한글 문자 데이터베이스 PHD08 구축

Construction of Printed Hangul Character Database PHD08

함대성, 이득용, 정인숙, 오일석
전북대학교 전자정보공학부 컴퓨터공학

Dae-Sung Ham(ham890@chonbuk.ac.kr), Duk-Ryong Lee(dr_lee@chonbuk.ac.kr),
InSuk Jung(msisjung@honbuk.ac.kr), Il-Seok Oh(isoh@chonbuk.ac.kr)

요약

문자 인식의 응용이 형식 문서의 인식 같은 고전적인 영역을 벗어나 웹 문서나 자연 영상의 문자 인식으로 확장되고 있다. 이러한 새로운 응용에서는 명조나 고딕같은 표준 글꼴뿐만 아니라 다양한 모양의 글꼴을 사용하는 것이 보편적이다. 기존의 데이터베이스들은 주로 표준 글꼴을 대상으로 제작되어 새로운 응용에 사용하는데 한계를 안고 있다. 본 논문에서는 완성형 2350자 각각을 대상으로 9종류의 글꼴에 대해 글꼴 크기, 품질, 해상도를 달리하여 243개의 이미지 샘플을 생성하였다. 또한 이들 샘플 각각에 대해 이진 임계치와 회전 각도를 달리하여 변형된 샘플을 얻었다. 이러한 과정으로 각 글자마다 2,187개의 샘플을 생성하였으며, 총 5,139,450개의 샘플을 갖는 인쇄체 한글 데이터베이스를 구축하였다. 데이터베이스에 대한 특성과 상용 OCR 소프트웨어에 대한 인식 성능 등을 제시한다.

■ 중심어 : | 문자인식 | 한글인식 | 데이터베이스 | 인쇄체 |

Abstract

The application of OCR moves from traditional formatted documents to the web document and natural scene images. It is usual that the new applications use not only standard fonts of Myungjo and Godic but also various fonts. The conventional databases which have mainly been constructed with standard fonts have limitations in applying to the new applications. In this paper, we generate 243 image samples for each of 2350 Hangul character classes which differs in font size, quality, and resolution. Additionally each sample was varied according to binarization threshold and rotational transformation. Through this process 2187 samples were generated for each character class. Totally 5,139,450 samples constitutes the printed Hangul character database called the PHD08. In addition, we present the characteristics and recognition performance by an commercial OCR software.

■ keyword : | Character Recognition | Hangul Recognition | Database | Printed Hangu |

I 서 론

지금은 문자 인식의 응용이 형식 문서의 인식 같은 고전적인 영역을 벗어나 웹 문서나 자연 영상의 문자 인식으로 확장되고 있다 [1-3]. 이러한 응용은 스팸메일 필터링, 문서의 저작권 보호, 문서영상 자동검색 등을 포함한다. 예를 들어 메일의 ASCII 부분은 스트링

접수번호 : #080730-002
접수일자 : 2008년 07월 30일
심사완료일 : 2008년 08월 29일
교신저자 : 함대성, e-mail : ham890@chonbuk.ac.kr

매칭을 통해 스팸 메일을 판단할 수 있지만 문서영상 형태의 메일은 문자인식 기능이 없다면 스팸을 판단할 수 없다. 이러한 새로운 응용에서는 명조나 고딕같은 표준 글꼴뿐만 아니라 다양한 모양의 글꼴을 사용하는 것이 보편적이다. [그림 1]은 몇 가지 웹 문서를 보여주는데, 사용된 글꼴이 다양함을 확인할 수 있다.

(a) 광고 이미지 (b) 제품 이미지

그림 1. 웹 문서

현재 국내에서 사용되고 있는 인쇄체 한글 인식기는 거의 표준 글꼴을 가진 글자 샘플을 가지고 훈련한 것들이다. 따라서 표준 글꼴의 글자에 대해서는 어느 정도 만족할 만한 성능을 제공하지만, [그림 1]과 같은 다양한 글꼴에 대해서는 쓸모없을 정도의 성능을 보이는 것이 현실이다.

