영상 대 영상 매칭을 이용한 한글 문서 영상에서의 단어 검색

박 상 철[†] · 손 화 정[††] · 김 수 형[†††]

요 약

본 논문에서는 두 단계 이미지 매칭을 이용하여 한글 문서영상에서 사용자 검색어를 빠르고 정확하게 검색할 수 있는 시스템을 제안한다. 본 시스템은 문자 분리, 검색어 영상 생성, 특징 추출 그리고 이미지 매칭 과정으로 구성된다. 매칭 과정에서 차원이 다른 두 가지 특징 벡터를 이용한다. 8쪽 분량의 문서 영상을 한국정보과학회 웹사이트에서 다운로드하였고, 그 문서로부터 1600개의 한글단어 영상을 획득하여 실험데이터로 사용하였다. 그 결과 제안한 시스템은 기존에 제안된 영상-기반 한글 단어 검색 시스템보다 성능이 크게 향상되었음을 알 수 있었다.

키워드 : 한글 문서영상 검색, 문서 영상처리, 광학문자인식

Keyword Spotting on Hangul Document Images Using Image-to-Image Matching

Sang Cheol Park[†] · Hwa Jeong Son[††] · Soo Hyung Kim[†††]

ABSTRACT

In this paper, we propose an accurate and fast keyword spotting system for searching user-specified keyword in Hangul document images by using two-level image-to-image matching. The system is composed of character segmentation, creating a query image, feature extraction, and matching procedure. Two different feature vectors are used in the matching procedure. An experiment using 1600 Hangul word images from 8 document images, downloaded from the website of Korea Information Science Society, demonstrates that the proposed system is superior to conventional image-based document retrieval systems.

Key Words : Keyword Spotting On Hangul Document Images, Document Image Processing, OCR

1. 서 론

컴퓨터 사용이 보편화되고 인터넷의 발전으로 전자 문서가 보편적으로 사용되어 종이 없는 시대의 도래가 예견되었던 것과 달리 종이 소비량은 오히려 증가하고 있으며, 많은 인쇄 문서나 도서의 생산으로 이어지고 있다. 종이문서와 더불어 오래된 고문서의 검색을 위해서는 이들 문서의 디지털화는 필수적이다[1].

문서 영상에서의 단어 검색 방법 중 하나는 OCR 소프트웨어를 이용한다[2]. OCR은 스캐너로부터 문서 영상을 획득하는 과정, 영상의 왜곡 등을 교정하는 전처리 과정, 문서의 구조를 분석하고 문자를 분할하는 과정 그리고 이를 인식하는 단계들로 구성된다. 하지만 문서의 획득 과정에서 발생하는 왜곡과 OCR 기술의 한계로 인해, 인식된 문자의 오류를 정정하는 후처리 과정은 필수적이다. 후처리 과정은 노동 집약적이기 때문에 이를 자동화하는 새로운 접근이 필요하다.

또 다른 검색 방법은 영상-기반 방법이다. 이 방법은 단어 검색이 수행되기 이전에 문서 영상과 단어 영상들의 특징 정보를 데이터베이스에 저장한다. 검색은 검색어에서 추출한 특징 정보와 데이터베이스에 저장된 단어 영상의 특징 정보를 비교하는 절차를 통해 이루어진다[2]. 이 방법은 문서 영상에 사용된 언어에 상관없이 전문 검색이 가능하게 한다. 또한 인터넷이나 전자 도서관에 많은 문서 영상이 있을 경우, 사용자는 문서 영상의 내용을 모르기 때문에 적어도 한번은 다운로드하여 원하는 문서인지 확인하여야 한다. 이러한 상황에서 영상기반 단어 검색이 지원된다면 모든 문

※ 이 논문은 2004년도 한국학술진흥재단의 지원(KRF-2004-041-D00631)에 의하여 연구되었음
† 준 회 원 : 전남대학교 자연과학대학전산학과 박사과정
†† 준 회 원 : 전남대학교 자연과학대학전산학과 박사과정
††† 정 회 원 : 전남대학교 자연과학대학전산학과 교수
논문접수 : 2004년 12월 17일, 심사완료 : 2005년 4월 18일

서를 다운로드 하지 않고 검색어가 포함된 문서만을 선택적으로 다운로드할 수 있다[1].

