



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2022-0168787
(43) 공개일자 2022년12월26일

(51) 국제특허분류(Int. Cl.)

G06K 9/00 (2022.01)

(52) CPC특허분류

G06V 30/40 (2022.01)

G06V 30/10 (2022.01)

(21) 출원번호 10-2021-0078719

(22) 출원일자 2021년06월17일

심사청구일자 2021년06월17일

(71) 출원인

한밭대학교 산학협력단

대전광역시 유성구 동서대로 125 (덕명동)

(72) 발명자

이충호

대전광역시 유성구 왕가봉로 23, 1103-204

(74) 대리인

조해연

전체 청구항 수 : 총 16 항

(54) 발명의 명칭 만주어의 글자 추출 방법 및 이를 수행하는 시스템

(57) 요약

본 발명은 만주어의 글자 추출 방법에 관한 것이다.

본 발명의 실시 예를 따르는 만주어의 글자 추출 방법은, 적어도 하나의 문자열을 포함하는 문서 데이터를 이치화하는 이치화 단계; 상기 이치화된 문서 데이터를 제1방향으로 투영하여 화소수를 도출하는 제1투영단계; 상기

(뒷면에 계속)

대표도 - 도5



제1투영단계에서 도출된 화소수가 제1기준에 해당하는 제1지점을 도출하고, 상기 제1지점을 기준으로 구분하여 문자열을 추출하는 문자열추출단계; 상기 문자열추출단계에서 추출된 문자열을 제2방향으로 투영하여 화소수를 도출하는 제2투영단계; 상기 제2투영단계에서 도출된 화소수가 제2기준에 해당하는 제2지점을 도출하고, 상기 제2지점을 기준으로 구분하여 단어를 추출하는 단어추출단계; 상기 단어추출단계에서 추출된 단어를 제2방향으로 투영하여 화소수를 도출하는 제3투영단계; 및 상기 제3투영단계에서 도출된 화소수가 제3기준에 해당하는 제3지점을 도출하고, 상기 제3지점을 기준으로 구분하여 글자를 추출하는 글자추출단계;를 포함한다.

명세서

청구범위

청구항 1

적어도 하나의 문자열을 포함하는 문서 데이터를 이치화하는 이치화 단계;

상기 이치화된 문서 데이터를 제1방향으로 투영하여 화소수를 도출하는 제1투영단계;

상기 제1투영단계에서 도출된 화소수가 제1기준에 해당하는 제1지점을 도출하고, 상기 제1지점을 기준으로 구분하여 문자열을 추출하는 문자열추출단계;

상기 문자열추출단계에서 추출된 문자열을 제2방향으로 투영하여 화소수를 도출하는 제2투영단계;

상기 제2투영단계에서 도출된 화소수가 제2기준에 해당하는 제2지점을 도출하고, 상기 제2지점을 기준으로 구분하여 단어를 추출하는 단어추출단계;

상기 단어추출단계에서 추출된 단어를 제2방향으로 투영하여 화소수를 도출하는 제3투영단계; 및

상기 제3투영단계에서 도출된 화소수가 제3기준에 해당하는 제3지점을 도출하고, 상기 제3지점을 기준으로 구분하여 글자를 추출하는 글자추출단계;를 포함하는,

만주어의 글자 추출 방법.

청구항 2

제1항에 있어서,

상기 제1지점은 화소수가 0인 지점인,

만주어의 글자 추출 방법.

청구항 3

제1항에 있어서,

상기 제2지점은 화소수가 0인 지점인,

만주어의 글자 추출 방법.

청구항 4

제1항에 있어서,

상기 제3지점은 화소수가 0을 초과하고, 특정 값 이하인 지점인,

만주어의 글자 추출 방법.

청구항 5

제1항에 있어서,

상기 글자추출단계는,

상기 만주어의 중심축 영역을 식별하는 단계; 및

상기 만주어의 중심축 영역에 화소수가 0인 경우에는 상기 제3지점에서 제외하는 단계;를 포함하는 만주어의 글자 추출 방법.

