로 분할하고, 각각의 단어 영상들의

※ 이 논문은 2004년도 산업자원부의 지역혁신인력양성사업의 지원에 의하여 연구되었음.
† 준 회 원 : 전남대학교 자연과학대학 전산학과 박사과정
†† 정 회 원 : 전남대학교 자연과학대학 전산학과 교수
††† 종신회원 : 전남대학교 자연과학대학 전산학과 교수
논문접수 : 2005년 6월 21일, 심사완료 : 2005년 8월 18일



(그림 1) 평균 문자 특징을 이용한 단어 영상 시스템

특징 정보를 데이터베이스에 저장하는 과정이다. 이는 검색 단계에서 처리 시간 비용을 절감하기 위한 방편이다. 검색 단계는 키워드 특징과 데이터베이스에 저장된 단어 영상의 특징을 비교하여 영상간의 유사성을 판단하는 단계이다[4].

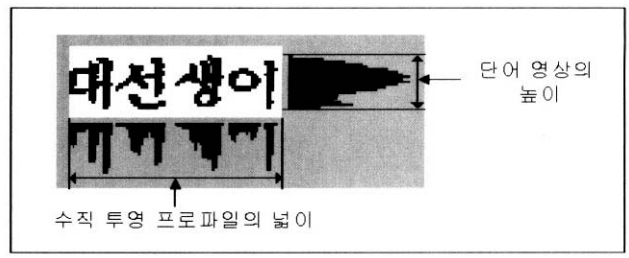
최근 몇 년간 영문 문서 영상에서의 영상-기반 키워드 검색에 대한 연구 결과들이 발표되었다[4-9]. 이들 중 대표적인 Lu 등의 시스템[6]은 단어 영상을 특징 기호 열(string of feature codes)로 표현하였고, 각 문서 영상을 단어 열의 연속으로 표현하며, 근사한 스트링 매칭(inexact string matching)을 적용하여 문서 내의 단어와 키워드를 매칭한다.

한글 문서 영상에서 영상기반 단어 검색에 대한 연구는 오일석 등[11, 12]의 연구가 있다. 오일석 등은 한글 단어를 검색하기 위해 곽희규 등[13]의 시스템을 이용하여 문서 영상으로부터 단어 영상을 추출한 후, 단어 영상 데이터베이스를 구성하였으며, 단어 영상은 다시 문자 단위로 분할된다. 문서 영상에 사용된 동일한 폰트를 참고하여 문서 편집기로 문자 집합을 생성하였고, 이를 인쇄한 후 스캔하여 질의 영상으로 사용하였다. 처리시간을 최소화하기 위해 두 단계 매칭 방법을 사용하였는데, 1단계에서는 프로파일 특징을 이용하였고, 2단계에서는 Harr 웨이블릿 계수 중 가장 큰 값을 갖는 30개의 특징을 선택하여 사용한다.

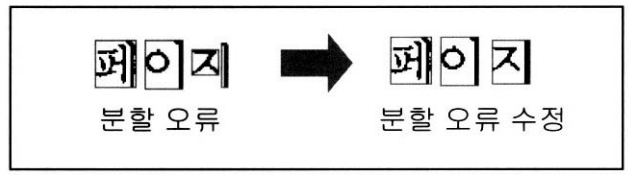
(그림 1)은 본 논문에서 제안한 키워드 검출(Keyword Spotting) 시스템의 블록 다이어그램이다. 본 논문에서는 정창부 등 [14]의 시스템을 이용하여 문서 영상으로부터 개별 단어 영상을 분할한 후, 이들 단어 영상을 데이터베이스에 미리 저장해 두었다고 가정한다. 단어 영상은 문자 영상으로 분할되고 일정한 크기로 정규화 된다. 정규화된 문자 영상은 메쉬 특징으로 표현되어 데이터베이스에 저장된다. 키워드가 입력되면 데이터베이스에서 키워드와 유사한 단어 영상을 검색한다.