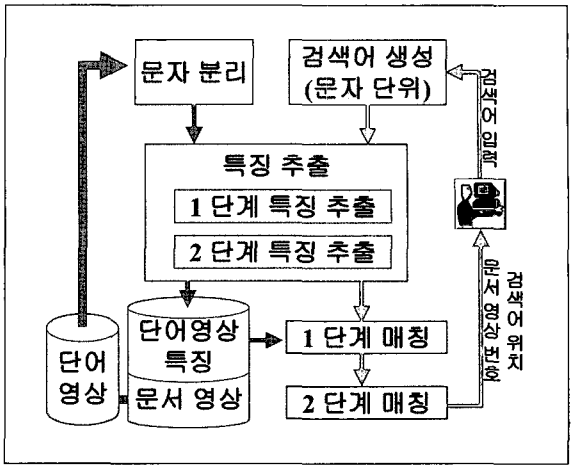
영상 기반 단어 검색 방법으로는 최근 몇 년간 영문 단어와 한자 단어 검색에 대한 연구 결과들이 발표되었다[1, 3-11]. 그러나 한글 단어 검색에 대한 연구는 오일석 등[12, 13]에 의한 연구 결과가 유일하다. 오일석 등은 한글 영상 문서에서 한글 단어를 검색하기 위해 곽희규 등[14]의 시스템을 이용하여 한글 문서에서 단어 영상을 추출하여 단어 영상 데이터베이스를 구성하였다. 검색어는 문서 편집기에서 데이터베이스에서 사용된 동일한 폰트를 생성하여 이를 프린트하고 다시 스캔하여 사용하였다. 단어 영상은 문자 단위로 분할하였다. 처리시간을 최소화하기 위해 두 단계 매칭 방법을 사용하였다. 1단계에서는 프로파일 특징을 이용하였고, 두 번째 단계에서는 Harr 웨이블릿 계수 중 가장 큰 값을 갖는 30개를 선택하여 사용하였다.

본 논문에서는 기존 방법과는 달리 문자 분리가 빠르게 수행될 수 있도록 문자수를 추정하여 문자 폭의 분산이 최소가 되는 분할 방법을 선택하였고, 검색어 영상을 시스템 인터페이스를 이용하여 자동으로 생성하도록 하였다. 문자 영상의 특징 벡터는 오일석 등의 연구에서 사용된 특징 벡터와 격자 방법을 이용하여 우수한 특징 벡터를 선택하였다. 매칭 방법은 여러 문헌에서 널리 사용된 2단계 매칭 방법[1, 13]을 이용하였다.

본 논문의 구성은 다음과 같다. 1장에서는 대규모 문서 영상 데이터베이스에서 단어 검색의 필요성과 기존의 문서 영상에서 단어 검색의 방법인 OCR의 설명과 문제점을 서술하였고 영상기반 단어 검색의 기존 연구를 살펴보았다. 2장에서는 두 단계 매칭을 이용한 단어 검색 시스템을 이루는 문자 분할, 검색어 영상 생성, 특징 추출, 매칭 방법에 대해 서술한다. 3장에서는 실험환경, 성능 및 오류에 대해 서술하며, 4장은 결론과 향후 연구 방향을 제시한다.

1. 영상기반 한글 단어 검색 시스템

(그림 1)은 두 단계 영상 매칭을 이용한 한글 문서 영상에서의 단어 검색 시스템의 블록 다이어그램이다. 단어 영상은 정창부 등의 시스템[15]을 이용하여 문서 영상에서 분리되어 데이터베이스에 미리 저장되었다고 가정한다. 단어 영상은 문자 영상으로 분리되고 두 가지 특징 벡터가 추출되어 문서 영상과 함께 데이터베이스에 저장된다. 사용자가 검색 시스템 인터페이스에서 검색어를 입력하면, 시스템이 지원하는 폰트를 이용하여 검색어 영상을 문자 단위로 생성하고, 검색어 영상에 대한 두 단계 특징 벡터를 추출한다. 첫 번째 매칭에서는 검색어와 데이터베이스 내의 단어와 비교하여 1단계 매칭의 임계값을 만족시키는 단어를 다음 단계의 후보로 선택한다. 2단계 매칭에서는 전 단계에서 선택



(그림 1) 영상기반 한글 단어 검색 시스템의 블록 다이어그램

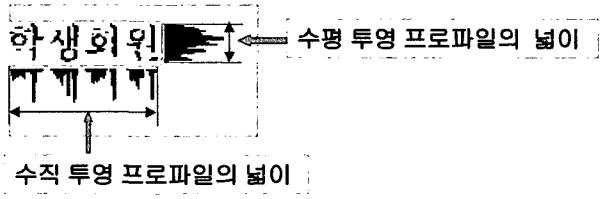
되어진 후보 단어와 검색어의 2단계 특징과 비교하여 해당 임계값을 만족하는 단어만을 선택한다. 양 단계의 임계값을 모두 만족한 단어 영상이 검색어와 동일한 단어 영상이다. 결과는 단어 영상이 포함된 문서 영상의 번호와 해당 문서 내의 단어 영상 위치이다.

