필기체 문자 인식을 위한 문자 영상 데이터 구축에 관한 연구

이향란(H. R. Lee)*, 고경철(K. C. Ko)*, 이말례(M. R. Lee)**

군산대학교 정보통신전파공학부,컴퓨터과학과 대학원*, 여수대학교 멀티미디어학부**

A Study of Construction of Character Image Data for Recognition Handwritten Text

ABSTRACT

In order to develop a character recognition system, it is an essential preceding work that gathers an image data of the standard. On this purpose a data of the digitized images of a handwritten characters was collected. The types of a gathered image data are Korean character, Chiness character, Numeral, English character, Special character, and so on. This paper deals with a handwritten character image data base, and the image data base different from the general storage structure of a large capacity multimedia was designed and builded.

1.서 론

인쇄체 문자에 비하여 필기체 문자가 더 복잡한 변형을 포함하고 있다. 따라서, 필기체 문자를 인식하기 위한 알고리즘을 개발하기 위해서는 문자인식 실험에 사용될 다양하게 변형된 문자 데이터를 수집할 필요가 있다. 왜냐하면 사람마다 필기형태가 다양하기 때문에 다양하게 필기된 문자 영상 데이터를 수집해야 하며, 인식 실험 시 다양하고 유용한 필기체 문자를 제공받아 인식률을 측정해야 효율적인 문자인식 알고리즘으로서 평가받을 수 있다. 그러므로 필기체 문자 데이터는 앞으로 문자인식 연구에 있어서 필수적 요소라 할 수 있다. 본 논문에서 구성하고자 하는 문자영상 데이터는 다음과 같은 특성을 반영해 줄 수 있는 필기체 문자 데이터를 구축하는데 있다.

첫째, 필기체 문자인식 실험에 필요한 다양한 데이터를 제공해줌으로써, 연구의 활성화와 체계적인 연구 환경을 조성하는 것이다. 오프라인 필기체 문자인식 알고리즘을 연구하기 위해 필요한 다양하고 유용한 데이터를 얻기 위해서 소모하는 시간을 줄임으로써 좀더 문자인식 알고리즘 개발에 전념할 수 있는 환경이 조성된다.

둘째, 무 제약의 서식문서 필기체 문자 데이터의 표준화이다. 문자인식 시스템간의 객관적인 성능비교를 위해서는 표준화된 필기체 문자영상 데이터베이스를 구축하는 것이 필수적이다.

셋째, 서식 문서 내에서 선에 겹쳐진 문자와 겹쳐지지 않은 문자들을 따로 분리하여 저장함으로써 서식 문서 내에서 선에 겹쳐진 글자의 선 제거 및 문자복원을 할 수 있는 알고리즘 개발용 데이터를 제공한다. 제약이 가해지지 않은 필기체 문자 데이터는 인식이 어려울 뿐만 아니라 서식문서에 작성하는 경우가 많기 때문에 인식실험을 위해서는 선 제거가 일차적인 목표가 되고 선 제거 후

분리된 글자의 복원이 이차적인 전처리 기법이 된다.

필기체 문자인식에 있어서도 좋은 연구결과들이 발표되고 있지만 이들 연구 대부분은 정해진 위치나 사각형 내에 필기하도록 표기 형태를 제한하고 있으며, 특히 오프라인 필기체 한글인식의 경우 연속 필기 문자열의 인식에 관한 연구는 없다. 따라서 본 논문에서 개발한 데이터는 오프라인 문자인식 실험에 필요한 충분한 양의 글자 수와 다양한 크기와 다양한 변형을 수용하는 무 제약의 필기체 문자들을 문자열 단위로 수집하여 저장시킨 오프라인 필기체 문자 영상 데이터이다.

본 논문의 구성은 다음과 같다. 2장에서는 필기체 문자 영상 데이터 구축에 필요한 연구 범위를 설명하고, 3장에서는 문자 영상 데이터 구축를 보이며 4장에서는 본 논문의 결론 및 추후 연구에 대해 제시한다.

