만주 글자의 단위를 추출하는 효율적인 방법

스노우버거 아론 다니엘 · 이충호*

한밭대학교

An Efficient Method to Extract Units of Manchu Characters

Aaron Daniel Snowberger · Choong Ho Lee*

Graduate School of Hanbat National University

E-mail : aaron.snowberger@gmail.com / chlee@hanbat.ac.kr

요 약

만주 문자는 세로로 씌여지며 한 단어 안에서는 띄어쓰기 없이 이어져 있기 때문에 문자를 인식하기 전에 글자영역 분리와 글자를 이루는 단위를 분리해 내는 전처리과정이 필요하다. 본 논문에서는 글자영역을 추출하고 글자의 단위를 끊어내는 전처리 방법을 기술한다. 기존 연구가 단어별 또는 문자단위로 인식하는 방법을 전제로 하거나, 이어져 있는 글자의 줄기를 없앤 후 남는 부분으로 인식하는 것과 달리, 본 방법은 인식 가능한 단위별로 글자를 끊어낸 다음 그 단위의 합성으로 글자를 인식하는 방법에 적용할 수 있다. 실험을 통하여 본 방법의 유효성을 검증하였다.

ABSTRACT

Since Manchu characters are written vertically and are connected without spaces within a word, a preprocessing process is required to separate the character area and the units that make up the characters before recognizing the characters. In this paper, we describe a preprocessing method that extracts the character area and cuts off the unit of the character. Unlike existing research that presupposes a method of recognizing each word or character unit, or recognizing the remaining part after removing the stem of a continuous character, this method cuts the character into each recognizable unit. It can be applied to the method of recognizing letters by combining the units. Through an experiment, the effectiveness of this method was verified.

키워드

Manchu Characters, Character Recognition, Preprocessing, Pattern Recognition

Ⅰ. 서 론

만주 글자를 인식하기 위한 연구는 2000년대 초반부터 활발히 이루어져 왔다. 만주글자를 인식하기 위해서는 만주글자를 글자 단위로 추출하여 인식하고자 하는 방법이 이루어졌으나 글자단위 추출의 오류 때문에 인식률을 어느 한계 이상으로 높이는데 한계가 있었다. 그 방법 중 하나로 만주글이 세로로 씌여진다는 점에 착안하여 중심 축을 일률적으로 제거하고 양쪽에 남은 조각으로 인식하는 방법이 제안되어 있다[1].

하지만 이 방법은 글자 단위에서 중심축을 지나는 획이 있는 경우에도 일방적으로 제거되어 단위 인식의 오류가 있을 수 있다.

최근에는 글자단위 인식의 오류를 피하고자 단어 단위로 인식하고자 하는 방법이 제안되기도 하였다[2]. 이 방법에서는 띄어쓰기로 분리된 독립된 단어를 인식의 대상으로 하고 있다. 이 경우에 기계학습을 적용하는 경우, 독립된 단어를 추출하고 이것을 인위적으로 변형하여 대량의 데이터를 생성하여 학습데이터로 쓰는 방법도 제안하고 있다[3].

본 논문에서는 만주어 인식률을 높일 수 있는

* corresponding author

만주어 단위를 추출하는 새로운 방식을 제안하고 있다. 기존의 방법과는 달리 중심 줄기를 제거하지 않고 가로방향으로 분리하는 방법이다. 이 방법을 사용하면 중심축이 그대로 남아 있기 때문에 중심을 지나는 획이 있는 경우에도 단위를 그대로 추출할 수 있다.

II. 기존의 만주 글자 단위 분리 방법

기존의 방법 중 하나는 만주글자가 세로로 씌여지며 가운데 중심 축이 있다는 데에서 착안하여 중심 축을 제거하고 양 방향에 남은 조각들을 단위로 인식하는 것이다. 그림 1과 같다[1]. 이경우에 곡선형태로 중심축을 지나는 아크(arc)형태의 패턴은 원래 1개의 패턴이 2개의 패턴으로 분리되어 오추출의 인식이 될 수 있다.

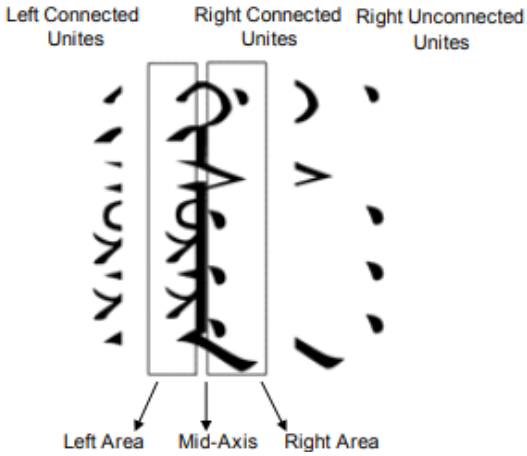

그림 1. 중심축을 제거하여 만주글자의 단위를 추출하는 방법

III. 제안하는 만주 글자 분리 방법

만주글로 씌여진 텍스트는 그림 2와 같이 세로로 씌여져 있다. 글자를 분리하는 대략적인 방법은 다음과 같다.