2.1 문자 분리

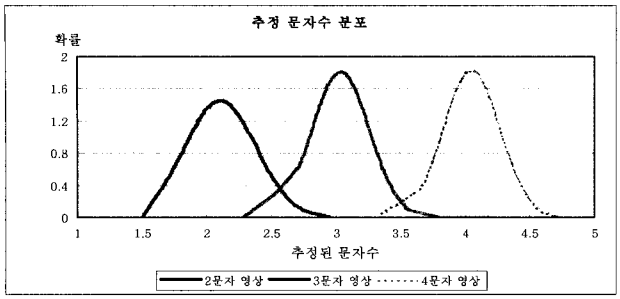
문자 분리 과정은 4단계로 구성된다. 첫 번째 단계에서는 단어 영상이 몇 개의 문자로 구성되었는지 그 문자수를 추정하고, 두 번째 단계에서는 추정된 문자수에 따라 문자 분할 점을 탐색한다. 세 번째 단계는 문자 분할이 애매하여 문자수가 2가지로 추정되었을 경우 분할된 문자 넓이의 분산이 최소가 되는 문자수를 선택한다. 네 번째 단계에서는 문자 분할 오류를 수정하는 후처리가 수행된다.

2.1.1 문자수 추정

한글 문자의 바운드 박스(최소 외곽 사각형)는 일반적으로 정사각형의 형태를 띄고 그 크기가 일정하다. 그러므로 단어 영상의 수평 투영 프로파일의 넓이로 수직 투영 프로파일의 넓이를 나누면 단어영상을 구성하는 문자수를 쉽게 추정할 수 있다. 하지만 이러한 방법으로 추정된 문자수는 절대적으로 신뢰하기는 어렵다. 따라서 추정 문자수가 (그림 3)에서 보는바와 같이 명확하게 결정하기 힘들 경우 양쪽에 걸쳐있는 문자수 2가지 모두를 추정 문자수로 한다. (그림



(그림 2) 단어 영상의 투영 프로파일 분석



(그림 3) 추정 문자수의 분포

3)은 3.2절의 <표 1>로부터 얻어진 실제 문자수에 따른 추정 문자수의 분포를 나타낸다.

2.1.2 분할 점 탐색

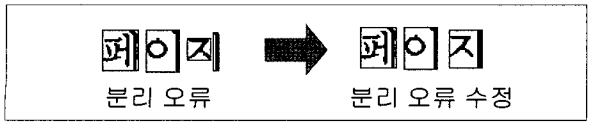
이 과정에서는 (1)에서 언급한 한글 문자의 형태적인 특성과 추정된 문자수를 이용하여 문자 분할 점을 탐색한다. 추정 문자 넓이만큼 이동하여 수직 투영 프로파일의 값이 0인 경우 해당 지점을 분할 점으로 선택하고 그렇지 않을 때는 투영 프로파일의 좌·우로 동시에 이동하면서 먼저 0이 발견된 지점을 분할 점으로 선택한다.

2.1.3 최적의 문자수 선택

추정된 문자수가 두 가지일 경우, (1)에서 언급한 한글 문자의 형태적 특징을 이용하여 최적의 문자수를 선택한다. 문자의 바운드 박스가 일정한 크기의 정사각형의 형태라 것은 추정 문자수에 따라 문자 분리가 올바르게 수행될 경우 문자 넓이의 분산은 그렇지 않은 경우보다 적은 값을 갖는다. 따라서 두 가지 추정 문자수로 문자 분할을 수행한 후, 그들의 문자 넓이의 분산이 더 적은 경우의 문자수를 선택한다.

