대용량 오프라인 한글 글씨 영상 데이터베이스 KU-1의 설계 및 구축

김 대인†, 김 상엽†, 이 성환††
†고려대학교 대학원 컴퓨터학과
†고려대학교 대학원 영상정보처리학과

Design and Construction of a Large-set Off-line Handwritten Hangul Character Image Database KU-1

Dae-In Kim†, Sang-Yup Kim†, Seong-Whan Lee††
†Dept. of Computer Science and Engineering, Korea University,
†Dept. of Visual Information Processing, Korea University

E-mail: {dikim, sykim, swlee}@image.korea.ac.kr

요약

최근 들어 인쇄체 문자 인식 기술의 발전에 힘입어 필기체 한글 인식에 관한 연구가 활발히 진행되고 있다. 인쇄체 문자와는 달리 자연스럽게 필기된 한글 글씨는 동일한 문자라 하더라도 같은 모양을 가지고 있다고 단정하는 것이 불가능할 정도로 필기자의 필기 유형에 따른 다양한 변형을 내포하고 있다. 따라서 효과적인 한글 글씨 인식기를 개발하기 위해서는 다양한 변형을 포함하는 대용량의 한글 글씨 영상 데이터베이스가 필수적이다.

본 논문에서는 시스템공학연구소 주관 국어 정보 베이스 개발 사업의 일환으로 고려대학교에서 구축 중인 오프라인 한글 글씨 영상 데이터베이스, KU-1에 대해 간략히 소개하고자 한다. 본 데이터베이스는 KS C 완성형 한글 사용 빈도순 상위 1,500자에 대하여 다양한 계층, 직업, 연령, 지역 분포를 고려한 1,000명 이상의 필기자가 정서체와 본인의 평소 자유 필체로 필기한 1,000벌의 명도 한글 글씨 영상으로 구성되어 있다.

I. 서론

오프라인 한글 글씨 인식에 관한 연구는 지난 20 여년간 국내의 대학, 연구소 및 기업체를 중심으로 꾸준히 진행되어 왔으나, 입력 문자 영상에 포함된 잡영이나 왜곡 그리고 다양한 필체의 변형 등을 해결해야 하는 어려움으로 인하여 실용화 단계에는 이르지 못하고 있는 실정이다. 최근 들어 오프라인 한글 글씨 인식에 관한 관심이 고조되어 많은 연구 결과들이 발표되고 있으나, 이러한 연구 결과들이 오프라인 한글 글씨 인식의 실용화를 이끌기에

는 아직도 많은 부분에서 취약성을 보이고 있다[이성환 93].

지금까지의 연구 결과를 분석해 볼 때 오프라인 한글 글씨 인식 기술이 아직도 실용화 단계에 이르지 못하고 있는 이유로는 인식 알고리즘의 개발에 필요한 한글 글씨 영상 데이터의 수집에 드는 비용과 노력으로 인하여 연구자들이 쉽게 오프라인 한글 글씨 인식 분야에 접근할 수 없었다는 점과 우수한 오프라인 한글 글씨 인식 알고리즘의 개발을 유도하고 그 성능을 객관적으로 평가해 볼 수 있는 공동의 한글 글씨 영상 데이터베이스가 부족했다는 점을 들 수 있다.

일본이나 미국 등에서는 이미 공동의 필기체 문자 데이터베이스가 구축되어 효과적으로 사용되고 있으며 [Fenri93, Saito85, Wilso90], 국내에서도 오프라인 한글 글씨 영상 데이터베이스의 구축이 시도된 바 있지만 수집된 데이터의 양이 부족하고 잘못 레이블링 되거나 문자 단위 분할 오류 등으로 인하여 사용하는데 상당한 불편함이 있는 것으로 알려져 있다[이성환94].

