DB 자동 구축을 위한 필기 형식문서 인식 시스템의 개발

김동준*, 조성정*, 류성호*, 이택헌*, 김진형*
*한국과학기술원 전자전산학과
e-mail : {djkim,sjcho,shryu,three,jkim}@ai.kaist.ac.kr

Development of Handwritten Form Recognition System for Automated Database Construction

Dong-Jun Kim*, Sung-Jung Cho*, Sung-Ho Ryu*, Taik-Heon Rhee*, Jin-Hyung Kim*

*Dept. of EECS, KAIST

요 약

형식문서는 현재 정보의 체계화된 표현 및 저장 수단으로서 널리 사용되어 왔다. 최근 이러한 형식문서들을 데이터베이스화해주는 시스템들이 보급되고 있다. 그러나 대부분 외국의 시스템을 기반으로 작성되어 한글, 영어, 숫자, 한자등 다양한 필기 문자들이 사용되는 국내 환경의 특수성을 적절히 반영하지 못하고 있다. 그 결과, 대부분의 경우 아직도 사람이 직접 자료를 입력해야만 한다. 본 논문에서는 이러한 국내 실정에 맞게 다양한 언어의 필기 문자 인식기를 결합하여 형식 문서의 정보를 자동으로 데이터베이스에 입력해 주는 시스템을 제안한다. 제안된 시스템은 영상을 인식한 뒤 그 결과를 검증하는 방법을 통하여 정보의 입력을 보다 효율적으로 수행할 수 있을 뿐 아니라, 전체 작업을 단계별로 분할하여 병렬적으로 수행할 수 있게 함으로써 처리율을 향상시킬 수 있게 하였다.

1. 서론

형식문서, 즉 각종 신청서, 보험청약서, 기업의 인사카드, 각급 학교의 학생카드, 은행의 전표 등과 같이 문서의 형식이 정형화된 문서는 사회 각 분야에서 널리 사용되고 있다. 이러한 필기 형식문서들은 그 내용을 수시로 참조해 보아야 하는 경우가 대부분이며, 원본은 일정기간 반드시 보관되어야 한다.

수년 전까지만 해도 이러한 필기 형식문서는 원본이 캐비닛 등에 보관, 관리되었으며, 내용 검색이 필요한 경우 사람이 직접 해당 문서를 찾아내어 확인해야 했다. 그러나, 종이를 저장하는 방식은 넓은 물리적 공간을 요구하며, 수작업에 의한 자료의 관리는 업무의 신속성, 효율성, 신뢰성에 문제를 드러내고 있어 기업 경쟁력 강화 측면에서 이들의 개선이 요구되고 있다.

이를 해결하기 위하여 최근 필기 형식문서의 영상을 데이터베이스화하여 저장하는 시스템이 개발되고 있다. 그러나, 아직도 형식문서에 포함된 정보는 사람이 직접 수작업으로 컴퓨터에 입력하고 있는 실정이다. 형식문서에 기입된 정보는 대부분 필기문자로 이루어져 있으나 기존에 공급된 제품들은 대부분 외국의 제품을 이식한 것에 불과하여, 불완전한 솔루션밖에 제공하지 못하고 있기 때문이다. 이는 한글, 영어, 한자, 숫자가 공존하는 국내 환경의 특성을 적절히 반영한 인식 솔루션을 제공하지 못하고 있다는 점과, 국내 필기 형식 문서 인식 시스템의 기술적 수준이 상대적으로 낙후한데 기인하고 있다.

본 논문에서는 다양한 언어의 필기 문자 인식기를 결합한 필기 형식 문서 처리 시스템을 개발함으로써, 문서 정보의 입력 및 검증과정을 자동화하여 보다 용이한 형식문서 정보의 데이터베이스화 작업을 가능하게 한다.

2. 형식문서 처리 작업 개요

필기 형식문서 처리시스템의 목적은 각종 신청서와 같은 필기 형식문서를 인식하여 자동으로 데이터베이스를 구축하는 것이다. 즉, 그림 1 에 나타난 바와 같이 기존의 수작업에 의존한 형식문서 처리 과정을 자동화하는 것을 목적으로 한다.