2. 문자 분할

한글 문서에서는 문자와 문자 사이에 공백이 존재한다. 그리고 문자의 넓이가 거의 일정하고 넓이와 높이의 크기가 같다. 본 논문에서는 위 사실에 근거하여 문자 분할 알고리



(그림 2) 4문자로 구성된 단어 영상의 투영 프로파일



(그림 3) 후처리에 의한 문자 분할 오류 수정

즘을 제안한다. 처리 과정은 다음과 같이 네 단계로 구성된다.

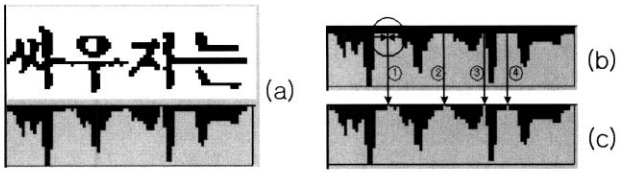
첫 번째 단계에서는 단어 영상이 몇 개의 문자로 구성되었는지 추정한다. 단어 영상의 높이로 수직 투영 프로파일의 넓이를 나누면 단어영상을 구성하는 문자수를 쉽게 추정할 수 있다. 이때 잡음이나 문자 획의 손상으로 인해 문자수 추정의 오류가 발생할 수 있으므로, 최대 2 가지의 추정치를 구한다. (그림 2)는 한글 단어 영상에서 수직 투영 프로파일의 넓이와 단어 영상의 높이를 나타낸다.

두 번째 단계에서는 추정된 문자수에 따라 문자 분할 점을 탐색한다. 추정된 문자수로 단어 영상을 균등하게 분할하여 분할 대상 점을 선택한다. 분할 대상 점의 수직 투영 프로파일 값이 0인 경우 해당 지점을 분할 점으로 선택하고, 그렇지 않을 때는 분할 대상 점으로부터 투영 프로파일을 좌·우로 동시에 이동하면서 투영 프로파일 값이 0인 지점을 찾아 분할 점으로 선택한다. 앞서 추정된 문자수가 2가지라면 두 가지 형태의 분할이 가능하다

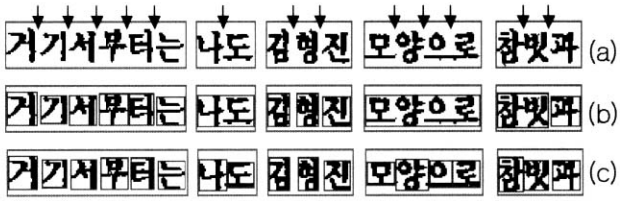
세 번째 단계에서는 분할 결과를 선택한다. 이는 한글 문자의 넓이가 일정하다는 가정으로부터 문자 분리가 올바르게 수행될 경우 문자 넓이의 분산은 그렇지 않은 경우보다 적은 값을 갖는 사실을 근거로 설계된 선택 기준이다. 즉, 문자수가 2가지로 추정되었을 경우 분할된 문자들의 넓이 값 분산이 최소가 되는 분할결과를 선택한다.

네 번째 단계에서는 문자 분할 오류를 수정하는 후처리가 수행된다. 특정 문자가 "ㅏ", "ㅐ", "ㅣ", 등과 같은 수직 모음을 포함하면 해당 문자는 두 조각으로 분리될 가능성이 있다. 따라서 분할된 문자의 넓이가 단어 영상에서 가장 큰 문자 넓이의 β% 이하의 크기이면 해당 문자를 그 좌측의 문자와 결합한다. β는 실험에 의하여 결정하였다. (그림 3)은 끝 문자의 모음 때문에 발생하는 문자 분할 오류를 위 방법으로 수정한 예이다.