2. 문자 영상 데이터 구성

2.1 영상 데이터

문자의 종류로는 한글, 한자, 숫자, 특수문자 들을 제약이 가해지지 않은 상태에서 서식문서 내에 자유롭게 필기하게 한 후 문자열 단위로 추출하여 저장한다. 본 논문에서 구현한 문자 영상 데이터 베이스에 저장되는 문자열의 단위는 한글이름, 한자이름, 주소, 우편번호, 서명, 성별 부분으로 총 6개 부분의 문자열 단위로 추출하여 저장시키는 데이터 베이스이다. 문자영상 데이터 베이스 구축 시 6개 부분 중 주소 부분 문자 영상은 한글 영상 데이터가 주종을 이루고, 간혹 한자로 표기된 문자 영상을 획득할 수도 있으므로 데이터 베이스 구축 시 한글 영상 데이터인 경우는 K(Korean character)마크를 넣어서 저장하고, 한자 영상 데이터인 경우는 C(Chinese character)마크를 넣어서 저장한다. 그리고 성별부분은 간혹 영문자로 표기하는 사람이 있어서 헤더 정보 란에 영문자 E(English character)로 표기해서 저장시켰다. 숫자 영상 데이터인 우편번호 부분은 숫자 영상인 경우만 존재하므로 헤더 정보 란에 영상의 표기 상태를 저장하지 않았다. 또한 특수문자 영상 데이터인 서명 부분도 숫자 영상 데이터의 경우와 마찬가지로 영상의 표기 상태를 저장하지 않았다. 영상 데이터 구축 시 고려한 중요한 사항으로는 무 제약의 필기체 문자 영상 데이터를 저장시키는 과정에서 형식을 갖춘 서식 문서 내에 필기자가 무 제약의 필기체 문자를 필기하면서 라인에 겹쳐지게 필기한 데이터와 라인에 겹쳐지지 않은 데이터를 따로 구분하여 저장시키는 방법이다.

2.2 영상 데이터 베이스 구축 시 고려사항

필기체 문자를 획득하는 과정에서 필기대상자들은 남녀간 비율, 다양한 직업들, 골고루 분포된 연령층을 모두 고려하여 어느 특정한 집단으로 치우치지 않고 각계 각층의 필기체 자료를 수집하였다. 왜냐하면 다양하게 변형된 데이터를 얻기 위해서는 다양한 계층의 자료 수집이 필수적이기 때문이다. 그러므로 다양한 계층의 자료를 획득하는 가장 좋은 방법은 어느 특정한 집단을 대상으로 하는 것 보다는 가족 단위로 자료를 수집한다면 가장 효율적인 필기체 영상 데이터가 될 수 있다. 가장 큰 비중을 두는 분야는 역시 다양하게 필기된 데이터의 수집이다. 이렇게 수집된 자료간에도 비슷한 글씨체로 쓰는 사람이 발생하게 되고, 독특한 글씨체로 쓰는 사람도 발생하게 된다. 본 논문에서는 수집된 자료를 대상으로 흘림의 정도, 나이, 겹침의 정도 등을 구분하여 표 1과 같이 분류 하였다.

Table 1 Classification of data by writter

분류	영상 데이터의 특성
CF00	심하게 흘려쓴 필기체 영상 자료
CF01	10대와 20대의 영상 자료
CF02	20대가 필기한 영상 자료
CF03	20대와 30대 이상의 영상 자료
CF04	라인에 겹쳐진 글자가 없는 자료
CF05	sign부분만 겹쳐진 영상 자료
.....	
...	

대상으로 하는 인식 시스템은 오프라인 필기체 문자인식 시스템으로 문자열 단위로 인식함을 대상으로 한다. 본 논문에서 구축한 영상 데이터 베이스를 대상으로 한 필기체 문자 데이터는 그림1과 같이 일정한 서식 문서에 필기한 필기체 데이터를 수집하였다.

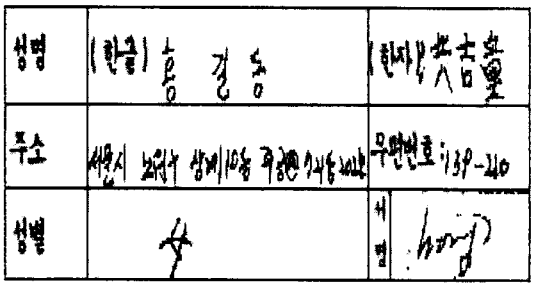

Fig .1 Example of handwritten

이렇게 저장되는 서식문서는 628장에 이른다. 서식문서 628장에 필기된 문자들은 설계된 방식에 따라 새로운 문자 영상 데이터를 만들어 CD-ROM에 저장하였다. 처음 만든 문자 영상 데이터는 가능하면 잡음이 섞이지 않은 선명한 영상 데이터이다. 선명한 영상 데이터를 얻기 위해서는 문자의 명도와 배경의 명도가 많이 차이 날수록 좋으므로 문자는 검게, 배경은 희게 하도록 했다.