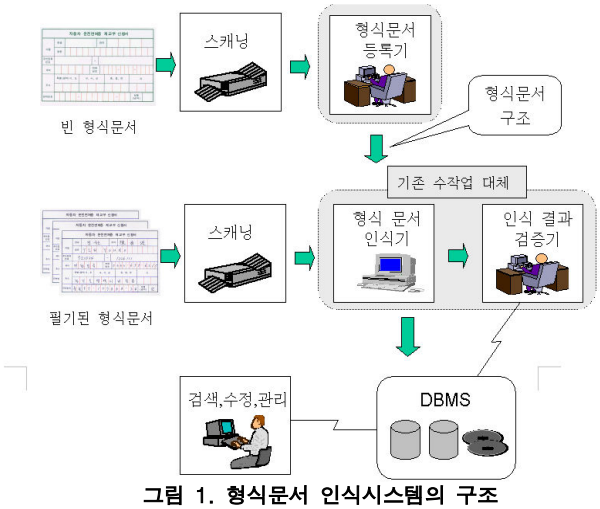

그림 1. 형식문서 인식시스템의 구조

이러한 형식문서의 처리 작업은 크게 형식문서의 등록,문서 영상의 인식, 인식 결과의 검증이라는 세가지 단계로 구분할 수 있다. 형식 문서의 등록 단계에서는 형식문서의 구조와 각각의 필드들이 나타내는 정보들을 입력한다. 문서 영상의 인식 단계에서는 등록된 형식문서의 정보를 바탕으로 사용자가 작성한 문서 영상을 일괄적으로 스캔하여 그 내용을 인식, 저장한다. 이렇게 저장된 인식 결과는 그 신뢰도를 바탕으로 사람이 원래 영상과 대조해가며 검증한 뒤, 적절한 변환을 거쳐 데이터베이스에 저장되게 된다.

이 중, 문서 영상의 인식과 검증 단계에서 전체작업 수행의 대부분의 시간이 소모되므로, 각 단계를 분리하여 병렬적인 처리를 가능하게 할 경우, 시스템의 전체적인 처리율을 향상시킬 수 있다.

3. 필기 형식문서 인식 시스템

본 논문에서 구현된 시스템에서는 형식 문서의 등록, 문서의 인식, 인식 결과의 검증과정을 독립된 시스템으로 구현하였다.

3.1 형식 문서 등록 시스템

형식문서 등록시스템은 인식을 수행하기 위해 필요한 정보들을 등록하는 프로그램이다. 여기에서 설정하는 정보는 다음과 같다.

1) 인식해야 할 필드의 위치 및 구조 지정
2) 필드의 언어 종류 및 후처리 설정
3) 필드 간의 상호 관계 표시
4) 형식문서에서의 필드와 데이터베이스간의 대응관계 표시

각 문서의 모든 위치 정보는 앵커(anchor)라고 불리우는 특수한 모양의 기준점을 바탕으로 기록된다. 동일한 형식문서를 바탕으로 작성된 문서라도 스캐너를 통해서 입력 받을 때에 다양한 종류의 변이가 발생하게 된다. 이를 해결하기 위하여 인식기에서는 문서 영상에서 앵커들을 우선 찾은 뒤 이들의 위치를 바탕으로 왜곡된 영상을 보정하는 방법을 사용하고 있다.

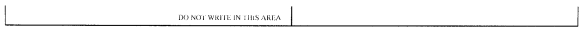

그림 2. 앵커 포인트

각각의 필드들에는 사용하는 한글, 영어, 한자등 사용되는 언어와 각 필드가 표현하는 내용에 관한 정보를 기록된다. 성별, 주소등과 같이 제한된 단어들만이 사용되는 필드의 경우 인식 결과로 나타날 수 있는 후보들의 사전을 미리 구축해 둠으로써 인식기에서 나타난 오류를 보정할 수 있다.

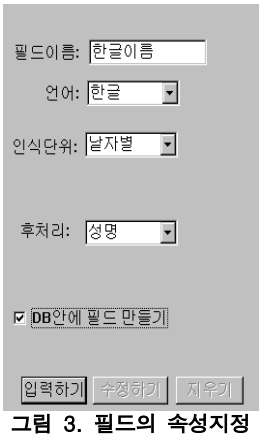

그림 3. 필드의 속성지정

각각의 필드들의 상호 의존 관계 역시 기록된다. 예를 들면, 이름의 경우 한글, 한자, 영문 세 가지로 표현한 이름 모두 동일한 발음을 나타낸다. 주소의 경우, 동/읍에 관한 정보는 이전에 나타난 시/도의 인식 결과에 의하여 결정된다. 이러한 정보들은 각 필드의 인식 결과를 상호 검증하여 인식기에서 발생한 오류를 보정하는데 사용될 수 있다.