(그림 4)의 (a)와 같이 두 문자가 서로 연결되었을 경우 위 제안방법에 의한 문자 분할은 실패할 수밖에 없다. 따라서 우리는 연결된 두 문자를 단절시키기 위해 (b)처럼 수직



(그림 4) 문자간 연결 성분 제거



(그림 5) 문자 분할 결과

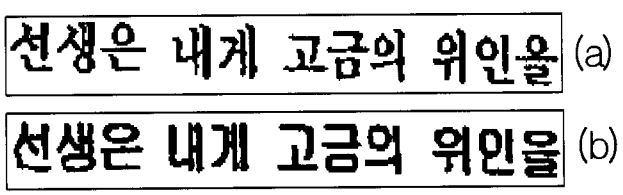
투영 프로파일의 상위 α만큼 삭제(α-cut)하였다. (b)는 α-cut을 적용한 투영 프로파일을 나타내는데, ①, ②, ③ 그리고 ④의 지점이 분할점 후보가 된다는 사실을 알 수 있다. (b)의 첫 번째 분할 점에 그려진 원안의 마주보는 기호인 →←는 두 세그먼트 사이의 간격이다. 이 부분을 이등분하여 두 세그먼트를 분할한다. α는 실험에 의하여 수직 투영 프로파일의 평균값의 7%로 결정하였다.

(그림 5)의 (a)는 입력 영상이다. 화살표는 문자의 분할 목표점이다. (b)는 α-cut을 적용하지 않은 문자 분할 결과이다. (c)는 α-cut을 적용한 후의 문자 분할 결과이다. 문자들이 서로 연결된 경우 α-cut은 문자 분할 목표점의 위치에서 문자가 분할되도록 돕는다.

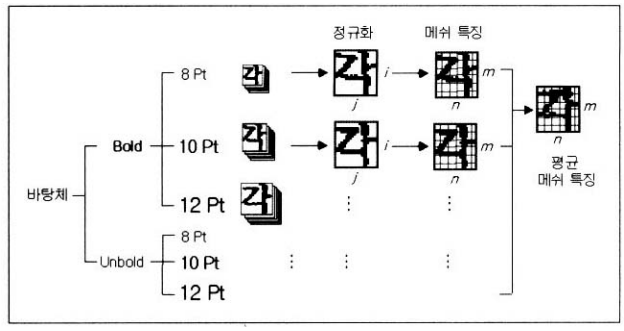
3. 키워드 특징 추출

명조체류와 고딕체류는 한글문서에서 널리 사용되는 서체이며 이들의 문자 형태는 현격히 다르다. 영상 기반 단어 검색은 영상의 형태에 의존하기 때문에 서체가 다른 문서 영상에서의 단어 검색 성능은 낮을 수밖에 없다. 따라서 검색 효율을 높이기 위해 서체를 분류한 후 검색하는 방법이 유용하다. 본 논문에서는 데이터가 미리 서체에 따라 분류되었다고 가정한다. (그림 6)의 (a)는 바탕체로 작성된 문서 영상의 일부이다. (b)는 (a)와 같은 내용의 굴림체 영상이다. 의미상으로는 같지만 서로 다른 서체로 작성된 문서 영상이기 때문에 획의 위치와 두께가 다르다.

문서를 여러 차례 복사하거나 스캔하다 보면 텍스트 획 정보가 손상되고 잡음이 추가될 수 있다. 문서 영상의 기울



(그림 6) 바탕체와 굴림체의 한글 영상



(그림 7) 바탕체 문자의 평균 특징

어짐 교정을 수행하는 과정에서도 추가적으로 정보의 변질이 생길 수 있다. 따라서 이러한 잡음의 영향을 최소화 하고 원래의 정보를 획득하는 방법이 필요하다. 본 연구에서는 각 서체별로 6가지 폰트 속성(2가지 두께 및 3가지 크기의 조합) 각각에 해당하는 문자 영상들을 수집하여 이들의 36차원 메쉬 특징을 평균하고, 이를 서체별 해당 문자 모델로 사용한다. 4절의 매칭 단계에서 사용되는 키워드 특징은 서체별 문자 모델의 결합으로 구성된다.

(그림 7)은 바탕체로 쓰여진 문자 "각"의 평균 특징을 추출하는 과정을 도시하고 있다. 문자 "각"의 6가지 폰트 속성에 해당하는 훈련데이터들을 $i \times j$ 크기로 정규화한다. 정규화된 영상에서 $m \times n$ 메쉬 특징을 추출한 후, 이들 메쉬 특징을 평균하여 바탕체 "각"을 표현하는 모델로 사용한다. 우리는 실험에서 i와 j를 각각 36으로 하였으며, m과 n을 6으로 하였다. 따라서 36차원의 메쉬 특징이 사용되었다.