2.1.4 후처리

이 단계에서는 한글의 형태적 특징과 데이터 분석 정보를 이용하여 문자 분리 오류를 정정한다. 마지막 문자가 "ㅏ", "ㅐ", "ㅣ", 등과 같은 모음을 포함하고 영상의 위와 아래 부분이 절단되어 추정 문자 넓이가 짧아지고, 수직 투영 프로파일에서 하나의 모음이 분리되면 이는 두 개의 문자로 분할될 가능성이 높다. 따라서 마지막으로 분할된 문자의 넓이가 단어 영상에서 가장 큰 문자 넓이의 80% 이하의 크기이면 그 앞의 문자와 결합한다. 만약 앞의 문자의 넓이가 가장 큰 문자 넓이의 80%이상 크기이면 마지막 분할 영역

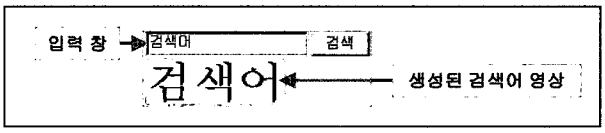


(그림 4) 후처리에 의한 문자 분리 오류 수정

을 잡음으로 간주하고 결합하지 않는다. (그림 4)는 끝 문자의 모음 때문에 발생하는 문자 분할 오류를 위 방법으로 수정한 예이다.

2.2 검색어 영상 생성

기존의 영상기반 단어 검색방법에서는 검색어의 특징을 추출하기 위해 문자 클래스 별로 특징 벡터를 훈련해 둔 후 사용자가 검색어를 입력하면 해당 문자의 특징들을 조합하여 검색어의 특징을 구성한다. 이러한 방식은 한글과 같이 많은 문자가 존재할 경우 데이터베이스를 생성하는 과정에서 많은 시간이 필요하다. 따라서 본 논문에서는 이러한 번거로운 문제를 해결하고 다양한 문자와 다양한 언어의 단어 영상을 쉽게 획득하기 위해 사용자가 검색어를 입력하면 시스템에서 제공하는 폰트를 사용하여 문자 영상을 직접 생성한다. 생성된 문자 영상은 32×32크기로 정규화되고 특징 벡터가 추출된다. (그림 5)는 본 논문에서 구현한 시스템의 검색어 입력 창이다.



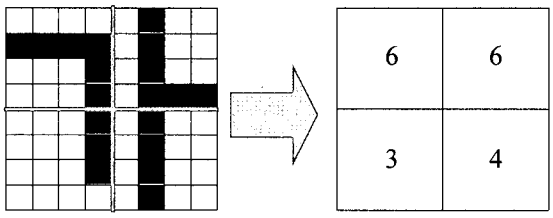
(그림 5) 시스템 인터페이스를 이용한 검색어 영상 생성

2.3 특징 추출

정규화된 문자 영상에서 격자(Mesh)와 프로파일[12], 웨이블릿(Wavelet)[13, 16]의 특징 추출 방법으로 해당 문자를 대표할 수 있는 특징 벡터를 추출한다.

2.3.1 격자 특징

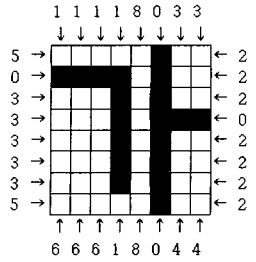
격자 특징 추출은 문자 영상을 N×M개의 사각형으로 구분하여, 각 셀에 포함된 흑화소의 개수를 셀의 특징 벡터로 이용한다. (그림 6)은 8×8로 정규화된 한글 문자 영상에서 2×2 격자를 이용하여 4차원 특징을 추출하는 모습을 나타낸다.



(그림 6) 격자 특징

2.3.2 프로파일 특징

프로파일 특징은 영상의 4방향(left, top, right, bottom)에서 얻어진다. 정규화된 문자 영상의 각 면을 따라가면서 반



$\overline{F_4}$=(3, 2, 2, 4)

$\overline{F_8}$=(3, 3, 1, 3, 1, 2, 5, 4)

(그림 7) 프로파일 특징

대 면까지 흰화소를 세어가면서 흰화소 런(white run)을 구한다. 흑화소를 만나거나 반대 면에 도달할 때까지 흰화소 런을 추적한다[13]. 한 면에서 구해진 흰화소 런의 길이를 평균한 값이 프로파일 특징의 한 구성요소이다. (그림 7)은 8×8로 정규화된 한글 문자 영상에서 4차원 프로파일 특징($\overline{F_4}$)을 추출하는 모습을 나타낸다. 8차원 특징($\overline{F_8}$)을 추출한다면 각 면을 둘로 나누어 4개의 흰화소 런의 길이를 평균한 값이 특징의 한 구성요소이다.