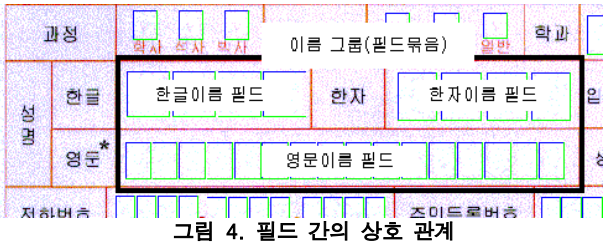

그림 4. 필드 간의 상호 관계

또한, 각각의 필드들과 데이터베이스와의 대응관계도 기록된다. 이후 여기에서 명시된 대응관계를 바탕으로 검증이 끝난 인식 결과들이 데이터베이스에 저장되게 된다.

3.2 형식 문서 영상 인식 시스템

형식문서 인식시스템은 등록된 형식문서의 구조를 참조하여 필기 문서 영상을 분석하여 필드를 추출하고 그 필드에 알맞은 인식기들을 호출하여 인식을 수행한다.

형식문서 처리는 한 번에 일괄적으로 스캔된 영상들을 단위로 이루어진다. 본 시스템에서는 이를 '작업'이라고 하여, 영상의 인식과 검증 작업을 수행하는 기본 단위로서 사용한다. 각각의 작업은 독특한 ID 와 이름을 가지며, 각자 고유의 디렉토리에 원본 영상을 저장한다.

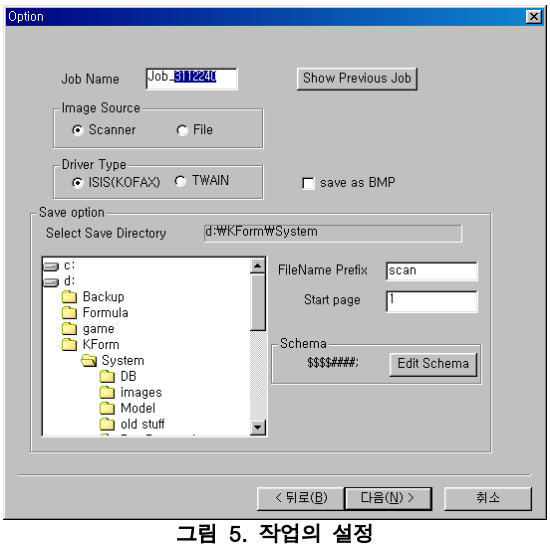

그림 5. 작업의 설정

문서 영상이 입력되면 이진화 과정을 거쳐 각각의 필드들의 위치를 찾는다. 각 필드를 구성하는 선분들은 스캐닝이나 이진화 과정을 거치면서 끊어지는 경우가 빈번히 발생하기 때문에, 전체 문서 영상을 blurring 한 뒤 Run-Length Smearing Algorithm 을 적용하여 직선 선분을 뚜렷이한 뒤 필드를 추출한다.

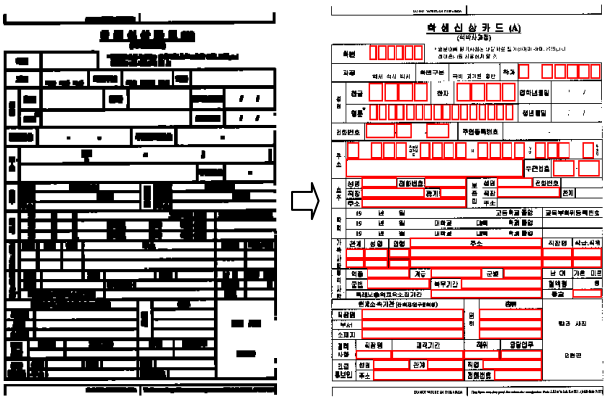

그림 6. 필드의 추출

필드의 위치가 파악되면, 각 필드의 언어 정보를 바탕으로 필드 별로 해당 인식기를 호출, 인식을 수행한다. 인식이 이루어진 뒤, 후보 사전이나 상호 참조 정보와 같은 별도의 정보가 존재할 경우 이들을 사용하여 후처리를 수행한다.