4. 단어 매칭

키워드 영상이 k개의 문자로 구성되었다고 가정하고 $Q(C_1^q, C_2^q, \ldots, C_k^q)$라고 표기하자. 이때 검색 대상 영상에서 매칭에 참여하는 k개의 문자 영상을 목적 영상이라 하고 $T(C_1^t, C_2^t, \ldots, C_k^t)$로 표기한다. 여기서 위 첨자 q와 t는 각각 키워드 영상과 목적 영상을 의미한다. C_i는 단어 영상의 i번째 문자에서 추출한 v차원 특징 벡터이고, $C_i = (c_{i,1}, c_{i,2}, \ldots, c_{i,v})$로 표기한다. 목적 영상과 키워드 영상의 i번째 문자의 유사성 판단은 식 (1)에 근거한다. T_c는 문자 단위 유사성 여부를 판단하기 위한 임계값이다.

$$\begin{cases} \text{if} & Dist(C_i^q, C_i^t) < T_c, \quad then \quad C_i^q = C_i^t \\ else & \qquad\qquad\qquad\qquad\quad C_i^q \neq C_i^t \end{cases} \tag{1}$$

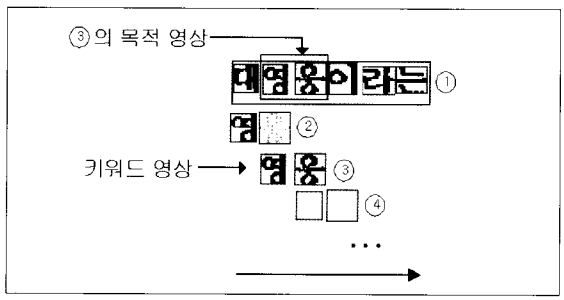
$$\text{where} \quad Dist(C_i^q, C_i^t) = \sum_{j=1}^{v} \left| c_{i,j}^q - c_{i,j}^t \right| \tag{2}$$

k개의 문자가 연속적으로 매칭된 경우 식 (3)에 근거하여 두 단어의 매칭 여부를 판단한다. T_w는 단어 단위 매칭을 위한 임계값이다.

$$\begin{cases} \text{if} & Dist(Q, T) < T_w, \quad then \quad Q = T \\ else & \qquad\qquad\qquad\quad Q \neq T \end{cases} \tag{3}$$

$$\text{where} \quad Dist(Q, T) = \frac{1}{k}\sum_{i=1}^{k} Dist(C_i^q, C_i^t) \tag{4}$$

(그림 8)은 키워드 영상보다 더 많은 문자를 갖는 영상과의 매칭 과정을 도시하고 있다. ①은 검색 대상 영상이고 ②는 첫 번째 문자가 서로 달라서 매칭이 실패한 경우이다. ③은 두 문자 각각이 매칭 조건을 만족하고, 단어간의 매칭 조건도 만족하여 매칭이 성공한 경우이다. ④ 이후에도 키워드와 동일한 목적 영상이 존재할 수 있으므로 매칭을 계속 수행한다.



(그림 8) 매칭 과정

5. 실험 결과

5.1 실험 환경

"백범일지" 일부를 마이크로소프트 워드를 이용하여 A4용지 20쪽 분량의 문서 파일로 만들었다. 이를 서로 다른 폰트 속성(서체: 바탕체, 굴림체; 크기: 8, 10, 12; 두께: bold, unbold)으로 편집하였다. 이 문서 파일을 삼성 ML-8065 프린터로 출력한 후, 제록스 Document Centre 285 PLUS G 복사기로 복사하되, 복사 결과물을 다시 복사하는 방식으로 8회 복사하였다. EPSON GT-30000 스캐너를 사용하여 200DPI로 스캔하여 저장하였다. 이 문서 영상을[14]의 시스템을 이용하여 단어 단위 영상으로 분할하였다. 2절에서 제안된 문자 분할 방법으로 단어 영상을 문자 단위로 분할하고 36×36의 크기로 정규화하였다. 정규화된 문자 영상은 36차원 메쉬 특징으로 표현된다. 전체 데이터에서 절반은 훈련데이터, 나머지는 테스트 데이터로 사용하였다. 실험에 사용된 기자재는 Pentium-4 CPU 2.80GHz와 1GB RAM 자원을 갖는 개인용 PC이다.