본 논문에서는 시스템공학연구소 주관 국어 정보 베이스 개발 사업의 일환으로 고려대학교에서 구축 중인 오프라인 한글 글씨 영상 데이터베이스 KU-1에 대해 간략히 소개한다. 오프라인 한글 글씨 영상 데이터베이스 구축은 그림 1에서 보여지는 바와 같이 크게 4 단계로 구성되며, 각 단계별 고려 사항 및 수행 과정들이 이후의 장들에서 기술될 것이다.

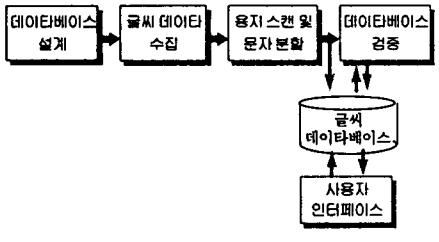


그림 1. 한글 글씨 영상 데이터베이스 구축 단계

II. 한글 글씨 영상 데이터베이스 설계

2.1 데이터베이스 구축시의 고려사항

한글 글씨 영상 데이터베이스는 광범위한 계층의 필기자를 선정하여 필기하도록 함으로써 특정 요소에 의해 편중되어 발생될 수 있는 국부적인 필기 특성을 배제하고 다양한 필기 변형을 충분히 포함하는 대용량의 글씨 데이터베이스이어야 한다.

오프라인 한글 글씨 인식기의 실용성을 높이기 위해서는 필기 수집 용지의 종류와 필기 도구에 구애를 받지 않아야 한다. 더욱이, 오프라인 한글 글씨 인식기는 기계로부터 읽어 들인 영상을 바탕으로 인식 작업을 수행하게 되므로 필기 수집 용지의 재질에 따라 성능이 좌우될 수 있으므로, 다양한 재질의 수집 용지를 사용하여 한글 글씨 영상 데이터를 수집해야 한다.

또한, 필기 도구는 동일한 사람이 같은 문자를 필기하더라도 사용된 필기 도구에 따라 서로 다른 특성을 가지게 되므로, 다양한 필기 도구로 필기된 한글 글씨 영상 데이터를 수집해야 한다.

뿐만 아니라 수집 용지로부터 문자 영상을 스캔하여 문자 단위로 분할 및 저장하는 방법에 있어서는 이진 영상으로 인한 정보 손실을 최소화하기 위해 칼라 또는 명도 영상으로 저장하는 것이 바람직하다.

2.2 설계 내용

본 연구에서는 국내외의 5 종류의 오프라인 글씨 데이터베이스 구축의 사례[김정규93, 방승양92, 양영규92, Fenri93, Saito85]를 비교 및 분석함으로써 오프라인 한글 글씨 영상 데이터베이스가 가져야 할 기준을 새로이 정립하였으며, 데이터베이스의 설계시에 다음의 요소들을 고려하였다.

- 필기 환경의 적합성
  - 수집 용지의 재질, 두께, 특성 및 형식
  - 필기구의 종류
  - 필기 형태
- 필기자의 다양성
- 데이터베이스의 완전성
  - 다양한 필기 형태의 포함 여부
  - 데이터베이스의 품질
  - 사용시의 편의성
- 인식 알고리즘 개발 및 성능 평가에의 활용 정도

본 연구에서는 고품질의 오프라인 한글 글씨 영상 데이터베이스의 구축을 위하여 현재 한글 글씨 인식 관련 연구를 수행 중에 있는 20여명의 연구자들을 초청하여 한글 글씨 영상 데이터베이스의 설계에 관한 자문회의를 개최하였으며, 자문회의 내용을 바탕으로 하여 데이터베이스를 설계하였다. 본 데이터베이스는 KS C 완성형 한글 사용 빈도순 상위 1,500자에 대하여 다양한 계층, 직업, 연령, 지역 분포를 고려한 1,000명 이상의 필기자가 정서체와 본인의 평소 자유 필체로 필기한 1,000벌의 명도 한글 글씨 영상으로 구성함을 목표로 한다.

