한글 온라인 필기 인식을 위한 전처리 모듈 개발

정민진*, 정다빈*, 이강은*, 김성석**, 양순옥***
*서경대학교 소프트웨어학과
**서경대학교 소프트웨어학과, 교신저자
***가천대학교 교양학부
e-mail:{alswls4390, zxzx1227, gang3039}@naver.com, sskim03@skuniv.ac.kr, soyang@gachon.ac.kr

Development of Preprocessing module for Korean online handwriting recognition

Min Jin Jeong*, Dabin Jeong*, Kang Eun Lee*, Sungsuk Kim**, Sun Ok Yang***
*Dept of Software Engineering, Seokyeong University
*Dept of Software Engineering, Seokyeong University, corresponding author
**College of Liberal Arts, Gachon University

요 약

본 논문은 개발하고자 하는 기계학습 기반 한글 필기 인식 시스템의 첫 연구 결과를 담고 있다. 즉, 기계학습을 위해서는 학습용 및 테스트용 필기 데이터가 아주 많이 필요하므로, 이를 수집하고 전처리하는 방법을 제안하였다. 한글의 한 글자는 자음과 모음을 결합하여 생성되는데, 실제 만 개 이상의 글자가 생성될 수 있다. 따라서 각각의 글자 데이터를 수집하는 대신, 수집한 글자 데이터로부터 초성, 중성, 종성을 구분하여 최종적으로 자음, 모음 데이터로 저장하고자 한다. 아직 초기 연구이므로, 다양한 경우에 대한 분석이나 실험 결과는 없지만, 이를 활용하여 온라인 필기 인식 모델에 적용하여 인식 성능을 높이기 위한 추후 연구의 기반으로 활용하고자 한다.

키워드 : 필기 인식, 한글 인식, 기계학습, 전처리

1. 서론

문자 인식은 인공지능 분야의 매우 오래된 연구 주제 중 하나이다. 초창기 연구에서는 주로 출력된 인쇄물의 폰트의 특징을 계산하여 글자를 인식하려고 하였다[1]. 규칙을 기반으로 한 인식은 각 폰트마다 별도의 알고리즘을 개발해야 하며, 다양한 언어가 섞인 경우에는 복잡도가 증가하고 특히 활용 범위가 넓은 필기 인식에는 적절하지 못한 단점이 있었다.

필기 인식은 사람들이 작성한 글자를 컴퓨터가 인식할 수 있는 알고리즘을 개발하는 것이다. 필기 인식은 다시 온라인 필기 인식과 오프라인 필기 인식으로 구분할 수 있다. 오프라인 필기 인식은 작성자가 작성한 필기를 이미지로 저장한 후 인식하는 방식이며, 온라인 필기 인식은 필기 도구를 이용하여 필기 하는 과정에서 추출된 여러 정보를 함께 활용하여 인식하는 것이다. 과거에는 필기를 인식하기 위해 각 글자의 특성이나 필기자의 습관 등을 파악하여 활용하는 방식을 취하였지만, 최근에는 기계학습이나 딥러닝 기술을 활용하여 범용적인 인식 기술을 개발하려는 연구가 많이 진행되고 있다[2].

우리 연구의 최종적인 목적은 온라인/오프라인 필기 인식을 모두 지원할 수 있는 기계학습 기반 시스템 개발이다. 이에 대한 동기는 유아용 한글학습 앱 중에서 받아쓰기 기능을 제공하기 위한 것이었다. 받아쓰기는 학습자의 필기가 문제와 일치하는지를 비교하는 것으로, 일반적인 필기 인식 문제에 비해 난이도가 상대적으로 낮은 응용이다. 하지만, 최근 온라인 학습에 대한 관심이 커지면서, 학습 결과에 대한 주관식 시험 등에 적용 가능하므로 그 응용범위는 적지 않다고 판단된다.

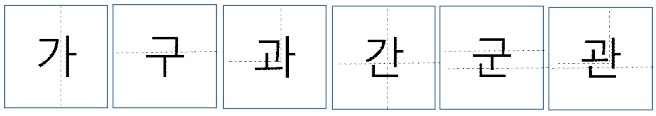
전체 연구 중에서 첫 번째 대상은 온라인 필기 인식을 위한 기계학습 기반 모델 개발을 목적으로 하며, 본 논문은 학습 및 테스트 데이터 획득용 프로그램 개발과 데이터 전처리 과정에 대한 결과를 담고 있다.

2. 온라인 한글 필기 인식을 위한 접근방법

기계학습이란 인공지능 알고리즘 중 하나로서, 사람의 추가 개입없이 시스템이 스스로 학습하고 경험을 통해 그 학습능력을 향상시킬 수 있도록 하는 기술이다[3]. 기계학습은 지도학습/비지도학습/강화학습으로 구분할 수 있으며, 인공신경망의 은닉층의 깊이를 깊게 한 딥러닝이 최근 많은 향상된 연구 성과를 보이고 있어 연구가 활발하게 진행되고 있다. 필기 인식을 위해서 본 논문에서는 지도학습 방식을 채택하였으며, 이를 위해 시스템이 학습할 수 있는 상당한 규모의 필기 데이터를 획득하는 것이 연구의 시작 지점이 된다.