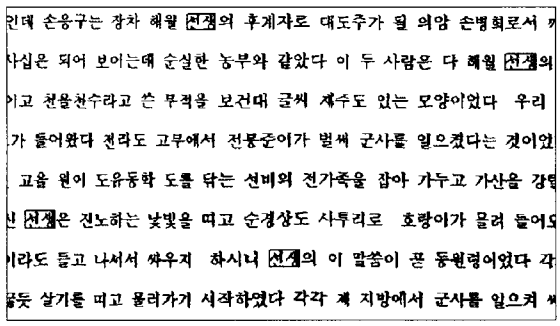
실험에 사용된 한글 문서 영상에서는 두 문자 단어(두 문자로 이루어진 단어)와 세 문자 단어(세 문자로 이루어진 단어)의 출현 빈도가 가장 높다. 따라서 두 문자 단어와 세 문자 단어로 30개 키워드를 구성하였다. 먼저 훈련 데이터에서 두 문자 단어와 세 문자 단어를 출현빈도에 따라 정렬하고, 두 문자 단어 중 출현빈도 상위 20개를 선정한다. 세 문자 단어도 같은 방법으로 10개를 선택한다. 한글 문서 영상에서는 두 문자 단어가 더욱 빈번하게 출현하므로 두 문자 단어에 대해서는 키워드 개수를 더 많이 포함하였다.

5.2 검색 성능

<표 1>은 테스트 데이터의 문서 영상을 문서 인식 소프트웨어인 아르미 6.0[15]로 인식한 후 검색한 결과와 제안 시스템을 이용한 검색 결과를 비교하고 있다. OCR 기반 검색의 경우 모든 데이터에 대해서 100% 정확율을 나타내고 있다. 이는 내용이 다른 단어가 키워드로 오인식될 경우 정확율에 영향을 주는데, 본 실험에서 사용한 30개의 키워드는 해당 사항이 없었기 때문이다. 이와 반대로 키워드에 대응하는 문자들이 오인식되면 재현율에 영향을 주는데, OCR을 이용한 검색 성능은 대부분 이 오류에 의해 결정된다.

<표 1> OCR 기반 검색 방법과 제안된 시스템의 검색 성능 비교

폰트			OCR 기반 검색 (%)			제안 시스템(%)			평가
서체	굵기	크기	재현율	정확율	조화평균 F (A)	재현율	정확율	조화평균 F (B)	B-A
바탕	굵게	**8**	**39.21**	**100**	**56.33**	**67.84**	**67.84**	**67.84**	**11.51**
		10	77.53	100	87.34	80.34	80.34	80.34	-7.00
		12	87.22	100	93.17	79.74	79.74	79.74	-13.43
	보통	8	55.95	100	71.75	70.80	70.80	70.80	-0.95
		10	81.94	100	90.07	71.68	71.68	71.68	-18.39
		12	92.51	100	96.11	76.21	76.21	76.21	-19.90
굴림	굵게	**8**	**25.11**	**100**	**40.14**	**78.76**	**78.76**	**78.76**	**38.62**
		10	**55.51**	**100**	**71.39**	**85.02**	**85.02**	**85.02**	**13.63**
		12	75.33	100	85.93	83.41	83.41	83.41	-2.52
	보통	**8**	**21.15**	**100**	**34.92**	**69.16**	**69.16**	**69.16**	**34.24**
		10	59.03	100	74.24	74.01	74.01	74.01	-0.23
		12	70.07	100	82.40	81.06	81.06	81.06	-1.34
평균			61.71	100.00	76.32	76.50	76.50	76.50	0.18



(그림 9) 검색 결과 영상

재현율(Recall)이란 데이터베이스 내에 검색어에 해당하는 단어 중에서 실제 검색된 단어의 비율을 말한다. 정확율(Precision)이란 검색된 결과 중에서 검색어와 일치하는 단어의 비율을 말한다. 조화평균-F(F-measure, harmonic mean F)는 재현율과 정확율을 결합한 단일 척도이다[16]. 식 5-7은 재현율, 정확율 그리고 조화평균-F를 나타내는 수식이다.