영상 데이터의 저장 방식은 256 gray level을 선택했다. 이는 영상의 명도를 깨끗하게 나타낼 수 있으며, 영상 정보를 충분히 손실 없이 얻을 수 있다. 영상 데이터의 해상도는 400dpi로 스캔하여 6개의 부분으로 구분하여 저장시켰다.

2.3 문자 영상 데이터의 구성 원칙

저장된 영상 데이터를 중심으로 설명하면 다음과 같다. 각각의 영상 데이터에 대해서 파일명을 부여하여 저장시키는데 6개의 모든 부분에 대해서 적용된 파일명 부여의 원칙은 다음과 같다. 파일명의 첫 번째 글자는 O또는 I가 온다. 영문자 O는 샘플된 영상 데이터가 서식 문서의 선에 겹쳐진 글자가 존재할 때 부여되는 글자이고, 선에 겹쳐진 글자가 존재하지 않을 때는 영문자 I를 표기한다. 파일명의 "_"다음의 두 개의 숫자가 오는데 이 숫자는 샘플된 영상 데이터의 순서를 나타낸다. 이때 O타입과 I타입은 각각 따로 순서를 부여하여 저장시킨다. 헤더 정보 란에 내용을 기입하지 않는 디렉토리는 특수문자 영상 데이터 부분인 서명 디렉토리로써 각각의 파일 내용은 서로 다르지만 특수하게 표기하는 문자 데이터가 많기 때문에 별도로 영상 데이터의 내용을 표기하지 않았다.

본 논문에서 구축된 오프라인 필기체 문자 영상 지식베이스는 영상 지식베이스가 구축되는 과정에서 스캐너가 붙어 있는 컴퓨터와 CD-ROM이 장착되어 있는 컴퓨터가 서로 네트워크로 연결되어 있어서 스캔된 영상 데이터의 효율적인 전송을 위해서 대용량의 영상 데이터를 ARJ를 사용하여 하나의 파일로 압축한 후에 전송하는 방법을 사용하였다. 전송된 파일은 원래의 영상 파일로 압축을 풀어서 CD-ROM에 저장하였다. 오프라인 필기체 문자 영상 데이터를 CD-ROM에 저장시키므로 실제 구현된 영상 데이터는 압축 알고리즘이 불필요하여 압축하지 않고 원래의 이미지 그대로 저장시켰다.

2.4 필기체 문자 영상 데이터 베이스의 구조

본 논문의 영상 문자 데이터 베이스의 구조는 6개의 field로 구성하였으며 이들의 구조는 다음과 같다.

NAME -->NAME_K(한글), NAME_C(한자)
ADDRESS --> AD1(시,도), AD2(구,시,군), AD3
(예외구), AD4(동,읍,면), AD5
(리, 기타 나머지 주소부분)
ZIP --> (우편번호)
SEX --> (성별)
SIGN --> (서명)
INDEX -->(전체5개 필드를 연결해 주는 인덱스)

2.4.1 각 디렉토리별 파일 설계

앞의 구조에 따라 영상 데이터의 header는 다음과 같이 정리하여 구성하였으며 그 내용을 정리해 보면 다음과 같다.

(1) 이름

No	Block_Id	Width	Height	N름Info	Null	Raster Image
int(2)	int(2)	signed long(4)	signed long(4)	Unsigned (10)	char(2)	variable size

<--------- 헤더(24 bytes) ---------->

Block_Id값에서 1은 한글이름, 2는 한자이름을 의미하고, 각자 ID아래의 숫자는 바이트의 크기를 나타낸다.

(2) 주소부:

전체주소(FULL SCAN 한 주소 영상)

No	Block_Id	Ma가_Form	Witdth	Height	시작주소의화일명	Null	Raster Image
int(2)	int(2)	K:한글 C:한자 E:영문	signed long(4)	signed long(4)	char(8)	char(2)	variable size

<----------헤더(23bytes)---------->

(3) 우편번호

No	Block_Id	Width	Height	ZIP_Info	Raster Image
int(2)	int(2)	signed long(4)	signed long(4)	char(6)	variable size

<--------헤더(18bytes)---------->

Block_Id 값이 5인 경우는 우편번호부의 숫자영상 데이터를 의미한다.