한글 글씨 영상 데이터 수집 용지는 다음과 같이 설계되었다.

- 크기 : A4
- 색상 : 백색
- 재질 : 양면 아트지, 건식 복사지, 갱지
- 형식 : 10 종류의 서로 다른 문자 배열로 구성
- 필기 칸
  - 크기 : 문자당 9mm x 9mm의 사각형
  - 색상 : 적색
  - 필기 칸 간의 간격 : 가로 2mm, 세로 5mm
- 예시 문자
  - 크기 : 10pt 고딕체
  - 색상 : 적색
  - 위치 : 필기할 칸의 상단에 인쇄
- 수집 용지에 포함된 필기자 정보 : 성별, 좌우 손잡이 여부, 필기구 종류, 필기 형태, 지역, 직업, 나이, 이름

한글 글씨 영상 데이터 수집 시에 사용된 필기구의 특성은 다음과 같다.

- 색상 : 흑색
- 종류 : 사인펜, 볼펜, 수성 마킹펜

수집 용지의 스캔 및 문자 단위 영상의 저장 과정에서 문자 추출 영역은 필기 칸 보다 우측과 하단이 1mm 큰 10mm x 10mm 크기로 설계되었고, 300DPI의 해상도와 256 단계(1 화소당 1 byte)의 명도 영상으로 스캔 및 저장된다. 문자 영상은 단위 문자에 대하여 위에서 아래로, 좌에서 우의 순서로 저장된다. 한글 글씨 영상 데이터 한 벌을 완성형 한글 코드순으로 하나의 화일에 저장하여 데이터베이스를 구성하였다. 이렇게 구성된 데이터베이스는 MYLEX DOC960S Disk Array(60GB)에 저장된다. 그림 2는 한글 글씨 영상 데이터 한 벌이 하나의 화일에 저장되는 순서를 나타내 준다.

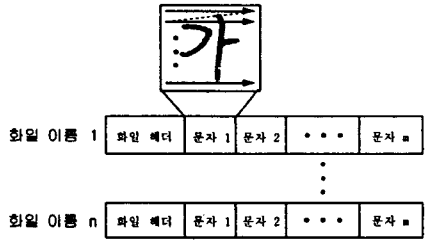


그림 2. 화일에 저장되는 순서

III. 한글 글씨 영상 데이터 수집

한글 글씨 영상 데이터의 수집 방법은 다음과 같다. 서로 다른 지역에서 광범위한 분포의 필기자를 선정한 다음

미리 준비한 일상 생활에서 많이 사용되고 있는 수집 용지와 필기구를 사용하였다.

필기자에게는 정서체와 자유 필체, 두 종류의 글씨체로 필기하도록 하였으며, 필기전에 "필기시의 주의 사항"을 제시하여 숙지하도록 함으로써 데이터베이스의 품질을 유지하고자 하였다. 그림 3은 정서체로 필기된 수집 용지의 한 예이다.

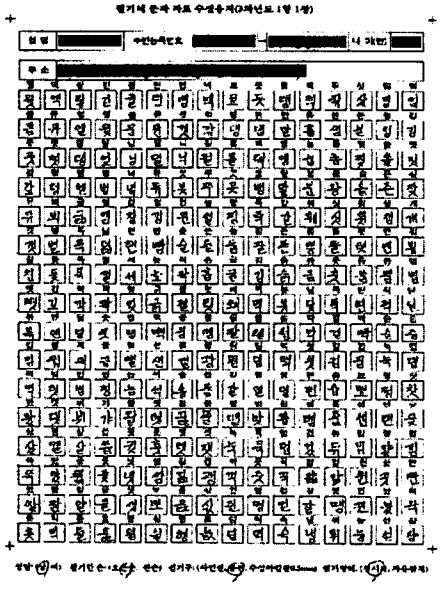


그림 3. 필기된 수집 용지의 예

3.1 필기자

필기자는 서로 다른 지역에서 다양한 계층의 연령, 성별 분포를 고려하여 1,000명 이상을 선정하였다. 필기자의 지역별 분포가 그림 4에 나타나 있다.