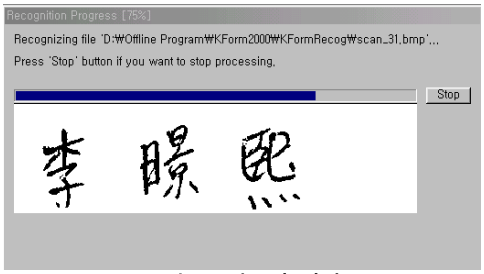

그림 7. 필드의 인식

인식된 결과는 아직 인식 오류가 존재할 수 있으므로, 일단 임시 저장공간에 저장된다.

3.3 인식 결과의 검증 시스템

인식 결과 검증 시스템은 인식기에서 얻어진 결과를 원본 영상과 대조해 가며 검증할 수 있게 도와주는 시스템이다. 이 단계는 상대적으로 사람의 참여가 많은 부분이며, 따라서 최종 입력자료의 질에 가장 큰 영향을 미치는 부분이다. 실제 작업에서도 검증기의 효율적인 디자인이 시스템의 전체적인 생산성에 가장 큰 영향을 미친다고 알려져 있다.

각 검증 작업은 스캔 시에 설정된 작업을 기준으로 실행된다. 따라서, 서로 다른 작업과정을 통하여 스캔된 영상들은 각자 다른 시스템에서 동시에 검증작업을 수행할 수 있다.

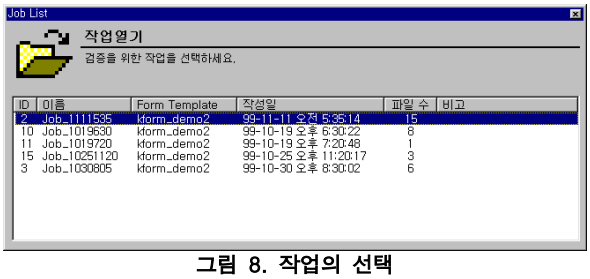

그림 8. 작업의 선택

본 시스템에서 사용된 검증기는 필드의 인식 결과를 선택하면 원 영상에서의 해당 필드를 자동으로 표시 해 준다. 각 필드들은 영상 단위로 정렬해서 보거나 [그림 9], 동일한 필드별로 정렬해서 보는[그림 10] 두 가지 종류의 방식을 통하여 검증할 수 있다.

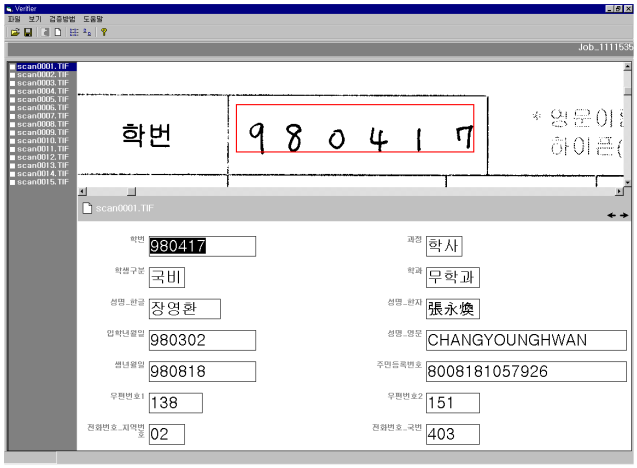

그림 9. 영상단위 검증

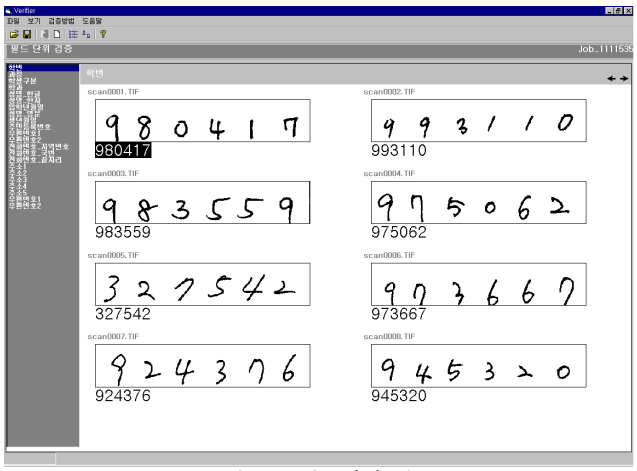

그림 10. 필드단위 검증

검증이 끝난 인식 결과는 기존에 명시된 대응관계에 따라 데이터베이스로 저장되게 된다.