2.1 한글 인식을 위한 접근 방법

한글은 자음 19개(겹자음 포함), 모음 21개(겹모음 포함)를 이용하여 하나의 글자를 구성한다. 하나의 글자에는 초성과 중성, 그리고 선택적으로 종성을 사용하며, 초성과 종성은 자음, 중성은 모음을 사용한다. 따라서 하나의 글자는 모음의 위치에 따라 (그림 1)과 같이 6가지 중 하나로 구분할 수 있다.



(그림 1) 모음의 위치에 따른 한글의 글자 패턴 6가지

그림에서 알 수 있듯이, 각 패턴에 가능한 글자의 개수를 모두 계산하면 이론적으로 초성 19자×중성 21자×종성 28자 = 11,172 글자가 가능해진다. 그런데 학습을 수행하기 위해서는 하나의 글자에 대해 아주 많은 수의 글자 데이터를 필요로 한다. 예를 들면, 0-9까지의 숫자를 인식하기 위한 연구로 유명한 MNIST 데이터베이스에서는, 학습용으로 60,000개 이미지, 테스트용으로 10,000개의 이미지를 확보하여 공개하고 있다[4](그림 2 참고). 이는 숫자 0을 학습하기 위해서 6,000명으로부터 숫자 '0' 에 대한 28×28 픽셀 크기의 필기 데이터가 확보되었음을 의미한다.

(그림 2) MNIST 데이터 집합 샘플

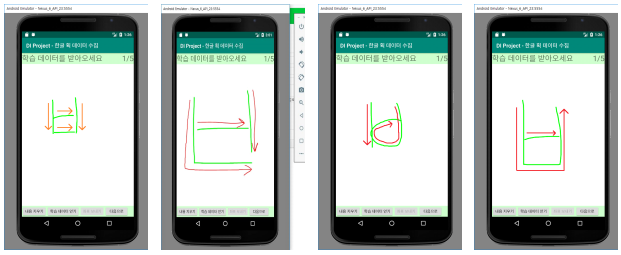
이런 접근방법이라면 11,000개가 넘는 한글의 한 글자당 1,000개 이상의 학습 데이터와 테스트 데이터를 확보해야 함을 의미하는데, 이는 하나의 연구실 차원에서는 접근하기 곤란하다. 대신 본 연구에서는 충분한 필기 데이터로부터 자음과 모음 데이터를 추출하여 학습시키는 접근방법을 채택하기로 한다. 즉, 학습자가 '간'이란 글자를 필기했다면 이 글자로부터 자음 'ㄱ', 'ㄴ', 모음 'ㅏ'을 추출하여 학습 데이터로 사용하고자 한다. 이렇게 할 경우, 일정 개수 이상의 필기 데이터만 확보하면 각각의 자음, 모음 데이터를 추출할 수 있어서, 상대적으로 적은 노력으로 충분한 학습 데이터를 추출할 수 있다.

2.2 글자로부터 초성, 중성, 종성 추출하기

우선, 필기 데이터를 확보하기 위하여 사용자용 앱과 이를 수집, 분석, 보관하기 위한 서버 시스템을 구현하였다. 앱에서 사용자에게 예를 들면 '반'이란 글자를 필기하라고 요구한다. 이때 서버는 '반'이라는 글자에 대하여 사전에 필요한 정보를 분석하여 저장하고 있다. 따라서 서버가 사용자로부터 필기한 획 데이터를 전달받으면, 이러한 정보를 활용하여 필기 데이터로부터 초성, 중성, 종성을 구분한 후 최종적으로 서버에 'ㅂ', 'ㄴ', 'ㅏ'를 각각 저장하도록 한다.

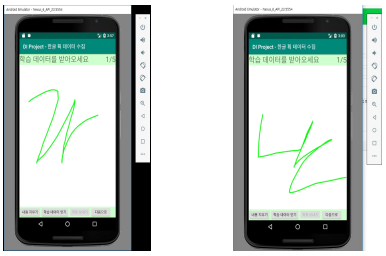
이전에 언급한 사전지식은 다음과 같이 적용한다. 필기할 데이터 '반'은 (그림 1)의 4번째 패턴이라는 사실을 알 수 있다. 또한 'ㅂ'은 4획, 'ㅏ'은 2획, 'ㄴ'은 1획으로 필기한다는 것도 알 수 있다. 사용자가 앱에 터치를 한번 하는 것이 곧 하나의 획을 긋는 것이므로, 터치 이벤트를 적절히 계산한다면 초성, 중성, 종성의 획을 추정할 수 있다. 따라서 그 획의 좌표 정보를 기반으로 초성, 중성, 종성을 다시 그려서 저장하고자 한다.