청구항 6

문서 데이터 입력부 및 연산부를 포함하고,

상기 문서 데이터 입력부 적어도 하나의 문자열을 포함하는 문서 데이터를 입력 받는 단계를 수행하고,

상기 제어부는 이치화 단계, 제1투영단계, 문자열추출단계, 제2투영단계, 단어추출단계, 제3투영단계 및 글자추출단계를 수행하는,

만주어의 글자 추출 시스템.

청구항 7

제6항에 있어서,

상기 이치화 단계는 상기 제어부가 상기 문서 데이터 입력부에서 입력 받은 문서를 이치화하는 것인,

만주어의 글자 추출 시스템.

청구항 8

제6항에 있어서,

상기 제1투영단계는 상기 제어부가 상기 이치화된 문서 데이터를 제1방향으로 투영하여 화소수를 도출하는 것인,

만주어의 글자 추출 시스템.

청구항 9

제6항에 있어서,

상기 문자열추출단계는, 상기 제어부가 상기 제1투영단계에서 도출된 화소수가 제1기준에 해당하는 제1지점을 도출하고, 상기 제1지점을 기준으로 구분하여 문자열을 추출하는 것인,

만주어의 글자 추출 시스템.

청구항 10

제6항에 있어서,

상기 제2투영단계는, 상기 제어부가 상기 문자열추출단계에서 추출된 문자열을 제2방향으로 투영하여 화소수를 도출하는 것인,

만주어의 글자 추출 시스템.

청구항 11

제6항에 있어서,

상기 단어추출단계는, 상기 제어부가 상기 제2투영단계에서 도출된 화소수가 제2기준에 해당하는 제2지점을 도

출하고, 상기 제2지지점을 기준으로 구분하여 단어를 추출하는 것인,
만주어의 글자 추출 시스템.

청구항 12

제6항에 있어서,
상기 제3투영단계는, 상기 제어부가 상기 단어추출단계에서 추출된 단어를 제2방향으로 투영하여 화소수를 도출하는 것인,
만주어의 글자 추출 시스템.

청구항 13

제6항에 있어서,
상기 글자추출단계는 상기 제3투영단계에서 도출된 화소수가 제3기준에 해당하는 제3지점을 도출하고, 상기 제3지지점을 기준으로 구분하여 글자를 추출하는 것인,
만주어의 글자 추출 시스템.

청구항 14

제13항에 있어서,
상기 제3지점은 화소수가 0을 초과하고, 특정 값 이하인 지점인,
만주어의 글자 추출 시스템.

청구항 15

제13항에 있어서,
상기 글자추출단계는,
상기 만주어의 중심축 영역을 식별하는 단계; 및
상기 만주어의 중심축 영역에 화소수가 0인 경우에는 상기 제3지점에서 제외하는 단계;를 포함하는
만주어의 글자 추출 시스템.

청구항 16

프로세서에 의해 실행됨으로써 제1항의 방법을 실행하는,
컴퓨터에서 판독가능한 저장매체에 저장된 프로그램.

발명의 설명

기술 분야

본 발명은 만주어의 글자 추출 방법에 관한 것이다.

[0001]

배경 기술

- [0003] 최근 문서를 스캔 또는 촬영하여 생성한 화상 데이터로부터 문자를 인식하거나, 카메라를 통해 촬영하면서 실시간으로 문자를 인식하는 기술이 활용되고 있다. 선행기술문헌인 한국 공개특허 제10-2017-0032347호는 기설정된 폭 값을 기준으로 문자 경계를 인식함으로써 문자를 인식하는 문자 인식 기술을 개시한다.
- [0004] 이러한 종래의 문자 인식 기술은 각 글자가 일정한 간격으로 찍여지거나 띄어쓰기가 되어 있는 경우에 적용이 가능하다. 상기 선행기술문헌의 경우에도 문서 데이터를 이치화한 후 화소값이 0인 지점을 경계로 삼고 있다.
- [0005] 만주어는 세로로 찍여지며 한 단어 안에서는 띄어쓰기 없이 이어져 있기 때문에 문자를 인식하기 전에 글자영역 분리와 글자를 이루는 단위를 분리해 내는 전처리과정이 필요하다. 상기 선행기술문헌 등의 종래의 기술로는 띄어쓰기 없이 이어져 쓰여진 만주어 문서에서 각 글자를 추출하는 것이 불가능하다.