4. 결론

사회가 정보화되어감에 따라 방대한 양의 자료를 컴퓨터에 입력하여 전산처리를 해야 하는 필요가 증가하고 있다. 현재까지 자료 입력의 대부분은 수작업을 통해 이루어지고 있거나 자료처리를 수작업에 의존하고 있어서 많은 인력과 시간이 소요되고 있다. 그러나, 지속적으로 증가해 가고 있는 데이터들을 효과적으로 처리하기 위해서는 자료의 입력과정의 자동화가 요구되고 있다. 형식문서 자동입력 시스템은 형식이 정해져 있는 문서에 사람이 필기한 각 필드를 분석, 인식하여 자동으로 DB 에 저장하는 시스템으로써 관공서, 기업, 학교 등에서 많이 사용되는 형식문서를 자동으로 입력하기 위한 시스템이다.

형식문서 자동 입력 시스템을 위해서는 문서영상 분석기술, 필드들을 인식할 수 있는 한글, 영어, 숫자, 한자 등 여러 언어에 대한 인식기술, 인식오류의 자동 보정을 위한 후처리기술, 그리고, 사용자가 편리하게 사용하기 위한 인터페이스 등이 필요하다.

본 논문에서 구현된 시스템은 형식문서 등록기, 형식문서 인식기 및 인식결과 검증기로 구성되어 있다. 사용자는 형식문서 등록기를 이용하여 형식문서의 구조를 등록한다. 형식문서 인식기는 대량의 필기된 형식문서를 연속으로 스캔한 후 등록된 형식문서의 정보를 이용하여 영상을 분석한다. 영상에서 필드를 추출하고 그에 알맞은 인식기를 이용하여 인식한 후 상호 의존 관계 및 후보 사전등을 이용하여 후처리 작업을 수행한다. 이렇게 인식기를 통해 인식된 정보에는 인식오류가 포함될 수 있으므로, 인식 결과를 검증기를 통해 원본과 비교해 가며 사람이 검증한다. 검증된 인식결과는 데이터베이스로 저장하여 이후 다양한 작업에 사용할 수 있게 한다.

참고문헌

[1] I.J..Kim, J.Kim "Stroke-guided Pixel Matching for Handwritten Chinese Character Recognition," Fifth International Conference on Document Analysis and Recognition, India, 1999, P.665 – 668

[2] C.E. Cheong, H.Y. Kim, J.W. Suh and J. Kim, "Handwritten Numeral String Recognition with Stroke Grouping," Fifth International Conference on Document Analysis and Recognition, India, 1999, P. 745 – 748

[3] C.L. Liu, I.J. Kim, and J. Kim "Model-Based Stroke Extraction and Matching by Heuristic Search for Handwritten Chinese Character Recognition," Proc. The 6th Int'l Workshop on Frontiers in Handwriting Recognition, Taejon, Korea, Aug. 12-14, 1998, P.547 – 556

[4] K. W. Kang, J. W. Suh and J. Kim "Skeletonization Of Grayscale Character Images Using Pixel Superiority Index " Proc. IAPR Workshop on Document Analysis Systems, Nagano,Japan, 1998

[5] C.L. Liu, I.J. Kim and J. Kim "High Accuracy Handwritten Chinese Character Recognition by Improved Feature Matching Method," Proceedings of the Fourth Int'l Conference on Document Analysis and Recognition, Ulm, Germany, August, 1997, Vol. 1, P. 1033 – 1037

[6] S. Mori, K.Yamamoto and M.Yasuda, “Research on machine recognition of handprinted characters”, IEEE Trans. PAMI, vol 6. no.4, pp.386-405, 1984

[7] J.Rocha and T. Pavlidis, “A shape analysis model with applications to a character recognition system”, IEEE Trans. PAMI, vol.16, no.4, pp.393-404,1994

[8] A.P. Dempster, N.M. Laird and D.B.Rubin, “Maximum likelihood from incomplete data via the EM algorithm”, J.Royal Stat. Soc. vol.39, pp.1-38, 1998