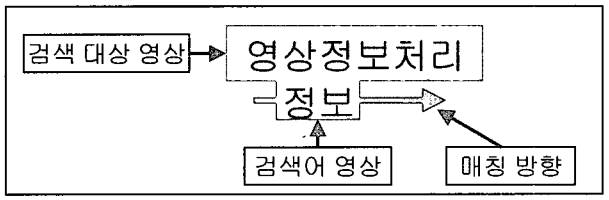
2.3.3 웨이블릿 특징

32×32의 정규화된 문자 영상을 Harr 웨이블릿을 이용하여 5단계 연속 변환한다. 그로부터 1개의 저주파 계수와 1023개의 고주파 계수를 얻을 수 있는데 계수의 절대값이 클수록 문자 영상의 정보를 많이 담고 있다. 따라서 값이 큰 몇 개의 계수를 문자 영상을 대표하는 특징 벡터로 사용하는데, 계수의 크기가 큰 순서대로 해당하는 위치와 값을 특징 벡터로 저장한다. 검색어 영상과 데이터베이스의 영상을 비교할 때, N차원의 특징 벡터를 사용한다면, N개의 특징 벡터와 검색대상 영상의 같은 위치에 있는 웨이블릿 계수를 비교한다.

2.4 매칭 방법

(그림 8)은 검색 대상 영상에서 검색어 영상을 매칭하는 모습을 보여준다. 두 영상의 매칭은 검색어 영상의 문자와 검색 대상 영상의 문자를 순차적으로 비교하여 연속적으로 매칭 규칙을 만족시키면 매칭이 성공한다고 본다.

검색어 영상이 k개의 문자로 구성되었다면 검색 대상 영상에서 매칭에 참여하는 k개의 문자 영상을 목적 영상이라고 하자. 그리고 목적 영상을 $D(C_1^d, C_2^d, \ldots, C_k^d)$, 검색어 영상을 $Q(C_1^q, C_2^q, \ldots, C_k^q)$라고 표기하자. 여기서 C_i는 단어 영상의 i번째 문자에서 추출한 m차원 특징 벡터이다. 이는 $C_i=(c_{i,1}, c_{i,2}, \ldots, c_{i,m})$로 표기한다. 목적 영상과 검색어 영상의 i번째 문자의 유사성 판단은 식 (1)에 근거하는데 T_c는 임계값이다.



(그림 8) 영상 매칭 과정

$$\left(Dist(C_i^q, C_i^d) = \sum_{j=1}^{m} |c_{i,j}^q - c_{i,j}^d|\right) < T_c \qquad (1)$$

k개의 문자가 연속적으로 같을 경우 식 (2)에 근거하여 두 단어의 유사성을 판단한다. T_w는 임계값이다.

$$\left(Dist(D,Q) = \frac{1}{k}\sum_{i=1}^{k} Dist(C_i^q, C_i^d)\right) < T_w \qquad (2)$$

매칭 단계에서 문자와 단어 매칭을 순차적으로 수행하기 때문에 고차원의 특징 벡터를 사용하게 되면 검색 속도가 현저하게 떨어진다. 그러나 속도를 고려하여 저차원의 특징 벡터를 사용하면 재현율과 정확율이 낮아진다. 따라서 본 논문에서는 두 단계 매칭 방법을 도입하여 1단계 매칭은 속도와 재현율을 높일 수 있도록 저차원 특징 벡터를 이용하고, 2단계 매칭은 1단계 매칭보다는 느리지만 재현율과 정확율을 높일 수 있도록 고차원의 특징 벡터를 적용한다.

3. 실험 결과

3.1 실험 환경

실험 데이터는 한국정보과학회[17]의 학술데이터베이스에서 다운로드한 학술 문서 영상 8페이지 분량이 사용되었다. 이 문서는 대부분 10포인트와 8포인트 크기의 신명조체 문자로 이루어져 있으며, 300 DPI로 스캔된 영상이다. 정창부 등[15]의 시스템을 이용하여 문서 영상에서 단어 영상을 분리하였고, 분리된 영상에서 영문, 특수 문자 그리고 숫자가 포함된 영상을 제외한 한글 단어 영상 1600개를 실험에 사용하였다. 단어 영상은 문자 단위로 분리되고 32×32의 크기로 정규화된다. 정규화된 문자 영상에서 2.3에서 서술한 특징 추출 방법으로 4, 8, 16 그리고 30차원의 특징을 각각 추출하였다. 다만 프로파일 특징의 경우 특징 추출의 특성 때문에 30차원에 가장 근사한 32차원 특징을 사용한다. 시스템 인터페이스를 이용하여 2문자 단어 영상 15개, 3문자 단어 영상 10개 그리고 4문자 단어 영상 5개, 총 30개의 검색어 영상을 생성하였다. 1600개 실험 영상에서 30개 검색 영상의 출현 횟수는 621회이다. 실험에 사용된 기자재는 Pentium-4 CPU 2.80GHz와 1GB RAM 자원을 갖는 개인용 PC이다.