선행기술문헌

특허문헌

- [0007] (특허문헌 0001) 한국 공개특허 제10-2017-0032347호

발명의 내용

해결하려는 과제

- [0008] 본 발명은 만주어로 쓰여진 문서 데이터로부터 각 글자 단위를 추출할 수 있는 만주어의 글자 추출 방법을 제공함을 목적으로 한다.
- [0009] 또한, 만주어 글자 출추함에 있어서 효율적이고 오류가 적다.

과제의 해결 수단

- [0011] 본 발명의 실시 예를 따르는 만주어의 글자 추출 방법은, 적어도 하나의 문자열을 포함하는 문서 데이터를 이치화하는 이치화 단계; 상기 이치화된 문서 데이터를 제1방향으로 투영하여 화소수를 도출하는 제1투영단계; 상기 제1투영단계에서 도출된 화소수가 제1기준에 해당하는 제1지점을 도출하고, 상기 제1지점을 기준으로 구분하여 문자열을 추출하는 문자열추출단계; 상기 문자열추출단계에서 추출된 문자열을 제2방향으로 투영하여 화소수를 도출하는 제2투영단계; 상기 제2투영단계에서 도출된 화소수가 제2기준에 해당하는 제2지점을 도출하고, 상기 제2지점을 기준으로 구분하여 단어를 추출하는 단어추출단계; 상기 단어추출단계에서 추출된 단어를 제2방향으로 투영하여 화소수를 도출하는 제3투영단계; 및 상기 제3투영단계에서 도출된 화소수가 제3기준에 해당하는 제3지점을 도출하고, 상기 제3지점을 기준으로 구분하여 글자를 추출하는 글자추출단계;를 포함한다.
- [0013] 상기 제1지점은 화소수가 0인 지점일 수 있다.
- [0015] 상기 제2지점은 화소수가 0인 지점일 수 있다.
- [0017] 상기 제3지점은 화소수가 0을 초과하고, 특정 값 이하인 지점일 수 있다.
- [0019] 상기 글자추출단계는, 상기 만주어의 중심축 영역을 식별하는 단계; 및 상기 만주어의 중심축 영역에 화소수가 0인 경우에는 상기 제3지점에서 제외하는 단계;를 포함할 수 있다.

- [0021] 본 발명의 실시 예를 따르는 만주어의 글자 추출 시스템은, 문서 데이터 입력부 및 연산부를 포함한다.
- [0022] 상기 문서 데이터 입력부 적어도 하나의 문자열을 포함하는 문서 데이터를 입력 받는 단계를 수행하고, 상기 제어부는 이치화 단계, 제1투영단계, 문자열추출단계, 제2투영단계, 단어추출단계, 제3투영단계 및 글자추출단계를 수행한다.

발명의 효과

- [0024] 본 발명의 실시 예를 따르는 만주어의 글자 추출 방법은, 만주어로 쓰여진 문서 데이터로부터 각 글자 단위를 추출할 수 있다.
- [0025] 또한, 만주어 글자 출추함에 있어서 효율적이고 오류가 적다.

도면의 간단한 설명

- [0027] 도 1은 만주어 문서 데이터를 도시한 것이다.
- 도 2는 도 1의 만주어 문서 데이터를 이치화한 후 X축 방향으로 투영하였을 때의 화소값을 나타낸 것이다.
- 도 3은 도 1에서 추출된 문자열을 도시한 것이다.
- 도 4는 도 3에서 추출된 단어를 도시한 것이다.
- 도 5는 도 4에서 추출된 글자를 도시한 것이다.
- 도 6은 본 발명의 실시 예를 따르는 만주어의 글자 추출 방법을 도시한 것이다.
- 도 7은 본 발명의 실시 예를 따르는 만주어의 글자 추출 시스템을 도시한 것이다.

