문자 인식 기술의 동향 연구

김영은* · 조범준**
조선대학교 컴퓨터공학과

A Study on Trend of Character Recognition Technological

Young-Un Kim* · Beom-joon Cho**

요 약

본 논문에서는 문자인식기술에 관련하여 최근 발표된 연구 사례를 종합하여 국내외 문자인식 기술의 현황을 알아보고, 향후 연구자들이 연구 방향을 설정하는데 도움이 되도록 함은 물론, 본 논문 분야 연구의 발전 방향을 모색하고자 한다. 사례 조사의 범위는 국내외 문자 인식 연구에 관련된 각 연구 기관과 기업체에 기술 수준에 대한 전반적인 동향으로 하였고, 이를 토대로 앞으로의 연구 방향을 제시하도록 한다.

ABSTRACT

This paper presents the trend of character recognition technology through uniting recently announced researches and also this paper can help researches to set their research direction. Range of the investigation is limited to general tendency of character recognition technology of research institution and business, and this paper presents forward research direction.

키워드
신경망, 문자인식, 문자패턴, 문자영상

Ⅰ. 서 론

문자는 음성과 함께 인간 상호간에 언어로 표현되는 정보 전달하기 위해 발명한 대표적인 부호체계[1]로써 정보 전달을 위한 서로간의 의사 결정 수단이다. 따라서, 문자가 탄생한 이래 인간의 문명은 급속한 고도의 성장을 가져왔고, 초창기에는 필기에 의한 문자 표현이 자주 사용되었으나 인쇄 기술의 발달에 따라 인쇄 문자로 정보를 교환하는 것이 일반화 되었다.

대부분의 사람들은 정보를 얻는 수단으로 눈을 통해서 사물을 보거나 귀로 소리를 들어 입력되는 패턴(영상 또는 음성)을 인식하고 그것을 지식으로 축적하는 연속적인 과정을 거친다[2].

패턴 인식의 핵심 연구 분야인 문자인식 기술은 시각 정보를 통하여 문자를 인식하고 나아가 의미를 이해하는 사람의 능력을 컴퓨터로 실현하려는 시도로서 현재까지 활발한 연구가 진행되고 있다.

그 결과 최근에는 고도 정보화 사회로의 여건 조성에 따른 산업 발전과 기술의 대형화, 고도화 등으로 매년 방대한 양의 정보가 처리되고 있다. 정보화 사회를 이루기 위해서는 대부분 종이로 기록되어 전해 오던 모든 정보를 컴퓨터에 저장하여, 이를 필요로 하는 사람이 적시적소에 사용할 수 있어야 하며, 사무자동화와 함께 보급된 개인용 컴퓨터의 빠른 확산으로 인하여 많은 양의 정보를 단시간에 처리할 수 있는 고도의 기술을 필요로 한다. 따라서 종이에 기록된 문자 데이터를 보다 효과적으로 컴퓨터에 입력하여야 한다.

현재 문자 인식 기술은 우편물 자동 분류를 위한 우편번호 인식, 산업 현장에서의 제품 검사나 분류, 문서 인식, 도면 인식, 팩스를 통해 전달 받은 영상에서의 문자인식, 워드프로세서 OCR, 금융기관에서의 전표 또는 수표의 자동입력 등 여러 분야에 걸쳐 실용화 되어 실생활에 효과적으로 사용되고 있다[3].

또한 문자인식기술은 이미 해외에서 신경망(Neural Network)을 이용한 영문자, 숫자, 일본의 가나문자 등에 대한 연구결과를 발표하고 있으며, 국내에서도 1970년대부터 인하대와 충남대에서 한글 문자인식을 중심으로 연구를 시작하여, 1980년대는 충남대, 광운공대, KIST, KAIST, 포항공대, 삼성, LG 등에서 집중적으로 연구 되었으며, 1990년

대는 영문서는 옛 서류, 문서 등의 글씨를 인식하여 웹문서로 자동 제작하고 있으며, 우리나라에서도 수준은 미흡하나 한글 문서를 인식하는 시스템들이 상용화하여 사용하고 있다.

따라서 본고에서는 컴퓨터에서 한글처리를 위한 문자인식 기술의 최근 국내·외 문자인식 기술의 동향과 향후 연구 및 개발 방향을 제시하도록 한다.