3.2 성능 평가

3.2.1 문자 분리

<표 1>은 실험에 사용된 1600개 단어 영상을 투영프로파일 정보에 근거하여 문자수를 추정하고 이을 분석한 자료이다. 실제 문자수가 1인 경우는 2개의 영상이고, 11인 경우는 1개로써 <표 1>에 수록하지 않았다.

본 논문에서 제안한 문자 분리 방법은 1600개 단어 영상에서 1577개를 정확하게 분리하여 98.56%의 문자 분리 성공률을 보였다. <표 2>에 나타난 영상은 문자 분리 오류의 예이다. 문자 분리 오류는 단어 영상에 노이즈가 포함되거나 영상 정보의 손실로부터 기인한다. 예를 들어 <표 2>의 1번 영상처럼 단어 영상의 위쪽이나 아래쪽에서 정보의 손실이 있어 과다 분리되거나, 2번 영상처럼 문자에 노이즈가 추가되어 두 문자가 연결된 경우이다.

〈표 1〉 추정 문자 수의 분석

실제 문자수	실험 단어 영상수	추정 문자수 최소 값	추정 문자수 최대 값	추정 문자수 평균	추정 문자수 표준편차
2	442	1.53	2.91	2.10	0.27
3	540	2.25	4.27	3.04	0.22
4	332	3.34	4.7	4.05	0.27
5	190	3.79	6.33	5.10	0.26
6	70	3.11	7.44	5.98	0.47
7	16	6.41	8.47	7.07	0.45
8	7	7.79	8.06	7.93	0.12

〈표 2〉 문자 분리 오류

번호	입력 단어 영상	문자 분리 오류
1		
2		

3.2.2 단일 및 2단계 특징 성능 분석

각 매칭 단계의 특징 벡터를 선택하기 위해 다양한 특징 벡터를 추출하여 각각의 성능을 분석하였다. <표 3>은 다양한 차원의 격자, 프로파일 그리고 웨이블릿 특징의 성능을 분석한 결과이다. 처리 시간이 상대적으로 빠르고 재현율이 높은 4, 8 그리고 16 차원의 프로파일 특징이 다른 특징 추출 방법보다 성능이 우수하여 이를 1단계 매칭의 특징 후보로 결정한다. 30차원 격자 특징 추출 방법이 다른 방법보다 같은 재현율과 정확율에 비해 처리속도가 우수하여 2단계 매칭에 이용하기로 한다.

<표 4>는 3가지 저차원 프로파일 특징을 1단계 매칭의 특징으로 하고 30차원 격자 특징을 2단계 매칭 특징으로 결합하여 성능을 분석한 자료이다. 세 가지 조합에 의한 시스

〈표 3〉 격자, 프로파일 그리고 웨이블릿의 특징 차원별 성능 분석

		재현율(%)	정확율(%)	초당 처리 단어 수 (words/sec)
4D	격자	97.10	2.75	600,000
	프로파일	97.26	7.22	600,000
	웨이블릿	97.10	2.90	244,800
8D	격자	97.10	5.70	545,454
	프로파일	97.10	7.97	533,333
	웨이블릿	97.10	3.18	200,470
16D	격자	97.10	5.09	462,720
	프로파일	97.10	12.94	450,000
	웨이블릿	97.10	3.52	189,333
30D	격자	89.69	89.69	377,739
	프로파일 32D	88.73	88.73	350,000
	웨이블릿	79.55	78.29	140,903

〈표 4〉 저차원 프로파일 특징과 고차원 격자 특징 결합의 성능 분석

	재현율 (%)	정확율 (%)	초당 처리 단어수 (words/sec)
4D 프로파일 + 30D 격자	89.69	89.84	519,951
8D 프로파일 + 30D 격자	89.86	89.71	467,673
16D 프로파일 + 30D 격자	90.02	90.16	447,604

템의 성능은 재현율과 정확율에 있어 서로 유사하지만, 4차원 프로파일 특징과 30차원 격자 특징의 조합이 처리시간에 있어 매우 우수하므로, 이들을 각 단계의 특징으로 최종 결정한다. 이 결합은 기존의 4차원 프로파일 특징과 30차원 웨이블릿 특징을 결합한 방식의 재현율 82.77%, 정확율 82.24% 그리고 초당 처리 단어 186,327개의 성능보다 우수하다.