3.2 한글 글씨 영상 데이터 수집 용지

한글 글씨 영상 데이터 수집 용지는 A4 크기의 백색

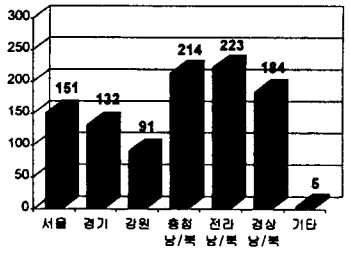
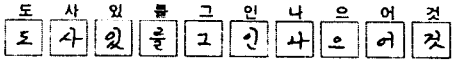


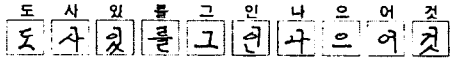
그림 4. 필기자의 지역별 분포

용지를 선택하였으며, 양면 아트지, 건식 복사 용지, 갱지, 세 종류의 재질을 사용하였다.

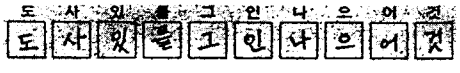
그림 5는 수집 용지의 재질 별로 스캔된 문자 영상의 상태를 나타낸다. 종이의 재질이 갱지인 경우, 명도 영상으로 스캔시 배경이 엷게 포함됨을 보여준다.



(a) 양면 아트지



(b) 건식 복사 용지



(c) 갱지

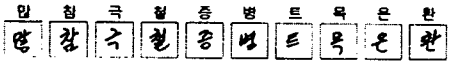
그림 5. 필기된 수집 용지의 재질별 명도 영상의 예

3.3 필기 도구

필기 도구는 일상 생활에서 일반적으로 많이 사용되는 볼펜, 싸인펜, 수성 마킹펜, 세 종류의 필기 도구로 제한하였으며, 색상은 흑색으로 정하였다. 필기 도구는 일괄적으로 필기자에게 제공하여 필기하도록 함으로써 필기 문자의 품질이 균일하게 하도록 유도하였고, 각 필기 도구의 종류별 분포 또한 균일하게 유지하였다. 그림 6은 필기 시에 사용된 필기 도구별 필기 결과를 보여 준다.

(a) 볼펜

(b) 싸인펜



(c) 수성마킹펜

그림 6. 필기 도구별 필기 결과의 예

IV. 한글 글씨 영상 데이터 수집 용지의 스캔 및 문자 분할

수집된 용지 내의 문자 영상 데이터는 스캐너를 통하여 입력되어 문자 단위의 분할 과정을 거쳐 일련의 구조를 갖는 화일로 저장되어야 한다. 본 연구에서는 수집 용지에 붉은 색으로 사각형의 필기 영역을 표시하여 문자 분할의 효율과 정확도를 높이고자 하였다. 일반적으로, 명도 영상으로 스캔하여 저장할 경우에는 스캔된 수집 용지의 인쇄된 붉은 색 부분이 어느 정도의 명도값을 가지게 되어 복잡한 문자 분할 작업을 요구한다.

명도 영상에서의 문자 분할을 어렵게 만드는 경우는 다음과 같이 분류할 수 있다.

- 필기 칸에 접촉되거나 벗어나게 필기된 경우
- 수집용지의 재질이 갱지인 경우
- 스캔된 영상이 기울어진 경우
- 잘못 필기된 경우

그림 7은 스캔시 수집 용지가 기울어져 스캔된 경우를 보여준다. 이 경우, 영상의 기울어짐으로 인하여 문자 단위 분할 시에 문자 분할 영역이 문자 영상을 모두 포함하지 못하거나 필기 칸의 선 부분이 문자 영상에 잡영으로 첨가될 수 있다. 특히, 필기된 수집 용지 내의 글자 크기가 필기 영역을 나타내는 사각형 보다 클 경우 분할의 어려움이 더하게 된다. 본 연구에서는 분할 영역의 크기를 좀 더 크게 설정함으로써 문자 영상이 잘리는 문제를 해결하였으며 검증 단계를 두어 분할 영역 내에 존재하는 필기 칸들을 효과적으로 제거하였다.