첫째 입력된 2치화 한다. 글자영역이 흰색이 되고 배경이 검정색이 되도록 반전시킨다. 이렇게 하는 것은 배경을 그레이레벨 0, 글자영역이 그레이레벨 255로 만들어 처리 시에 사람의 직관적 인식과 일치시키기 위한 것이다. 그 다음, 침식과 팽창으로 솔트 앤 페퍼 에러(salt and pepper)를 제거한다.

둘째, 세로 영역으로 투영하여 0이 되는 부분의 중심에서 세로로 글자영역을 끊어낸다.

셋째, 방금 위에서 끊어낸 세로 1줄을 가로방향으로 투영한다. 그레이레벨이 0이 지속되는 지점을 추출하여 그 정 가운데에서 끊어낸다. 이 때 단어가 추출된다.

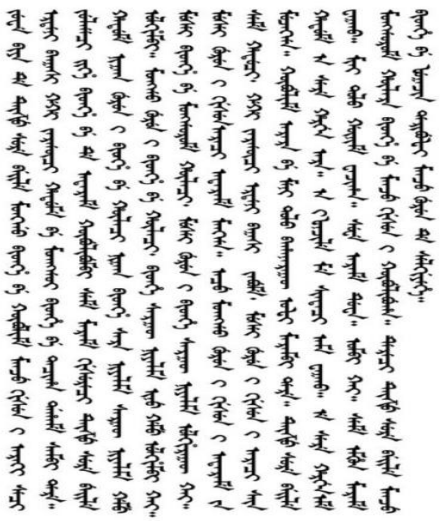

그림 2. 만주글자로 씌여진 텍스트

IV. 실험 결과

만주글자 단위 추출 알고리즘은 파이썬 프로그램으로 작성하였다. 그림 2의 가로 세로 459x748 크기의 텍스트를 세로로 투영하였을 경우의 누적 화소수는 아래 그림 3과 같다.

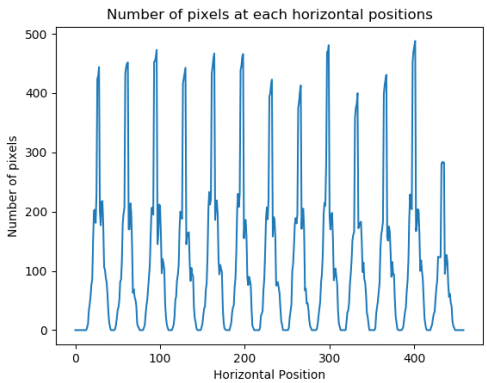

그림 3 수평 방향의 글자영역 해당 화소 수

그림 2에서 세로로 분리하는 지점은 6, 44, 78, 112, 146, 180, 214, 248, 281, 316, 350, 383, 417, 453으로 총 13개 열로 분리가능하다.

같은 방식으로 가로방향으로 투영한 후에 가로로 자르면 그림4와 같은 단어를 추출할 수 있다.

그림 4. 맨 왼쪽 열의 첫 번째 단어

그림 4와 같은 단어를 다시 가로로 투영하면 글자들이 모두 연결되어 있기 때문에 화소수가 0이 되는 부분은 존재하지 않는다. 그래서 경험적으로 임계치를 정하여 각 단어를 분리하는 방법을 취하였다.

최종적으로 얻어지는 결과는 그림 5와 같다. 끊어낸 위치는 7, 13, 20, 30 이다.

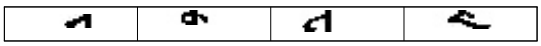

그림 5. 추출된 글자 단위

Ⅴ. 결 론

만주글자로 씌여진 텍스트에서 만주글자 단위를 효과적으로 추출하는 새로운 방법을 제안하였다. 기존의 방법과는 달리 중심축을 제거하지 않고 가로로 분리하는 방법을 제안하였다. 만주글자 텍스트를 2치화하고 반전시킨 후에 글자영역을 세로로 투영하여 세로 글자 영역을 추출하고, 다시 가로로 투영하여 가로 글자 단어를 추출하였다. 그 다음에 가로부분의 가늘어진 부분에 임계치를 적용하여 글자단위를 추출하였다. 제안된 방법의 유효성을 실험을 통하여 검증하였다.

References

[1] G.-Y. Zhang, J.-J. Li, A.-X. Wang, “A New Recognition Method for the Handwritten Manchu Character Unit,” in *Proceedings of the Fifth International Conference on Machine Learning and Cybernetics,* Dalian, pp. 13-16 August 2006.

[2] M. Li, R. Zheng, S. Xu, Y. Fu, “Manchu Word Recognition Based on Convolutional Neural Network with Spatial Pyramid Pooling,” in *2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics*, 978-1-5386-7604-2/18/, IEEE, 2018.

[3] R. Zheng, M. Li, J. He, J. Bi, B. Wu, “Segmentation-free Multi-font Printed Manchu Word Recognition Using Deep Convolutional Features and Data Augmentation,” *2018 11th International Congress on Image and Signal Processing,* 978-1-5386-7604-2/18, IEEE, 2018.