아무리 좋은 특징 추출 알고리즘과 분류 알고리즘을 사용하더라도 그것의 학습에 사용되는 훈련 데이터베이스의 품질이 낮으면 원천적으로 높은 성능을 기대할 수 없다[4]. 데이터베이스는 패턴 원천 (patterm source)을 충분히 대표할 정도로 다양한 샘플을 포함하여 질적 품질을 만족해야 한다. 또한 양적으로도 충분해야 한다[5][6]. 그동안 인쇄 한글을 위한 데이터베이스가 여럿 만들어졌지만 새로운 패턴 원천, 즉 웹 문서나 자연 영상과 같은 응용을 대표할 정도의 품질은 갖추지 못한 것이 사실이다. 다시 말하면 기존 데이터베이스는 새로운 응용에 대처하기에는 매우 낡았다고 할 수 있다.

이 논문은 이러한 새로운 응용에 활용할 수 있는 인쇄 한글 데이터베이스 구축을 목표로 한다. 이 논문이 제시하는 'PHD08' 데이터베이스는 완성형 2,350자를 대상으로 9종류의 글꼴에 포함한다. 각 글꼴에 대해 글자 크기, 인쇄 품질과 해상도를 달리하여 243개의 샘플을 생성하였다. 따라서 글자 부류별로 243*9=2,187개의 샘플을 수집하였다.

PHD08은 부류별로 2,187 샘플을 가지고 있어 기존 데이터베이스에 비해 양적으로 우수하다. 또한 글자 크기, 인쇄 품질, 그리고 해상도를 달리하여 생성하였기 때문에 질적인 면에서도 뛰어나다고 볼 수 있다. 이 논문은 PHD08을 어떻게 만들었는지 그 과정을 상세하게 소개한다. 데이터베이스의 특성의 하나로서 기존 상용 인식기로 인식을 시도하였고 그 결과로 얻은 성능도 제시한다.

이 데이터베이스는 웹에 공개하고 있으므로 누구나 쉽게 다운 받아 사용할 수 있다. 이 연구 결과는 웹영상 및 문서영상 등의 문자를 인식하기 위한 훈련 데이터로 사용될 수 있다. 인식에 있어서 실험데이터의 다양성은 매우 중요하다. 지금까지의 인쇄체 한글 인식에 대한 연구들은 대부분 실험과 성능 측정을 위한 데이터베이스를 별도로 제작하였다. 데이터베이스를 제작하는데 많은 시간과 노력이 필요하므로 효율적이지 못하였다. 이러한 측면에서 다양한 데이터베이스가 공유되어야 한다.

II. 기존 한글 데이터베이스

1장에서 언급했듯이 고성능 인식기를 만들기 위해서는 훈련 데이터베이스의 품질이 매우 중요하다. 이런 의미에서 인쇄 한글 인식 문제를 다룬 논문들이 어떤 데이터베이스를 사용했는지 살펴보는 작업이 중요하다. [표 1]은 1990-2006년에 발표된 몇 가지 인쇄 한글 인식 논문들에 대해 그들이 사용한 데이터베이스를 분석하고 있다. 특히 데이터베이스가 다룬 글꼴을 상세하게 분석한다. [그림 2]는 이들 논문에 국한하여 그들이 사용한 글꼴의 빈도수를 보여 주고 있다.

표 1. 기존 시스템들이 사용한 훈련 데이터베이스

구 분	데이터베이스 글꼴
조성배92[7]	명조
이판호94[8]	세명조, 중고딕, 중명조
이진수96[9]	명조, 고딕, 신명조, 중고딕, 궁서
정지호01[10]	바탕
임길택03[11]	우편봉투 인쇄체 문자
이성훈06[12]	비디오 뉴스 자막, 저해상도 스캔 영상
김병기06[13]	고딕, 굴림, 궁서, 명조, 신명조, 바탕, 중고딕

현재 공개되어있는 한글 인식을 위한 데이터베이스는 [표 2]와 같다. 이러한 데이터베이스들은 인식 목적에 따라 필기체와 인쇄체로 구분되어 있다. 필기체 데이터베이스인 KU-1은 1,000명 이상을 대상으로 수집되었으며[14], SERI95a는 무제약 또는 제약으로 필기체 데이터를 수집하였다[15]. PE92데이터는 554명의 필기자를 대상으로 수집되었으며[16], Kaist DB는 우편주소, 영수증, 보고서등의 환경에서 무제약으로 수집하였으며 문자당 벌수가 다르다[15].