II. 문자 인식 기술의 개요

문자 인식은 패턴 인식 기술의 핵심이며, 지난 40년간의 연구 결과에 힘입어 문자인식을 이용한 업무 자동화는 사무실, 공장, 가정 등 우리의 전 생활공간에서 널리 이용되고 있다.

다음은 문자 인식 체계로써 [그림 1]과 같이 분류할 수 있다.

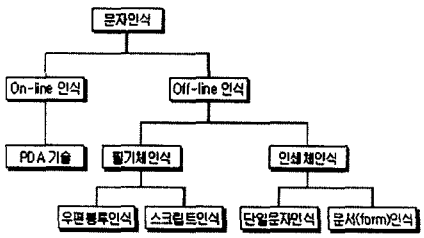

그림 1. 문자인식체계

문자 인식기술은 입력방법에 따라 온라인 인식(On-line recognition)과 오프라인 인식(Off-line recognition)으로 크게 둘로 나눈다.

온라인 인식(On-line recognition)은 필기자가 전자펜을 이용하여 필기하는 과정을 실시간으로 입력받아 인식하는 방법으로 입력 패턴의 시간적 정보, 위치상의 공간적 정보, 압력 정보 등을 분석하여 인식한다. 반면에 오프라인 인식(Off-line recognition)은 필기자가 노트위에 작성해 놓은 필기체 문서나 책 등의 인쇄된 문서들 스캐너 등 영상 입력 장치로 복사하듯 입력받아 인식하는 방법으로 입력패턴의 위치상 공간적 정보를 분석하여 인식 하는 기술이다.

오프라인 인식은 다시 필기체 인식과 인쇄체 인식으로 분류된다. 필기체 인식과 인쇄체 인식을 통틀어 OCR(Optical Character Recognition)기술이라고 부른다. 필기체 인식의 응용 분야로는 우편 분류 자동화와 스크립트 인식을 들 수 있다. 우편 분류 자동화 기술은 다양한 형태의 사람의 필체를 인식해야 하기 때문에 가장 힘든 기술이라고 여겨지고 있다.

그러나, 영어와 같이 일차원적인 문자의 조합으로 이루어진 문자들에 대해서는 어느 정도 인식 기술이 올라와 있는 상태이고 또한 우편 분류 자동화를 미국과 독일 등 부분적으로 사용되고 있다.

반면, 한글과 같이 동양권 문자들은 단순히 알파벳만을 사용하는 것만이 아니라 알파벳을 2차원으로 조합한 글자를 사용하고 있으므로 글자의 수효가 11,172자에 이르며 분류해야 할 클래스의 종류도 엄청나게 많은 실정이기 때문에 현실화 단계에는 아직 미치지 못하고 있다.

인쇄체 인식은 문서 인식(Form processing)에 많은 진보를 이루게 하였다. 따라서 디지털 도서관과 같이 일일이 스캔을 받아 큰 용량의 그림 파일로 서비스하는 대신에 문서 인식을 통하여 적은 용량의 텍스트(text) 파일로 전송해 줄 수 있기 때문에 용량 면에서나 전송 시간 면에서나 큰 역할을 감당할 것으로 기대 되어 진다. 이러한 인쇄체 인식은 상품화가 나와 있을 정도로 높은 인식률을 가지고 있지만 인쇄한 문서의 스캔 정도와 다양한 폰트에 따라 인식률이 차이가 나는 것이 사실이다. 따라서 각 폰트에 따라 학습시키는 과정은 중요한 이슈(issue)로 대두 되고 있다.

문서분석은 크게 하향식 접근(top-down approach) 방식과 상향식 접근(bottom-up approach) 방식으로 나눌 수 있다.

하향식 접근 방식은 지식 기반 형으로 특정한 형식을 갖는 문서에 대해 문서의 계층적 구조에 맞게 문서를 분할해 가는 것으로 처리 속도가 빠른 장점이 있다. 상향식 접근 방식은 기본적인 요소, 예를 들면 연결요소(connected component) 혹은 단일 문자로부터 출발하여 기본적 요소들을 그룹화 하여 전체적인 문서 구조를 만들어 가는 것으로 많은 처리시간이 요구되나 다양한 문서에 적용할 수 있는 장점이 있다.