하지만 사람들은 필기를 할 때 항상 올바르게 정자를 쓰는 것이 아니다. (그림 3)은 빈번하게 'ㅂ'을 필기하는 방법 4가지를 파악해보았다. 물론 그 외의 방법도 있지만, 그림이 주는 의미는 'ㅂ'은 최소 2획에서 최대 4획을 이용하여 필기할 수 있다는 사실이다. 따라서 각 자음, 모음에 대하여 미리 최소, 최대 획수를 파악해둠으로써 다양한 필기 방식에 대처하고자 한다.



(그림 3) 'ㅂ'을 필기하는 경우

이와 같이 필기 데이터를 수집할 때 발생할 수 있는 문제는 사람들이 글씨를 쓸 때 초성과 중성, 혹은 중성과 종성을 이어 쓰는 경우(흘려쓰기)가 있다는 것이다.



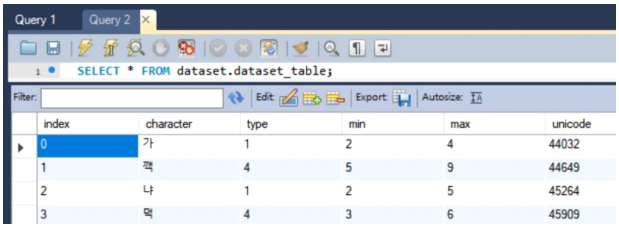
(그림 4) 흘려쓰기 예 – '가', '난'

(그림 4)는 '가'와 '난' 글자를 흘려 쓴 필기 데이터를 보여주고 있다. 이처럼 상대적으로 구분하기 곤란한 필기 데이터는 연구 초기부터 처리하기에는 곤란하므로, 이러한 데이터는 서버에서 사전에 별도의 공간에 저장하도록 배

제하였다. 추후 연구의 기술적인 수준이 높아지면 흘려쓴 필기 데이터를 인식하기 위해 사용하기 위해 별도로 저장하는 것이다.

2.3 시스템 구성

필기 데이터 수집, 분석, 관리를 담당하는 서버는 MySQL을 이용하여 수집할 학습 데이터를 미리 저장해둔다. (그림 5)를 보면, 각 글자에 대하여 패턴(type), 최소 획수(min) 및 최대 획수(max), 유니코드 값 등을 저장해 두었다가 사용자가 앱을 수행하면 일정 개수를 전달하도록 한다.



index	character	type	min	max	unicode
0	가	1	2	4	44032
1	각	4	5	9	44649
2	나	1	2	5	45264
3	역	4	3	6	45909

(그림 5) 수집할 데이터에 대하여 미리 분석한 테이블

사용자 앱은 (그림 6)과 같이 동작한다. 처음 시작하면 서버에 접속하여 일정 개수의 필기할 글자와 관련된 데이터를 받아온다. 사용자가 요구한 글자를 필기할 때 발생하는 터치 이벤트를 받아서 저장하여 서버에게 전달한다. 이러한 데이터는 2.2절에서 설명한 전처리 모듈에 의해 자음과 모음 데이터로 구분하여 28×28 픽셀의 이미지로 저장한다.

(그림 6) 사용자 앱의 동작 방식

3. 결론

본 연구의 목적은 한글 필기 인식 시스템 개발을 목적으로 하며, 그 중 초기 연구결과물인 데이터 수집 및 전처리 모듈에 대한 내용을 본 논문에 기재하였다. 이렇게 수집한 필기 데이터로부터 자음과 모음 데이터로 구분하여 저장한다.

추후 연구는 수집한 자음 및 모음 데이터가 모이게 되면, 기계학습의 학습용 및 테스트용 데이터로 활용하고자 한다. 즉, 우리 글자는 주로 가로 획, 세로 획, 빗금, 동그라미로 주로 구성된다. 이러한 특징을 정확하게 인식할 수 있는 인식 모델을 수립한 후 기계학습에 동작하여 온라인 필기 인식 시스템을 개발하고자 한다.

참고문헌

[1] 김두식, 이성환, "계층적 신경망 분류기를 이용한 다양한 언어, 크기 및 활자체 문자 인식," 한국정보과학회 논문지(B), 제25권, 5호, pp. 792-801, 1998

[2] 순환신경망을 이용한 한글 필기체 인식, 정보과학회 컴퓨팅의 실제 논문지, 제23권 제5호, pp. 316-321, 2017

[3] P. Louridas,C. Ebert, “Machine Learning”, IEEE Software, Vol. 33, Issue 5, pp. 110-115, 2016

[4] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition." Proc. of the IEEE, 86(11), pp. 2278-2324, 1998