$$Recall = \frac{|Ra|}{|R|} \times 100 \tag{5}$$

$$Precision = \frac{|Ra|}{|A|} \times 100 \tag{6}$$

$$F-measure = \frac{2 \times Recall \times Precision}{Recall + Precision} \tag{7}$$

여기서 $|R|$는 데이터베이스의 단어 중 검색어에 해당하는 단어의 개수이고, $|Ra|$는 검색 결과 중 검색어에 해당하는 단어의 개수이다. 그리고 $|A|$는 검색된 결과 중 검색어에 해당하는 단어의 개수와 그렇지 않은 단어의 개수의 합을 의미한다.

제안 시스템의 성능은 재현율과 정확율이 같은 값을 갖도록 하면서 측정하였다. 키워드 검출 시스템에서 재현율과 정확율은 반비례 관계에 있다. 따라서 재현율과 정확율이 같을 때를 기준으로 검색 성능을 비교할 수 있다. 제안 시스템과 OCR 기반 검색 사이의 성능 차이는 <표 1>에서 평가 항목이 양의 값을 가질 때 제안 시스템이 우수하다고 본다.

OCR을 이용한 검색 성능은 굴림체일 경우와 문자 크기가 작을수록 현저하게 낮다. 그 반면에 본 논문에서 제안한 키워드 검출 시스템은 바탕체 보다 굴림체에서 높은 성능을 보여준다. 또한 문자 크기가 작을 경우 OCR을 이용한 방법에 비해 월등히 우수하다. 결론적으로 한글 문서 영상에서 굴림체이거나 글자 크기가 작고 문서의 상태가 좋지 않은 경우 OCR을 이용한 검색 보다 단어 검출에 의한 검색이 훨씬 더 유리하다. (그림 9)는 실험데이터에서 "선생"을 검색한 결과이다.

6. 결 론

본 논문에서는 한글 문서 영상을 위한 키워드 검출 시스템을 제안하고 OCR 기반 문서 검색 시스템과 비교하였다. 단어 검출 시스템은 문자 분할, 키워드 특징 추출 그리고 단어 대 단어 매칭으로 구성된다. 문자 분할 단계에서는 인접한 두 문자간의 연결을 효과적으로 분할하면서 문자 넓이 값의 분산이 최소가 되도록 하는 문자 분할 방법을 제안하였다. 키워드 특징은 서체별 문자 모델의 결합으로 구성하였다. 단어 매칭 단계에서는 문자 매칭에 기반한 단어 대 단어 매칭 방법을 적용한다. 본 논문에서 제안한 단어 검출 시스템의 성능을 평가하기 위해 한글 문서 영상을 대상으로 OCR 기반 문서 검색 시스템과 비교하였다. 그 결과 한글 문서 영상에서 굴림체이거나 글자 크기가 작고 문서의 상태가 좋지 않은 경우 OCR을 이용한 검색 보다 키워드 검출에 의한 검색이 더 유리함을 입증하였다.

문서 영상의 품질은 원문의 상태와 기계의 상태에 따라 다양하게 나타난다. 특히 고문서나 원본 자체의 품질이 낮은 문서의 경우 어떤 고성능의 기계장치로도 좋은 품질의 영상을 획득하기 어렵다. 이러한 저 품질의 문서 영상은 OCR의 에러를 유발한다. 제안된 시스템은 이러한 저 품질 문서 영상의 내용기반 전문 검색을 위한 도구로써 활용 가능하다. 향후 연구 내용으로 키워드 검색 시스템에 적합한 분류기를 연구하고자 한다.

참 고 문 헌

[1] 오일석, 김수형, 유태웅, 곽희규, "문서 영상 처리 기술과 디지털 도서관", 정보과학회지, 제20권 제2호, pp.24-34, 2002.