오프라인 문자 인식의 기본적인 과정을 살펴보면 Scanner나 Camera를 사용하여 입력 영상을 만들고 주어진 영상에 대해 전처리 과정을 거치게 된다. 전처리는 노이즈를 제거하고 인식에 필요한 정보들을 추출하게 된다. 전처리가 끝난 영상은 문자열을 추출하고 문자를 분리해 내는 세그멘테이션 과정을 거치게 되는데, 높은 인식률을 얻기 위해서는 세그멘테이션의 신뢰성이 높은 분할이 필수적이다. 인쇄체 문자 인식의 경우는 글자들의 크기가 균일하고 글자 간에 겹침이 별로 없기 때문에 세그멘테이션에서 신뢰도가 높은 분할을 할 수 있으나 필기체 문자 인식의 경우 글자들의 크기에 변화가 심하고 글자간의 연결이 많기 때문에 신뢰도가 높은 분할을 하기가 쉽지 않다. 따라서 특정한 폼의 문서를 분석하는데 사용되는 하향식접근 방식이 적합하다. 그러므로 전체적으로는 먼저 문자열을 나타내는 선을 분리한 후 각 문자열로부터 각 문자를 분리하여 인식기에 활용하는 방법을 채택하는 것이 필요하다.

이와 같은 방법으로 가장 적합한 방식이 투영(projection)에 의한 방법이다. 이 방법은 우선 간단하여 연결요소 방식보다 속도가 빠르다. 이러한 Off-line 인쇄체 인식은 [그림 2]와 같은 과정으로 수행된다.

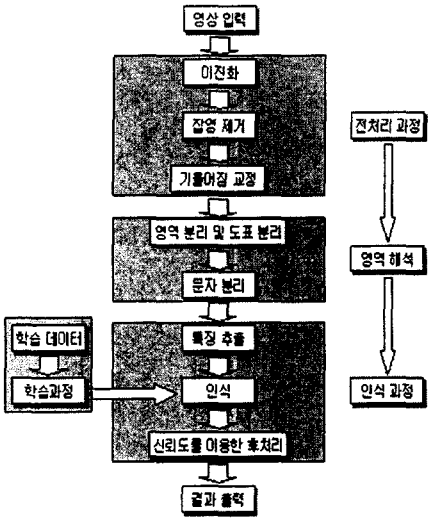

그림 2. 문자 인식 구성도

즉, 이진화 및 전처리 과정, 영상의 영역을 분리하고 해석하는 과정, 문자의 특징을 추출하는 과정, 분할된 세그먼트를 이용한 학습과정, 동적 프로그래밍 기법을 이용한 인식 정합과정 마지막으로 인식 정확도에 기반 한 후처리과정을 거치게 됨으로써 인식 결과를 도출하게 된다[4].

III. 국내외 문자 인식 기술 동향

1. 국외 경우

문자패턴 인식 연구는 [표 1]과 같이 세계의 여러 유명기관에서 활발히 진행되고 있으며 연구팀이 없는 대학이 없을 정도이다. 20여년전부터 독일, 일본, 미국, 프랑스, 이태리 등이 우편물 분류시장에서 경쟁하고 있으며 이제는 세계 곳곳의 우체국에 편지 봉투 분류기가 배치 되어있다. 처음에는 단순한 인쇄된 우편 번호만을 인식했지만 기술이 진전됨에 따라 필기 숫자는 물론 이제는 필기한 거리 이름도 인식한다.

현재 미국의 우체국의 경우에는 규격 봉투 우편물의 55%를 자동 처리하는 수준에 이르렀고 대봉투, 잡지, 신문, 소포의 자동처리도 손익분기점을 넘어서고 있다. 미국 SUNY Buffalo대학에 설치된 CEDAR는 미국우정성의 연구소 역할을 하고 있으며, University of Washington, 독일의 국가 인공지능센터인 DFKI, 프랑스의 출연 기관인 INRIA 등이 활발한 연구와 인재를 양성하고 있다. 일본은 NEC, Toshiba 등 대기업들이 커다란 문자인식 팀을 유지하고 있다. 미국의 정부 기관인 NIST는 인식 연구에 필요한 방대한 필기 문자 및 숫자 데이터베이스를 구축 관리하고 있으며, Washington 대학에서는 대량의 문서 영상 데이터베이스를 일본의 출연 기관인 ETL 및 대만의 ITRI는한자 데이터베이스를 구축하여 연구 지원하고 있다[5].