(4) 성별

No	Block_Id	Ma가_Form	Witdth	Height	Sex_Info	Null	Raster Image
int(2)	int(2)	K:한글 C:한자 E:영문	signed long(4)	signed long(4)	char(8)	char(2)	variable size

<---------헤더(21bytes)---------->

성별부분의 영상 데이타는 Block_Id의 값이 8이고, 문자 영상의 표기상태가 한글, 한자, 영문자인 경우로 구분하여 작성한다.

(5) 서명

No	Block_Id	Witdth	Height	Raster Image
int(2)	int(2)	signed long(4)	signed long(4)	variable size

<-----헤더(12bytes)-------->

서명 부분의 영상데이터는 Block_Id의 값이 14인 특수문자 영상 데이터이다. 데이터의 내용은 헤더에 저장시키지 않았다.

(6) INDEX

이 디렉토리는 실제의 영상 데이터가 저장되어 있는 디렉토리가 아니고 영상 데이터를 연결해서 화면에 보여주기 위해서 실제 영상 데이터가 저장되어 있는 위치정보를 가지고 있다.

3. 오프라인 필기체 문자 영상 데이터 구현

3.1 환경

본 논문의 무제약 필기체 문자 영상은 256 gray영상을 획득하여 C언어를 이용해서 Raster Image부분만 가져와서 앞 절에서 설계된 파일 형태에 따라 새로운 헤더 정보를 영상 데이터 앞부분에 입력시킴으로써 새로운 파일을 생성하여 각각의 해당 디렉토리에 저장시켜주는 필기체 문자 영상 데이터 저장 시스템이다.

256 Gray영상을 양자화 함수를 적용하여 원하는 단계의 영상을 획득할 수 있다. 여기서는 2단계 양자화 함수를 적용하여 흑, 백의 이미지로 영상을 화면에 보여 주었다. 이때 이미지 변환의 경계 값이 되는 임계값은 230(0.9)을 사용하였다. 즉, 픽셀의 회색 영역 값이 230보다 크면 흰색으로 변환되고, 회색 영역 값이 230보다 작거나 같으면 검은색으로 변환되어 표현된다. 그림 2는 헤더 부분에 대하여 획득한 영상에 대한 정보를 입력하여 주는 과정이다.

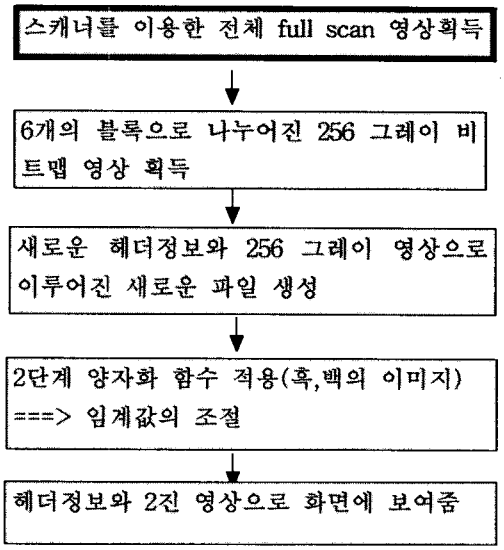

Fig. 2. Algorithm of DB

그림 3는 그림 2의 과정을 통해서 필기체 문자 영상 데이터 베이스 구현 결과를 나타낸다.