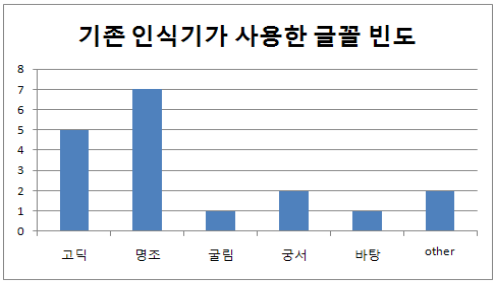


그림 2. 기존 인식기가 사용한 글꼴 빈도

인쇄체 데이터베이스인 ETRI DB는 문서작성 소프트웨어에서 사용되는 바탕체, 돋움체, 궁서체, 굴림체, 명조체, 고딕체의 6종류를 위주로 제작되었다[15]. 이 데이터베이스는 부류별로 100 샘플만을 가지므로 양적으로 충분하다 보기가 어렵다. CBNU한글00은 고딕과 명조의 2종류로 제작되었다[17]. 부류별로 1000개의 샘플을 가지므로 양적으로는 ETRI 데이터베이스를 능가하나 글꼴이 두 종류라는 한계를 안고 있다.

[표 2]의 마지막 줄에 제시한 PHD08은 양적으로나 질적으로 이 두 데이터베이스를 능가한다. 3장부터는 PHD08에 대해 생성 과정, 인식 난이도 평가 결과 등을 상세하게 제시한다.

Ⅲ. PHD08 생성

PHD한글08 데이터베이스는 인공적으로 생성하였다. 많은 다양한 문서 영상을 충분히 구할 수 있다면 실제 문서에서 수집하는 것이 더 높은 품질을 보장하겠지만 그러한 가정은 현실적으로 매우 어렵다.

표 3. PHD08이 사용한 글꼴과 모양

글꼴	모양	악어
바다	한글 데이터베이스	B
돋움	한글 데이터베이스	D
고딕	한글 데이터베이스	G
한양해서	한글 데이터베이스	H1
헤드라인	한글 데이터베이스	H2
명조	한글 데이터베이스	M
나무	한글 데이터베이스	N
샘물	한글 데이터베이스	S
엽서	한글 데이터베이스	Y

3.1 샘플 수집 과정

표 2. 한글 인식을 위한 데이터베이스

	구분	문자수	벌수	크기(wi_he)	저장형태	글꼴
KU-1	필기	1,500	1,000	-	Gray	-
SERI95a	필기	520	1,000	100_100	Gray	-
PE92	필기	2,350	100	100_100	Gray	-
Kaist DB	필기	422	-	-	Binary	-
ETRI DB	인쇄	2,350	100	100_100	Binary	6종류
CBNU한글00	인쇄	2,350	1,200	32_32	Binary	2종류
PHD08	인쇄	2,350	2,187	-	Binary	9종류

실제 상황에서 발생하는 다양한 변형을 시뮬레이션하기 위해 여섯 가지 측면에서 변화를 시도하였다.

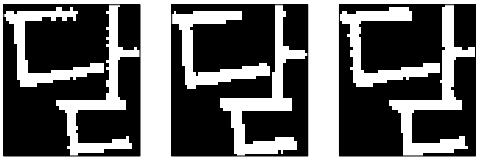
1. 글꼴 9가지: [표 3]에서 보는 바와 같이 대표적인 9가지 글꼴에 대해 샘플을 생성한다.
2. 글자 크기 3가지: 종이에 인쇄할 때 12, 13, 14의 세 가지 글꼴 크기로 인쇄하였다.
3. 노이즈 레벨 3가지: 노이즈 레벨을 세 가지로 시뮬레이션하기 위해 인쇄 원본, 한번 복사, 재 복사의 세 가지를 시도하였다.
4. 스캔 해상도 3가지: 스캔을 200, 240, 280의 세 가지로 하여 명암 영상을 얻었다.
5. 이진화 3가지: 명암 영상을 이진화하는 단계에서 임계값을 140, 180, 220으로 세가지를 사용하였다.
6. 회전 3가지: 이진 영상 각각에 대해 -3^{o}, 0^{o}, 3^{o}의 세 가지 회전을 가하였다.