발명을 실시하기 위한 구체적인 내용

- [0028] 이하, 첨부된 도면을 참조하여 본 발명의 바람직한 실시 형태들을 다음과 같이 설명한다. 그러나, 본 발명의 실시 형태는 여러 가지 다른 형태로 변형될 수 있으며, 본 발명의 범위가 이하 설명하는 실시 형태로 한정되는 것은 아니다. 또한, 본 발명의 실시 형태는 당해 기술분야에서 평균적인 지식을 가진 자에게 본 발명을 더욱 완전하게 설명하기 위해서 제공되는 것이다. 따라서, 도면에서의 요소들의 형상 및 크기 등은 보다 명확한 설명을 위해 과장될 수 있으며, 도면 상의 동일한 부호로 표시되는 요소는 동일한 요소이다. 또한, 유사한 기능 및 작용을 하는 부분에 대해서는 도면 전체에 걸쳐 동일한 부호를 사용한다. 덧붙여, 명세서 전체에서 어떤 구성요소를 "포함"한다는 것은 특별히 반대되는 기재가 없는 한 다른 구성요소를 제외하는 것이 아니라 다른 구성요소를 더 포함할 수 있다는 것을 의미한다.
- [0030] 도 6은 본 발명의 실시 예를 따르는 만주어의 글자 추출 방법을 도시한 것이다.
- [0031] 도 6을 참조하면, 본 발명의 실시 예를 따르는 만주어의 글자 추출 방법은, 적어도 하나의 문자열을 포함하는 문서 데이터를 이치화하는 이치화 단계; 상기 이치화된 문서 데이터를 제1방향으로 투영하여 화소수를 도출하는 제1투영단계; 상기 제1투영단계에서 도출된 화소수가 제1기준에 해당하는 제1지점을 도출하고, 상기 제1지점을 기준으로 구분하여 문자열을 추출하는 문자열추출단계; 상기 문자열추출단계에서 추출된 문자열을 제2방향으로 투영하여 화소수를 도출하는 제2투영단계; 상기 제2투영단계에서 도출된 화소수가 제2기준에 해당하는 제2지점을 도출하고, 상기 제2지점을 기준으로 구분하여 단어를 추출하는 단어추출단계; 상기 단어추출단계에서 추출된 단어를 제3방향으로 투영하여 화소수를 도출하는 제3투영단계; 및 상기 제3투영단계에서 도출된 화소수가 제3기준에 해당하는 제3지점을 도출하고, 상기 제3지점을 기준으로 구분하여 글자를 추출하는 글자추출단계;를 포함한다.