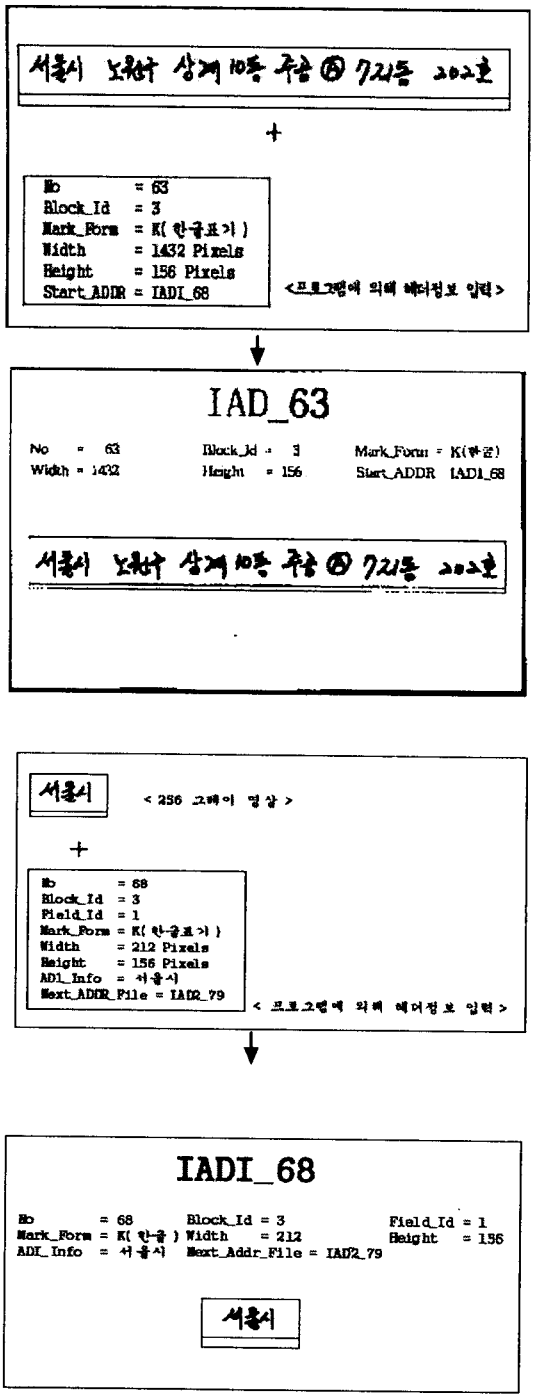

Fig. 3 Address scan result and Display

주소부분의 full scan 영상과 시, 구등의 주소로 나뉘어진 부분을 나타낸다.

4. 결론

본 논문에서 구축된 오프라인 필기체 문자 영상 데이터가 구축되기까지의 배경은 제약이 가해지지 않은 무 제약의 필기체 문자인식을 위한 알고리즘 개발이 아직까지 이루어지지 않고 있으므로 앞으로의 문자인식 연구는 무 제약의 필기체 문자를 인식하기 위한 알고리즘 개발에 주력할 것이 예상되었기 때문이다. 개발된 알고리즘이 얼마나 효율적인지 인식시스템의 인식률을 측정해서 성능을 객관적으로 인정받기 위해서는 실험에 사용될 표준화된 영상 데이터의 제공이 필수적이다.

본 논문에서 대상으로 이루어진 영상 데이터는 오프라인 필기체 문자 영상 데이터로서 일반적인 영상 데이터를 저장시키는 대용량 멀티미디어 저장구조와는 다르게 다루어져야만 한다. 그러므로 서식 문서에 필기된 자료를 스캐너로 읽어들여서 각각의 목적에 맞게 설계된 방식에 따라 영상 파일을 저장시켜 나갔다. 또한 본 논문에서 구축된 영상 데이터 베이스의 대상 문자 데이터는 한글뿐만 아니라 영문자, 한자, 숫자, 특수 문자 등 다양한 종류의 문자를 대상으로 한 오프라인 필기체 문자 영상 데이터 베이스이다. 이렇게 다양한 종류의 문자들로 구성된 오프라인 필기체 문자들을 인식할 수 있는 알고리즘 개발 시에 표준화된 영상 데이터로서의 역할을 충분히 해낼 수 있으리라 기대된다.

참고문헌

[1] 이성환, "문자인식", 홍릉과학 출판사, 1993.

[2] 김상욱, "멀티미디어 데이터베이스 시스템에서의 대형 멀티미디어 객체 관리 기법", 정보과학회지, 제14권 9호, pp. 31-41, 1996.

[3] E. Cohen, "Understanding Handwritten Text in a Structured Environment: Determining ZIP Codes from Address", In Character & Handwriting Recognition, pp.221-264, 1991.

[4] Billris, A., "An Efficient Database Storage Structure for Large Dynamic Objects," Proc. IEEE Inter. Conf. on Data Engineering, pp. 301-308, 1992.