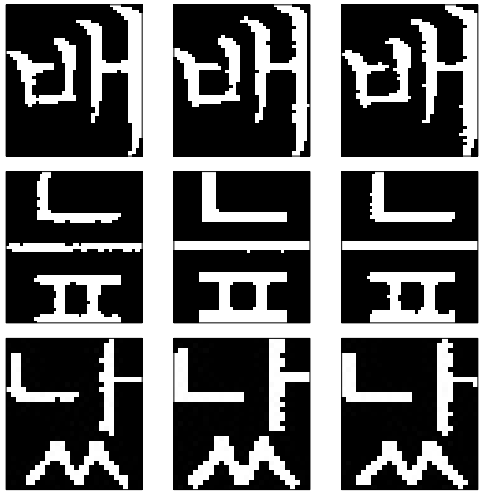
이런 과정을 거쳐 글자 부류별로 $9*3^{5}$=2187개의 샘플을 수집하였다. 글자 부류가 2,350개이므로 총 5,139,450개의 샘플을 갖는 데이터베이스가 완성되었다.

프린터는 HP LaserJet 2430, 스캐너는 Epson expression 1680을 사용하였다.

[그림 3]은 몇 가지 부류에 대해 노이즈 레벨에 따른 변형을 보여준다. 복사의 단계를 거치며, 일정하지 않은 노이즈가 첨가됨을 볼 수 있다. [그림 4]는 스캐닝 단계에서의 해상도에 따른 변형을 보여 준다. 높은 해상도에 따라 글자가 좀 더 세밀하게 표현됨을 볼 수 있다.

[그림 5]는 명암 영상을 이진화하는 단계에서 임계값을 달리하였을 때 나타나는 변형을 보여 준다. 이진 임계치가 높을수록 글자의 윤곽이 얇고 분명해지며, 낮을수록 글자 획의 사이가 가까워지는 것을 관찰할 수 있다. [그림 6]은 회전을 가하였을 때 나타나는 변형을 보여 준다. -3^{o}, 0^{o}, 3^{o}의 회전값에 따라 글자의 변형을 관찰할 수 있다.





(a) 원본 (b) 한번 복사 (c) 두번 복사

그림 3. 복사에 따른 변형

(a) 200 dpi (b) 240 dpi (c) 280 dpi

그림 4. 해상도에 따른 변형

(a) t = 140 (b) t = 180 (c) t = 220

그림 5. 이진 임계치에 따른 변형

(a) r = −3° (b) r = 0° (c) r = 3°

그림 6. 회전에 따른 변형

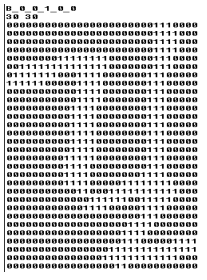
3.2 샘플 색인

3.1절에서 수집한 샘플 각각은 아래 인코딩 규칙에 따라 비트맵 상단에 두 줄로 생성 과정을 기록한다.