3.2.3 검색어 각각의 성능 분석

<표 5>는 30개의 검색어를 시스템에 입력하여 문자의 개수별로 성능을 분석한 결과이다. 2 문자 검색어는 자신 보다 긴 목적 단어와 매칭을 많이 하기 때문에 3, 4문자의 검색어보다 다른 영상과 매칭할 기회가 많아지고 오판할 가능성이 높아 검색 성능이 상대적으로 낮다.

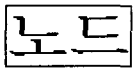
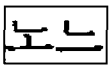
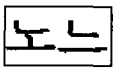
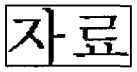
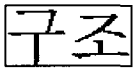
〈표 5〉 검색어 각각의 성능 분석

검색어	출현 빈도	1 단계				2 단계			
		검색 결과	정답 개수	재현율 (%)	정확율 (%)	검색 결과	정답 개수	재현율 (%)	정확율 (%)
공유	48	433	47	97.92	10.85	57	46	95.83	80.70
구현	18	115	18	100.00	15.65	17	17	94.44	100.00
성능	31	407	30	96.77	7.37	30	30	96.77	100.00
향상	11	788	10	90.91	1.27	12	10	90.91	83.33
노드	72	85	67	93.06	78.82	66	65	90.28	98.48
응용	25	491	24	96.00	4.89	60	22	88.00	36.67
통신	24	399	23	95.83	5.76	12	12	50.00	100.00
논문	24	173	23	95.83	13.29	23	23	95.83	100.00
환경	7	662	7	100.00	1.06	2	2	28.57	100.00
구조	17	58	17	100.00	29.31	15	15	88.24	100.00
벡터	12	499	12	100.00	2.40	9	9	75.00	100.00
자료	11	634	11	100.00	1.74	7	7	63.64	100.00
최대	9	492	9	100.00	1.83	10	7	77.78	70.00
할당	4	644	4	100.00	0.62	4	4	100.00	100.00
방법	5	651	5	100.00	0.77	5	5	100.00	100.00
2문자 검색어 성능	318	6,531	307	96.54	4.70	329	274	86.16	83.28
메모리	53	207	51	96.23	24.64	46	46	86.79	100.00
시스템	16	69	16	100.00	23.19	16	16	100.00	100.00
페이지	74	155	74	100.00	47.74	77	72	97.30	93.51
데이타	17	175	17	100.00	9.71	15	15	88.24	100.00
동기화	9	107	8	88.89	7.48	8	8	88.98	100.00
일관성	9	341	9	100.00	2.64	8	8	88.89	100.00
원거리	13	315	13	100.00	4.13	13	13	100.00	100.00
무효화	4	61	4	100.00	6.56	4	4	100.00	100.00
상업용	3	194	3	100.00	1.55	6	3	100.00	50.00
국부성	4	64	4	100.00	6.25	4	4	100.00	100.00
3문자 검색어 성능	202	1,688	199	98.51	11.79	197	189	93.56	95.94
클러스터	24	32	24	100.00	75.00	22	22	91.67	100.00
네트워크	7	10	7	100.00	70.00	7	7	100.00	100.00
디렉토리	3	19	3	100.00	15.79	2	2	66.67	100.00
프로세스	63	61	60	95.24	98.36	60	60	95.24	100.00
운영체제	4	30	4	100.00	13.33	3	3	75.00	100.00
4문자 검색어 성능	101	152	98	97.03	64.47	94	94	93.37	100.00
모든 검색어 성능	621	8,371	604	97.26	7.22	620	557	89.39	89.84

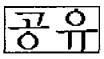
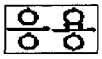
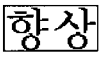
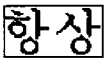
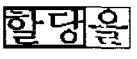
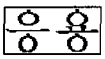
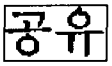
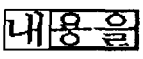
3.3.4 검색 단계 오류 분석

검색 단계에서 나타나는 오류는 두 가지로 나누어 볼 수 있다. 하나는 같은 영상인데 다르다고 판단한 경우인데, 실험에서 64개의 에러가 출현하여 89.69%의 재현율을 보였다. 또 다른 하나는 다른 영상인데 같다고 검색한 경우인데, 본 실험에서 63개가 나타났다. 동일 영상을 검색하지 못한 경우는 폰트가 다른 경우와 영상의 정보가 많이 훼손된 경우이며, 다른 영상을 같다고 한 경우는 제안된 특징 벡터가 두 단어 영상을 분류할 수 없기 때문이다. <표 6>과 <표 7>은 검색 단계에서 나타나는 두 가지 오류 각각의 예를 보여준다.