표 1. 국외 관련 업체

지역	기관명	사업 및 연구내용	비고
독일	Siemens	우편물 분류기 시장 선두주자: 시간당 36,000장 우편물 처리기 생산: 년 DM 2.5 Billion의 매출에 년6%의 성장률, 실청서 자동처리 등의 형식문서 인식분야로 확장 중	4500명
일본	NEC	동양권 우편분류기 시장 석권, 자동경시시스템(환경, 교통, 공장자동화) 분야 사업	500명
미국	CEDAR, SUNY at Buffalo	우정국의 전폭적 지원을 받는 미국 최대의 우정관련 연구센터: 우편봉투 자동처리 시스템, 필기인식 등을 연구	300명
미국	National Institute of Science and Technology	정부/산업계에 패턴인식 기술을 제공하기 위하여 설립: 영상인식기술 개발, 평가, 산업계 전수가 목표: 얼굴인식, 지문인식, 문서처리 시스템 등을 연구: 문자 DB 제공	200명
캐나다	CENPARMI, Concordia University	1988년 정부지원으로 설립된 패턴인식 전문 연구기관: 패턴인식/인공지능 분야의 인력 양성, 산학협동이 목표: 문서분석, 수표 자동처리, 숫자/단어/문자열 인식	400명
일본	Electrotechnical Laboratory	자동 영상통로 및 인맥상, 영상 매장 기술을 연구: 표준문자 DB를 구축 제공	400명
독일	German Research Center for Artificial Intelligence GmbH	우편물 처리, 문서분석기술 등을 연구개발: 고성능 OCR 시스템 개발, 문서구조분석, 구조화된 문서의 인맥상, 분류, 저장 기술의 개발	300명

2. 국내 경우

국내에서 개발된 대표적인 문자 인식 상용화 시스템은 "글눈(글을 읽는 눈을 뜻함)"이다.

글눈은 세계 최초의 다국어 인식기로 한글(11,172자), 한자(1만 6천여 자), 영어, 일어, 독어, 불어 등 14개국 문자를 인식할 수 있으며, 영어 99% 이상, 한글 98% 이상, 한자 90% 이상, 일어 95% 이상의 인식률로 초당 300자씩 고속으로 문자를 인식하며, 세계 유일을 학습기능을 제공하고 있다[6].

또한 이 제품은 2백 쪽 분량의 책자 내용을 입력하는데 10분 정도 시간이 걸리고 영상 편집기가 내장돼 있어 입력된 내용을 편집할 수 있을 뿐만 아니라 입력된 문서를 아래아 한글, 훈민정음 등의 워드프로세스 파일로 옮겨 작업 할 수도 있다.

표 2. 국내 관련 업체

산업명	제품 또는 공정명	관련 산업체 및 공공기관명
공공관련	우편물 자동처리 시스템	전자통신연구원 우정기술연구부
패턴인식, 문서처리 일반	문서처리 시스템, 핸컴퓨터	LG종합기술원 정보기술연구소
패턴인식, 문서처리 일반	문서처리 시스템, 핸컴퓨터	삼성전자
공공부문 인식산업	수납장표 자동처리 시스템	쌍용정보통신
공공부문 인식산업	수납장표 자동처리 시스템	효성데이터시스템
영상인식 기초연구	얼굴/제스쳐 인식	고려대학교 전자계산학과
문자인식 기초연구	카드슬립 자동처리	연세대학교 전산학과
문서인식 산학협력	핸컴퓨터용 인식기, 한글 필개인식, 형식문서 인식	KAIST 인공지능연구실

한편 국내 연구진의 해외 교류도 활발해짐에 따라 [표 2]와 같이 국내에서도 전자통신연구원, 삼성전자, LG전자 등 많은 연구소와 대기업이 10여 많은 연구소와 대그룹이 연구팀 운영으로 국내에 학술 대회 유치와 각종 논문 발표로 세계적인 문자인식 기술력을 키우기 위해 경쟁력을 키우기 위해

노력하고 있다. 하지만 아쉽게도 이 분야에 대한 연구 투자가 조직적이지 못하고 그 규모도 크지 않았기 때문에 경제적 이익을 동반하는 대형사업에 나설 수가 없었고 따라서 실용적인 연구는 선진국에 비하여 미흡한 실정이다.

우편 봉투의 주소를 인식하여 자동 분류하는 우편자동화 사업의 경우, 정보통신부의 자료에 따르면 연간 1,500억원의 경비 절감효과로 2003년에는 22개 우편집이 중국 건설에 1조 7000억원이 투자되었다. ETRI에서는 우편물 자동 분류 시스템 사업팀을 구성하고 독일 Siemens사와의 기술 협력 및 국내의 관련 연구팀과의 공동연구를 수행하고 있다.