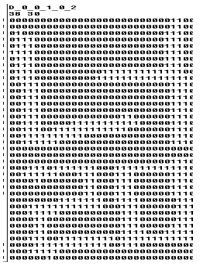
FO_FS_CP_RE_TH_SL
HE_WD

- FO: 글꼴의 종류 (표 3의 약어로 코딩)
- FS: 글꼴의 크기 12, 13, 14 (0, 1, 2로 코딩)
- CP: 복사 횟수 (0, 1, 2로 코딩)
- RE: 해상도 200, 240, 280 (0, 1, 2로 코딩)
- TH: 이진 임계치 140, 180, 220 (0, 1, 2로 코딩)
- SL: 기울기 -3°, 0°, 3° (0, 1, 2로 코딩)
- HE: 비트맵의 세로 화소 수
- WD: 비트맵의 가로 화소 수

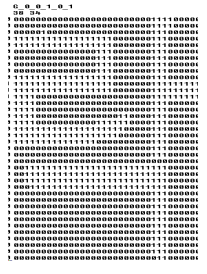
예를 들어 B_2_0_1_1_2와 38_32는 바다체, 글자 크기 14, 복사 횟수 0번, 해상도 240, 이진 임계치 180, 기울기 3°이면 비트맵은 38*32임을 나타낸다. [그림 7]은 9가지 글꼴 각각에 대해 데이터베이스에 들어있는 샘플 하나를 예시하고 있다.



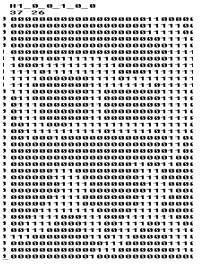
(a) 바다



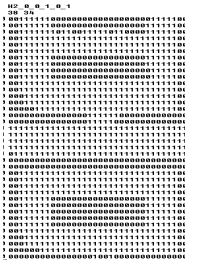
(b) 돋움



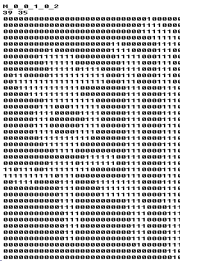
(c) 고딕



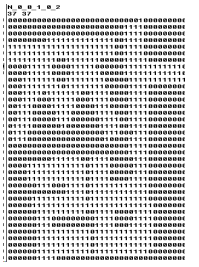
(d) 한양해서



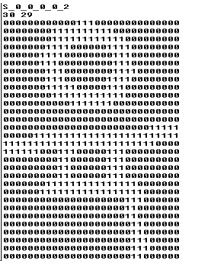
(e) 헤드라인



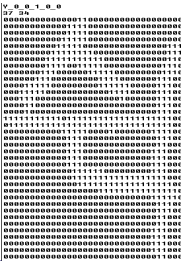
(f) 명조



(g) 나무



(h) 샘물



(i) 엽서

그림 7. 샘플 예제

부류 정보는 따로 기록하지 않고 파일 이름으로 구분하고 있다. 예를 들어 부류 '가'는 '가.txt'라는 파일에 기록되어 있다. 즉 가.txt 파일은 부류 가에 해당하는 샘플 전체, 즉 2,187개의 샘플을 포함하고 있다. 따라서 총 2,350 개의 파일이 존재한다. 파일 크기는 대략 3~4M바이트이고, 데이터베이스의 총 크기는 약 7G바이트이다.

파일 크기를 줄이기 위해 압축 버전도 제공한다. 압축은 간단한 런 길이 방식을 사용한다. [그림 8]에서 보여주는 바와 같이 압축은 줄 단위로 한다. 압축 버전은 총 2G바이트를 차지한다.

PHD08 데이터베이스는 웹사이트[17]에서 무료로 제공하고 있다.

00000011000 --> 6 2 3
11000111111 --> 0 2 3 6
10000011100 --> 0 1 5 3 2
00000111000 --> 5 3 3

그림 8. 런 길이 압축

Ⅳ. 상용 인식기에 의한 인식률

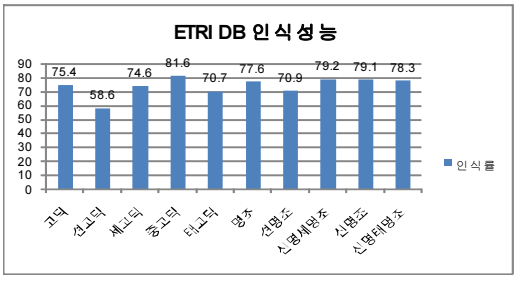
PHD08에 대한 특성으로서 기존 한글 인식기에 대한 인식 성능을 평가하였다. 또한 [표 2]에 제시된 기존의 두 가지 데이터베이스, ETRI DB와 CBNU한글00에 대해서도 같은 조건으로 성능 실험을 수행하여 난이도를 비교 평가하였다.