- [0033] 본 발명의 실시 예를 따르는 방법들은 만주어의 글자 추출 시스템을 통해 수행될 수 있다. 도 7은 본 발명의 실시 예를 따르는 만주어의 글자 추출 시스템을 도시한 것으로, 도 7을 참조하면, 상기 시스템은 문서 데이터 입력부 및 연산부를 포함한다. 상기 시스템은 통상적으로 정보의 저장 및 연산을 수행하기 위해 사용하는 컴퓨터, 노트북 컴퓨터, 휴대용 단말기, 서버 등일 수 있다.
- [0035] 상기 문서 데이터 입력부는 외부로부터 만주어를 포함하는 문서 데이터를 입력 받을 수 있다. 상기 문서 데이터는 만주어로 된 문자열을 적어도 하나 포함하는 것으로, 만주어 문서를 스캔하거나 카메라로 촬영하여 생성된 문서 데이터일 수 있다. 또는, 카메라로 촬영하고 있는 중에 실시간으로 전송됨으로써 컴퓨터의 메모리 반도체 칩에 일시적 또는 영구적으로 저장된 문서 데이터일 수 있다. 상기 문서 데이터 입력부는 데이터를 입력 받을 수 있는 것이면 특별히 제한하지 않는다.
- [0037] 상기 연산부는 입력 받은 문서 데이터를 처리하고 연산하는 작업을 수행한다. 이러한 작업들은 컴퓨터에서 관독 가능한 저장매체에 저장된 프로그램이 컴퓨터의 프로세서 또는 반도체 칩에 의해 실행됨으로써 수행하는 것일 수 있다. 상기 연산부는 프로세서 또는 반도체 칩을 적어도 하나 포함하는 것으로 통상적으로 소프트웨어 프로그램을 실행하고 정보를 처리하여 사용자에게 다양한 정보를 제공하는 것일 수 있다.
- [0039] 이치화란 스캔 또는 카메라로 촬영하여 얻어진 화상 데이터를 일정한 기준으로 구분하여 표현한 것을 의미한다. 일 예로, 그림 또는 문서 데이터를 '흑' 또는 '백'으로 구분하고, 각각 '0'과 '1' 또는 '1'과 '0'의 비트 데이터를 표현할 수 있다. 이러한 이치화 방법은 본 기술분야에서 일반적으로 사용하고 알려진 이치화 프로그램 등에 의해 수행할 수 있으며 특별히 제한하지 않는다. 이치화를 통해 각 지점 또는 영역의 화소수를 셀 수 있다.
- [0041] 본 발명의 실시 예에서, 글자 영역이 흰색이 되고, 배경색이 검정색이 되도록 반전하여 글자 영역이 그레이레벨 255, 배경색이 그레이레벨 0이 되도록 하였다. 다음으로, 침식과 팽창으로 솔트 앤 페퍼 에러(salt and pepper error)를 제거하였다.
- [0043] 상기 제1투영단계 및 문자열추출단계는 이치화된 문서 데이터를 일 방향으로 순차적으로 화소수를 세고, 화소수를 기준으로 문자열을 추출하는 단계이다. 도 1을 참조하면, 만주어는 세로로 쓰여 있기 때문에 가로 방향인 x축 방향으로 투영을 함으로써 문자열을 구분하는 제1지점을 도출할 수 있다.
- [0044] 도 2는 도 1의 문서 데이터를 이치화한 후 x축 방향으로 투영하여 추출한 누적 화소수를 나타낸 그래프이다. 도 2를 참조하면, 일정한 간격마다 화소수가 0인 지점이 나타나는 것을 알 수 있으며, 이 지점이 문자열의 경계와 동일함을 알 수 있다. 보다 구체적으로, 도 2에서 세로로 분리하는 지점은 6, 44, 78, 112, 146, 180, 214, 248, 281, 316, 350, 383, 417, 453으로 총 13개 열로 분리가능하다.
- [0045] 본 단계에서는 화소수가 0인 지점을 제1지점으로 판정하고, 이 지점을 기준으로 나눔으로써 각각의 문자열을 추출할 수 있다. 도 3(a)는 위의 방법으로 추출된 문자열 중 하나를 표현한 것이다.
- [0047] 상기 제2투영단계 및 단어추출단계는 각각의 문자열 데이터를 일 방향으로 순차적으로 화소수를 세고, 화소수를 기준으로 단어를 추출하는 단계이다. 도 3(a)를 참조하면, 만주어에서 문자열은 여러 단어에 의해 구별됨을 알 수 있다. 각각의 단어는 세로로 쓰여져 있기 때문에 세로 방향인 y축 방향으로 투영을 함으로써 단어를 구분하는 제2지점을 도출할 수 있다.
- [0048] 도 3(b)는 도 3(a)에서 제2지점을 표시한 것이다. 도 3(a)를 y축 방향으로 투영하는 경우, 각 단어의 사이마다 화소수가 0인 지점이 나타나는 것을 알 수 있으며, 이 지점이 단어의 경계와 동일함을 알 수 있다.
- [0049] 본 단계에서는 화소수가 0인 지점을 제2지점으로 판정하고, 이 지점을 기준으로 나눔으로써 각각의 단어를 추출할 수 있다. 도 4(a)는 위의 방법으로 추출된 단어 중 하나를 표현한 것이다.