금융기관 수납장표 정보화시장은 수납장표처리를 자동화하여 입력, 보관 및 정보추출을 효율적으로 하기 위한 사업이다. 1999년부터 시작된 4년간의 초기 투자가 4천7백억원 이루어졌으며, 연간 1천1백억원 규모의 시장[7]에 이르고 있다. 금융결제원에서 1999년에 실시한 공개 입찰에는 국내·외의 관련 업체로 구성된 9개의 컨소시엄이 참여하여 경합하기도 하였다.

형식 문서(Form) 자동 입력 시스템은 일정 형식이 있는 문서에 인쇄 혹은 손으로 쓴 내용을 각 항목 별로 추출, 인식하여 데이터베이스화 하는 기술이다. 아직 국내시장은 활성화되지 않았으나 그 시장 규모는 일본에 비추어 볼 때 연간 약 600억원으로 추정 된다. 또한 종이 문서의 전자 문서화 작업의 시장규모가 연간 약 1500억원에 달할 것으로 예상된다[8]. 세계문서기술의 시장의 규모는 1998년 132억 달러(약 15조원), 2003년에는 415억 달러(약 50조원)이며 계속해서 증가할 것으로 기대된다[9]. 형식 문서 자동 입력 시스템은 응용분야가 다양하여, 인식능력의 향상에 따라서 그 시장이 산업계 전반으로 급속하게 확장될 것으로 전망 된다. 본 기술이 확보될 경우 같은 동양 언어권인 중국 및 일본의 시장 진출의 가능성이 높아진다.

마지막으로 온라인 필기 문자 인식 기술이 필요한 PDA, Tablet PC, eBook 시장규모가 빠른 속도로 증가하고 있다. 2000년 추계 컴덱스에서는 Bill Gates가 Tablet PC를 홍보하였으며, IGI 그룹이 밝힌 '무선 웹 시장 전망'이란 보고서에서는 PDA의 경우 2002년 1600만대, 2003년 3500만대가 팔렸고, PDA와 휴대용 전화기가 결합된 스마트 폰은 2002년에 1억 7500만대, 2003년에는 3억 3300만대 이상이 출하 되었다[10].

IV. 결 론

문자 인식 연구는 지난 20세기 초부터 현재까지 타 분야에 비해 비교적 활발히 진행되고 있다. 최근 들어 산업 전반에 걸쳐 응용 사례가 늘고 있으며, 응용 가치 또한 날로 증대하고 있다. 하지만, 문자인식 기술은 해결해야 할 많은 과제를 안고 있는 어려운 기술로 앞으로 지속적인 연구가 필요하다. 따라서 지금까지 연구 경향을 보면 학교, 연구소 등을 중심으로 소규모의 문제를 독립적으로 해결하고자 노력해 왔다. 이러한 기술들이 우리나라의 문서 인식의 근간이 된 것은 사실이나 실용적인 문서인식 시스템을 개발하기 위해서는 보다 체계적인 연구가 필요하다.

문자 인식 기술은 국가적 전략기술이며, 국제적 경합 기술임으로 국가적으로 매우 중요한 기술로서 산업 전반에 걸쳐 폭넓게 활용되기 위해서는 보다 많은 사람들의 관심과 국가적인 지원 속에서 보다 체계적이고, 집중적인 연구가 절실히 필요하다.

참고문헌

[1] 오영환, "패턴 인식론", 정익사, p.63, 1991.
[2] 이성환, "오프라인 필기체 문자 인식 기술의 현황", 정보과학회, 11권 5호 p.51, 1993.
[3] 안창, 이상범, "한글처리-문자 중심 인식 기술 고찰", 정보처리학회, 5권 5호, p.52-53 1998.
[4] http://sheep.kangnam.ac.kr/
[5] http://ai.kaist.ac.kr/~jkim/profile.htm
[6] 이인동, 문자인식기술, 정보처리학회, 6권 4호, pp.11-16, 1994.
[7] "금융 결제원과 한국 은행의 '금융기관수납장표의 정보화 추진'에 관한 보고서"
[8] 전자신문 1997년12월8일, 1998년3월17일자 기사
[9] AIIM International의 Gartner그룹, "The State of the Document Technologies Industries", 1999-2003.
[10] 전자신문 2000년 1월 4일자 기사