〈표 6〉 동일 단어 영상인데 검색하지 못한 경우

검색어	동일 단어 영상인데 검색하지 못한 경우	
		
		
		

〈표 7〉 다른 영상을 동일 단어 영상으로 오인식

검색어	다른 영상을 동일 단어 영상으로 오인식	
		
		
		

4. 결론 및 향후 연구 방향

본 논문에서는 두 단계 이미지 매칭을 이용하여 한글 문서영상에서 사용자 검색어를 빠르고 정확하게 검색할 수 있는 시스템을 제안하였다. 이를 위해 문서 영상에서 추출된 단어 영상을 문자 단위로 분리하는 알고리즘을 제안하여

98.56%의 높은 분리 성공률을 획득하였다. 격자와 프로파일 그리고 웨이블릿을 이용하여 4, 8, 16 그리고 30차원의 특징 벡터를 추출하여 저차원 특징 벡터 중에서 처리 시간이 빠르고 재현율을 높일 수 있는 4차원 프로파일 특징 벡터를 1단계 매칭에 이용하고 재현율과 정확율을 높일 수 있는 30차원 격자특징 벡터를 2단계 매칭에 이용하였다. 그 결과 89.69% 재현율과 89.84%의 정확율을 보였고, 초당 처리 단어는 519,951개를 나타냈다. 이 결과는 기존에 제안된 영상-기반 한글 검색 시스템보다 재현율과 정확율이 더 높고, 초당 처리할 수 있는 단어량도 더 많다. 제안된 시스템은 전자 도서관 구축에 있어 과도한 수작업이 요구되는 OCR의 최대 문제점을 해결하는 대안이 된다.

향후 연구는 노이즈가 있거나 정보 손실이 있더라도 올바르게 문자를 분할할 수 있는 새로운 한글 문자 분할 알고리즘 연구가 필요하며, 검색 대상이 되는 영상의 폰트 정보를 추출하여 검색어 생성에 적용하고, 재현율과 정확율을 동시에 향상시키는 새로운 특징 추출 방법의 연구가 필요하다. 또한 문자를 분할하지 않고 단어 영상만을 이용한 검색의 연구가 필요하다.

참 고 문 헌

[1] Y. Lu and C.L. Tan, "Chinese word searching in imaged documents," *International Journal of Pattern Recognition and Artificial Intelligence*, Vol.18, No.2, pp.229-246, 2004.

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

[3] 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.

[4] Y. Lu, L. Zhang and C.L. Tan, "Retrieving Imaged Documents in Digital Libraries Based on Word Image Coding," *International Workshop on Document Image Analysis for Libraries*, USA, pp.174-187, 2004.

[5] 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.

[6] J. DeCurtins and E. Chen, "Keyword spotting via word shape recognition," *Proc. SPIE Document Recognition II*, pp.270-277, 1995.

[7] A. Kolcz, J. Alspector, M. Augusteijn, R. Carlson and GV Popescu, "A line-oriented approach to word spotting in handwritten documents," *Pattern Analysis and Applications*, Vol.3, No.2, pp.153-168, 2000.

[8] R. Manmatha, Chengfeng Han, and E. M. Riseman, "Word spotting: A new approach to indexing handwriting," *Proc. Computer Vision and Pattern Recognition Conference*, pp.631-637, 1996.

[9] T. Syeda-Mahmood, "Indexing of handwritten document images," *Proc. Workshop on Document Image Analysis*, Puerto Rico, pp.66-73, 1997.

[10] F.R. Chen, L.D. Wilcox, 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.

[11] F.R. Chen, L.D. Wilcox and D.S. Bloomberg, "Detecting and locating partially specified keywords in scanned images using hidden Markov models," *Proc. Document Analysis and Recognition*, pp.133- 138, Oct., 1993.

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

[13] 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.

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

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

[16] C.E. Jacobs, A. Finkelstein, and D.H. Salesin, "Fast multiresolution image querying," Proc. *22nd annual conference on Computer graphics and interactive techniques*, pp.277-286, Sep., 1995.

[17] http://www.kiss.or.kr/

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

손 화 정

e-mail : sonhj@iip.chonnam.ac.kr
2001년 전남대학교 통계학과(학사)
2004년 전남대학교 전산학과(이학석사)
2004년~현재 전남대학교 전산학과 박사 과정
관심분야 : 패턴인식, 문자인식, 영상처리

김 수 형

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