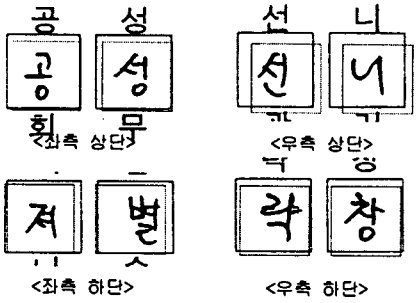


그림 7. 스캔된 영상이 기울어진 경우의 예

그림 8은 필기 시에 필기자의 부주의로 인하여 예시 문자와는 다른 문자를 필기했을 때 이를 수정하기 위하여 X 표로 표시한 경우를 보여 준다. 이 경우, 수집 용지의 여분 필기 칸에 다시 필기된 문자로 대치시켜야 하므로 별도의 처리 과정을 거치게 된다.

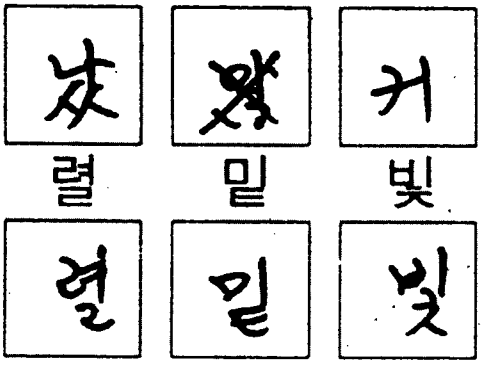


그림 8. 필기자의 부주의로 인하여 잘못 필기된 문자 영상의 예

V. 데이터베이스 검증

문자 분할 과정을 통하여 얻은 문자 영상 데이터들로부터 수집 용지의 필기 영역을 나타내는 사각형을 제거하고, 필기자의 실수로 인하여 잘못된 문자들을 교체하였다 하더라도 저장된 문자 영상 데이터들에는 많은 문제들이 내포될 수 있다. 따라서, 문자 분할 과정만을 거친 문자 영상 데이터들을 직접 문자 인식 연구에 사용하기에는 많은 어려움이 있다.

예를 들어, 수집된 용지를 스캔하여 저장하는 과정에서 실수로 헤더 정보가 잘못 입력될 수 있고, 필기된 문자가 필기 영역을 나타내는 사각형에 접촉된 경우 문자 영상만을 정확하게 분리할 수 없기 때문에 분할 오류가 발생할 수 있다. 또한, 필기자가 두종류의 서로 다른 문자를 습관적으로 유사하게 필기하거나 또는 전혀 다른 문자로 필기함으로써 오류가 발생할 수 있고, 문자 영상 데이터와 레이블된 코드가 서로 일치하지 않는 레이블링 오류가 발생할 수 있다. 따라서, 구축된 데이터베이스의 품질을 향상시키기 위해서는 데이터베이스 내에 저장된 문자 영상 데이터를 검증하여 이러한 오류를 찾아내고 이를 수정해야 한다.

본 연구에서 고려한 데이터베이스내의 오류 항목과 그 원인을 요약하면 다음과 같다.