[2] M. Ohta, A. Takasu, and J. Adachi, "Retrieval methods for English-text width missrecognized OCR characters," *Proceedings of 4th International Conference on Document Analysis and Recognition*, Vol.2, pp.950-955, 1997.

[3] K. Marukawa, T. Hu, H. Fujisawa, and Y. Shima, "Document retrieval tolerating character recognition errors-evaluation and application," *Pattern Recognition*, Vol.30, No.8, pp.1361-1371, 1997.

[4] D. Doermann, "The retrieval of document images: a brief survey," *Proc. ICDAR97*, Ulm, pp. 945-949, 1997.

[5] F. Chen, L. Wilcox, and D. Bloomberg, "Word spotting in scanned images using hidden markov models," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pp.1-4, 1993.

[6] Y. Lu and C. L. Tan, "Word searching in document images using word portion matching," *Fifth IAPR International Workshop on Document Analysis Systems*, USA, pp.319-328, 2002.

[7] Y. Lu, L. Zhang, and C. L. Tan, "A search engine for imaged documents in PDF files," *27th Annual International ACM SIGIR Conference*, UK, 2004.

[8] J. DeCurtins and E. Chen, "Keyword spotting via word

shape recognition," *Proc. SPIE Document Recognition II*, pp.270-277, 1995.

[9] F. R. Chen, L.D. Wilcox, and D.S. Bloomberg, "A comparison of discrete and continuous hidden Markov models for phrase spotting in text images," *Proc. Document Analysis and Recognition*, Vol.1, pp.398-402, 1995.

[10] C. L. Tan, W. Huang, Z. Yu, and Y. Xu, "Image document text retrieval without OCR," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, Vol.24, No.7, pp.838-844, July, 2002.

[11] 김혜금, 양진호, 이진선, 오일석, "웨이브렛을 이용한 영상 기반 인쇄 한글 단어 검색", 한국정보과학회 논문지, 제28권 제2호, pp.91-103, 2001.

[12] I. S. Oh, Y. S. Choi, J. H. Yang, S. H. Kim, "A Keyword Spotting System of Korean Document Images," *Proc. 5th International Conference on Asian Digital Libraries*, Singapore, p.530, Dec., 2002.

[13] 곽희규, "문서 영상의 단어 단위 분할 및 단어 영상의 속성 추출에 관한 연구," 전남대학교 전산통계학과 박사학위논문, 2001.

[14] C. B. Jeong and S. H. Kim, "A Document Image Preprocessing System for Keyword Spotting," *Proc. International Conference on Asian Digital Libraries*, China, pp.440-443, Dec., 2004.

[15] http://www.perceptcom.com/

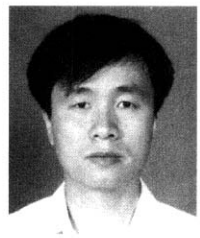
[16] R. B. Yates and B. R. Neto, "Modern Information Retrieval," ACM press, pp.75-82, 1999.

박 상 철

e-mail : sanchun@iip.chonnam.ac.kr
1999년 조선대학교 전자계산학과(학사)
2001년 조선대학교 전자계산학과(이학석사)
2003년~현재 전남대학교 전산학과 박사과정
관심분야 : 패턴인식, 문서영상 정보검색, 의료영상

김 수 형

e-mail : shkim@chonnam.ac.kr
1986년 서울대학교 컴퓨터공학과(학사)
1988년 한국과학기술원 전산학과(공학석사)
1993년 한국과학기술원 전산학과(공학박사)
1993년~1996년 삼성전자 멀티미디어 연구소 선임연구원
1997년~현재 전남대학교 전자컴퓨터정보통신공학부 교수
관심분야 : 인공지능, 패턴인식, 문서영상 정보검색, 유비쿼터스 컴퓨팅



최 덕 재

e-mail : dcho@chonnam.ac.kr
1982년 서울대학교 컴퓨터공학과(학사)
1984년 한국과학기술원 전산학과(공학석사)
1995년 Missouri-Kansas대학교 컴퓨터공학(공학박사)
1996년~현재 전남대학교 전자컴퓨터 정보통신공학부 교수
관심분야 : 서비스 관리, 망 관리, 유비쿼터스컴퓨팅