인식기로는 국내에서 널리 사용되는 상용화 인쇄 한글 인식기를 사용하였다. 데이터베이스의 샘플을 가지고 이미지 파일을 만들어 그것을 인식기의 입력으로 사용하였다. 글자 분할 과정에서 오류를 범하지 않도록 샘플 간의 여백을 넉넉히 배치하였다.

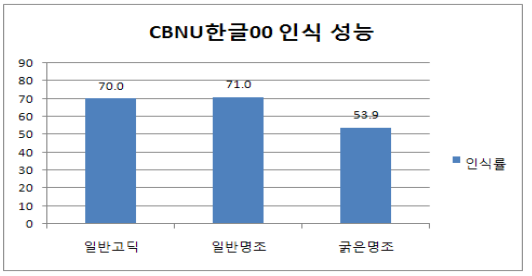
[그림 9]는 데이터베이스 별 인식 성능을 보여 준다. [표 4]는 세가지 데이터베이스에 대한 인식률을 보여 준다. [그림 9](c)의 PHD08에서는 명조, 고딕, 돋움의 인식 성능이 비슷하며, 바다, 헤드라인, 한양해서, 엽서체의 인식 성능이 현저히 낮은 것을 볼 수 있다. 즉 현재 한글 인식기는 다양한 글꼴에 적절히 대응하지 못하며 새로운 글꼴을 인식하기 위한 노력이 필요함을 의미한다.

표 4. 세가지 데이터베이스의 인식률

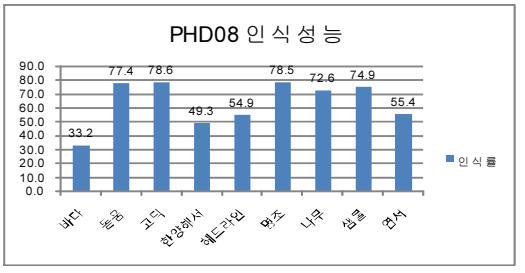
데이터베이스	ETRI	CBNU00	PHD08
인식률	74.59%	64.94%	63.87%



(a) ETRI DB



(b) CBNU한글00



(c) PHD08

그림 9. 데이터베이스별 인식률

Ⅴ. 결론

본 논문은 웹 문서나 자연 영상과 같은 새로운 영역에서 발생하는 문자 인식에 적합한 한글 데이터베이스를 구축한 과정과 활용 방법에 대하여 기술하였다. 이 데이터베이스는 9가지 종류의 글꼴에 대한 데이터를 포함하고 있으며 여러 변형과정을 거쳤으므로 양적으로나 질적으로 우수한 데이터베이스이다. 현재 국내에서 많이 쓰이는 한글 인식 소프트웨어에 대한 성능은 앞으로 PHD08을 이용하여 인쇄체 한글 인식기를 만드는 연구자에게 성능 기준을 제시해 준다. 이 데이터베이스는 웹에 공개하여 누구나 쉽게 쓸 수 있게 하였다.

참 고 문 헌

[1] 정인숙, 함대성, 오일석, “웹 이미지의 텍스트 추출을 위한 색 분산 방법의 실험적 평가”, 영상처리 및 이해에 관한 워크샵, p.36, 2008.
[2] 김지훈, 이택헌, 김기웅, 김진형, “상향식 접근 방