- 헤더 정보 오류
  - 문자 영상이 저장된 화일의 헤더 정보가 잘못 입력된 경우로서 데이터 저장 과정에서 발생한다.
- 문자 단위 분할 오류
  - 문자 영상의 일부가 잘려서 저장된 경우로 문자 분할 시에 사용된 임계값이 문자 영상 스캔시에 적합하지 않을 경우 발생한다.
  - 잡영 또는 필기 영역을 나타내는 사각형의 일부가 남아 불필요한 공백이 들어간 경우 문자 분할 시에는 이 부분도 문자 영상으로 간주하여 데이터베이스에 저장될 수 있다.
- 레이블링 오류
  - 필기자가 두개의 서로 다른 문자를 습관적으로 유사하게 필기한 경우 발생한다.
  - 필기자의 실수로 전혀 다른 문자로 필기한 경우 발생한다.

본 연구에서는 구축된 데이터베이스의 오류를 검증하고 이를 효과적으로 교정하기 위하여 검증 및 교정 도구를 개발하여 사용하였다. 그림 9는 구축된 데이터베이스의 검증 시에 사용된 도구를 이용하여 불필요한 잡영을 제거하는 처리 과정의 예를 보여 준다.

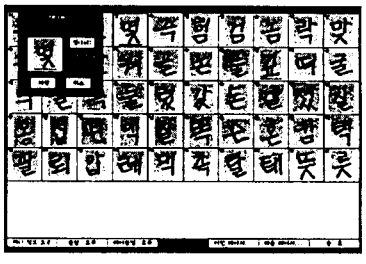


그림 9. 데이터베이스 검증의 예

또한, 교정이 불가능한 오류 데이터에 대해서는 해당하는 문자 영상을 공백으로 처리함으로써 데이터베이스 품질 향상은 물론 잘못된 문자에 대한 훈련이나 인식을 통하여 인식시스템의 성능이 저하되는 것을 방지하고자 하였다.

VI. 사용자 인터페이스

본 연구에서 구축된 오프라인 한글 글씨 영상 데이터베이스는 한글 글씨의 인식에 관한 연구를 수행하는 연구자들이 사용하기 편리한 형태로 구성되었다. 구축된 데이터베이스는 새로운 데이터의 삽입 또는 기존 데이터의 삭

제 등이 거의 발생하지 않는 특성이 있으므로 사용자들이 필요로 하는 데이터의 검색뿐만 아니라 데이터베이스에 대한 일반적인 특성들이 이해하기 쉬운 형태로 구성되어야 한다.

본 연구에서는 WWW(World Wide Web)의 HTML을 이용하여 편리한 사용자 인터페이스를 구현하였다. 본 사용자 인터페이스의 두 가지 주요 기능은 다음과 같다. 하나는 본 오프라인 한글 글씨 영상 데이터베이스에 대한 일반적인 소개 기능이다. 다른 하나는 KS C 완성형 한글 사용빈도순 상위 520자중 사용자들이 필요로 하는 각 문자에 대한 50개의 영상데이터를 보여줌으로써 본 데이터베이스에 대한 품질을 평가할 수 있는 기능을 갖는다. 그림 10은 50개의 '가'에 대한 데이터영상의 예를 보여준다.