법을 사용한 강인한 간판 인식", 한국컴퓨터종합학술대회 논문집, 제34권, 제1호, pp.234-235, 2007.
[3] 최영우, 김길천, 송영자, 배경숙, 조연희, 노명철, 이성환, 변혜란, "계층적 특징 결합 및 검증을 이용한 자연이미지에서의 장면 텍스트 추출", 정보과학회논문지, 제31권, 제4호, pp.420-438, 2004.
[4] 오일석, *패턴인식*, 교보문고, 2008.
[5] 허기수, 오일석, "전주 한옥마을에서 수집한 간판 영상 데이터베이스", 콘텐츠학회논문지, 제6권, 제11호, pp.243-248, 2006.
[6] J. H. Jonathan, "A Database for handwritten Text Recognition Research," IEEE Tr. on Pattern Analysis and Machine Intelligence, Vol.16, No.5, 1994.
[7] 조성배, 김진형, "인쇄체 한글 문자의 인식을 위한 계층적 신경망", 한국정보과학회논문지, 제17권, 제3호, pp.306-316, 1990.
[8] 이판호, 장희돈, 남궁재찬, "동적자소분할과 신경망을 이용한 인쇄체 한글 문자인식에 관한 연구", 한국통신학회논문지, 제19권, 제11호, pp.2133-2146, 1994.
[9] 이진수, 권오준, 방승양, "개선된 자소 인식 방법을 통한 고인식률 인쇄체 한글 인식", 정보과학회논문지, 제23권, 제8호, pp.841-851, 1996.
[10] 정지호, 최태영, "원형 패턴 벡터를 이용한 인쇄체 한글 인식", 전자공학회논문지, 제38권, SP편, 제3호, pp.33-45, 2001.
[11] 임길택, 김호연, "문자형식 분류 기반의 인쇄체 문자인식에 관한 연구", 전자공학회논문지, 제40권, CI편, 제5호, pp.26-39, 2003.
[12] 이성훈, 조규태, 김진식, 김진형, 정철곤, 김상균, 문영수, 김지연, "저해상도 인쇄체 한글 인식을 위한 자소 분할 방법", 한구컴퓨터종합학술대회 논문집, pp.382-384, 2006.
[13] 김병기, "유형의 상대적 크기를 고려한 한글문자의 유형 분류", 한국컴퓨터정보학회논문집, 제11권, 제6호, pp.99-106, 2006.
[14] 김대인, 김상엽, 이성환, "대용량 오프라인 한글 글씨 영상 데이터베이스 KU-1의 설계 및 구축", 제9회 한글 및 한국어정보처리 학술발표 논문집, pp.152-159, 1997.
[15] http://ai.kaist.ac.kr/
[16] 김대환, 방승양, "한글 필기체 영상 데이터베이스 PE92의 소개", 제 4회 한글 및 한국어 정보처리 학술발표 논문집, pp.567-575, 1992.
[17] http://cv.chonbuk.ac.kr/

저 자 소 개

함 대 성(Dae-Sung Ham) 정회원

- 2007년 2월 : 전북대학교 컴퓨터공학(공학사)
- 2007년 3월 ~ 현재 : 전북대학교 컴퓨터공학 석사과정

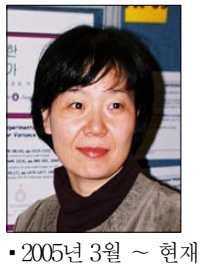
<관심분야> : 컴퓨터비젼, 패턴인식, 한글인식

이 득 용(Duk-Ryong Lee) 정회원

- 2004년 8월 : 전북대학교 컴퓨터공학(공학사)
- 2006년 8월 : 전북대학교 컴퓨터공학 석사
- 2006년 9월 ~ 현재 : 전북대학교 컴퓨터공학 박사과정

<관심분야> : 컴퓨터비젼, 패턴인식

정 인 숙(InSuk Jung) 정회원



- 1989년 2월 : 이화여자대학교 수학과(이학사)
- 1991년 2월 : 포항공과대학교 수학과(이학석사)
- 1991년 2월 ~ 1998년 : ETRI 인공지능연구실 선임연구원
- 2005년 3월 ~ 현재 : 전북대학교 컴퓨터공학 박사과정

<관심분야> : 컴퓨터비젼, 패턴인식, 로봇비젼

오 일 석(Il-Seok Oh) 정회원

- 1984년 : 서울대학교 컴퓨터공학과(공학사)
- 1992년 : KAIST 전산학과 박사
- 1992년 9월 ~ 현재 : 전북대학교 컴퓨터공학 석사과정

<관심분야> : 컴퓨터비젼, 패턴인식