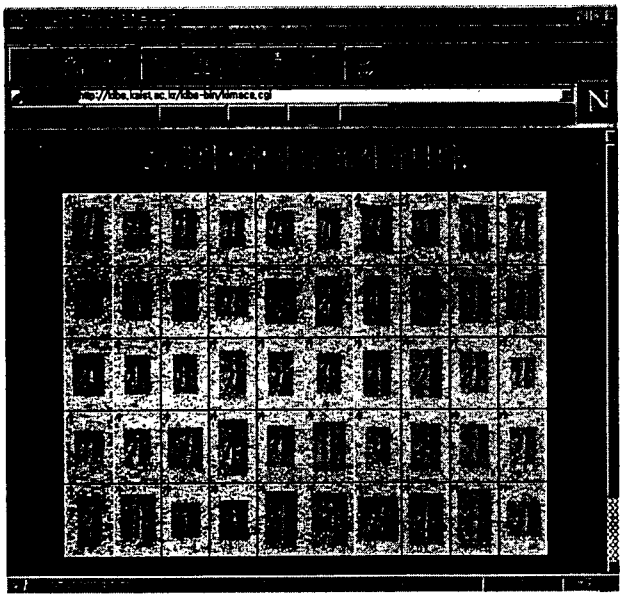


그림 10. 50개의 '가'자에 대한 데이터 영상

VII. 결론

본 논문에서는 시스템공학연구소 주관 국어 정보 베이스 개발 사업의 일환으로 고려대학교에서 구축 중인 오프라인 한글 글씨 영상 데이터베이스 KU-1에 대해 간략히 소개하였다. 본 연구에서는 다양한 변형을 갖는 글씨체의 수집을 데이터베이스 구축시 가장 고려해야 할 요소로 삼았으며, 고품질의 일관성 있는 대용량 오프라인 한글 글씨 영상 데이터베이스를 효율적으로 구축하기 위하여 수집 용지를 명도 영상으로 스캔하고 문자 단위로 분할한 다음 이를 정해진 데이터 화일로 저장하는 과정과 구축된 데이터베이스에 대한 검증을 수행하여 불필요한 잡영을 제거하고 레이블링 오류 및 문자 단위 분할시의 오류를 교정하는 과정을 도구화하였다.

현재 구축 중에 있는 한글 글씨 영상 데이터베이스는 조만간 국내의 오프라인 한글 글씨 인식 연구자들에게 이용하기 편리한 형태로 제공될 예정이며, 이는 국내의 오프라인 한글 글씨 인식에 관한 연구를 활성화시켜주는 계기가 될 것으로 기대된다.

더욱이, 문자 인식 시스템의 상용화를 앞당기기 위해서는 한글 뿐 만 아니라 우리 나라 사람들이 필기한 숫자, 한자 등에 대한 낱자 데이터베이스는 물론, 연속 필기시의 글자 변형을 포함하는 연속 필기 데이터베이스가 필요하므로 이들에 대한 데이터베이스 구축 또한 조속히 이루어져야 할 것으로 판단된다.

감사의 말씀

본 연구는 시스템 공학 연구소 주관 국어 정보 베이스 구축 사업중 "오프라인 한글 글씨 데이터베이스 구축" 과제의 연구비 지원을 받았음.

참고문헌

[김정규93] 김정규, 강태호, 조성익, 양영규, "필기 한글 문자 데이터베이스의 품질평가 방안 연구," 제1회 문자인식 워크샵 발표논문집, 청주, 1993년 5월, pp. 61-66.

[방승양92] 방승양, 한글 필기체 영상 데이터베이스의 구축, 한국전자통신연구소 제출용 최종 보고서, 포항공과대학, 1992년 6월.

[양영규92] 양영규 등, VIP 공동 연구 환경 기반 구축 (III), 과학기술처 제출용 최종보고서, 시스템 공학연구소, 1992년 9월.

[이성환93] 이성환, 박희선, "한글 인식 사례 연구 : 최근 5년 동안의 연구 결과를 중심으로," 제 1회 문자인식 워크샵 발표논문집, 청주, 1993년 5월, pp. 3-46.

[이성환94] 이성환, "오프라인 필기체 문자 데이터베이스 구축의 사례 연구," 한국정보과학회 한국어 정보처리 연구회 소식지, 2권 1호, 1994년 10월, pp. 2-13.

[Fenri93] R. Fenrich and J. J. Hull, "Concerns in Creation of Image Databases," Proc. of 3rd Int. Workshop on Frontiers in Handwriting Recognition, Buffalo, New York, USA, 1993.

[Saito85] T. Saito, H. Yamada, and K. Yamamoto, "On the Data Base ETL 9 of Handprinted Characters in JIS Chinese Characters and Its Analysis," 일본 전자통신학회 논문지, Vol. J68-D, No. 4, Apr. 1985, pp. 757-764.

[Wilso90] C. L. Wilson and M. D. Garris, "Handprinted Characters Database," NIST, 1990.