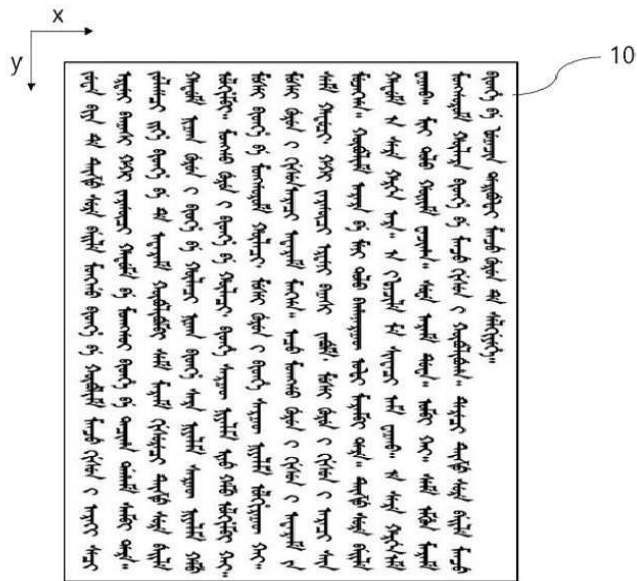
- [0051] 상기 제3투영단계 및 글자추출단계는 각각의 단어 데이터를 일 방향으로 순차적으로 화소수를 세고, 화소수를 기준으로 글자를 추출하는 단계이다. 도 4(a)를 참조하면, 만주어에서 단어는 여러 글자가 연속적으로 연결된 것임을 알 수 있다. 각각의 글자는 세로로 쓰여 있기 때문에 세로 방향인 y축 방향으로 투영을 함으로써 글자를 구분하는 제3지점을 도출할 수 있다. 그러나, 앞선 방법과 달리, 만주어에서 글자는 서로 연결되어 쓰이며 가운데 부분에 위치하는 중심축이 존재한다. 따라서, 단순히 화소수가 0인 지점을 기준으로 제3지점을 도출할 수 없다.
- [0052] 본 단계에서는 상기 제3지점은 화소수가 0을 초과하고, 특정 값 이하인 지점일 수 있다. 즉, 화소수에 대한 상한 및 하한의 범위를 미리 설정하여 두고, 단어를 y축 방향으로 투영하여 획득한 화소수가 상기 상한 및 하한의 범위에 해당하는 경우를 제3지점으로 도출할 수 있다.
- [0053] 도 4(b)는 화소수가 일정한 범위에 해당하는 지점을 제3지점으로 도출하여 표시한 것이고, 도 5(a) 내지 (d)는 앞서 설명한 방법에 의해 추출된 글자를 각각 표시한 것이다. 보다 구체적으로, 도 4(b)에서 세로로 분리하는 지점은 7, 13, 20, 30으로 총 4개 열로 분리가 가능하다.
- [0055] 도 4(a)에 표시된 단어의 아래 부분을 참조하면, 중심축이 없는 영역에서도 화소수가 상기 일정범위에 해당하는 것으로 나타날 우려가 있다. 따라서, 이러한 문제를 해결하기 위해 상기 글자추출단계는, 만주어의 중심축 영역에 화소수가 0이거나 일정한 범위(제3지점을 판단하는 기준이 되는 화소수 범위)를 벗어나는 경우에는 상기 제3지점에서 제외하는 단계;를 포함할 수 있다.
- [0056] 상기 중심축 영역을 식별하는 단계는, 사용자에 의해 중심축의 위치를 입력 받는 단계를 포함하거나, 단어의 전체 폭 및 화소수를 기초로 연산하여 중심축의 좌표를 도출하는 포함할 수 있다. 이러한 방법을 통해 중심축의 화소수가 0이거나 일정한 범위를 벗어나는 경우에는, 해당 행의 화소수가 제3지점을 판단하는 기준이 되는 화소수 범위에 해당하더라도 제3지점에서 제외할 수 있어, 보다 정확하게 글자를 추출할 수 있다.
- [0058] 본 발명은 상술한 실시 형태 및 첨부된 도면에 의해 한정되는 것이 아니며 첨부된 청구범위에 의해 한정하고자 한다. 따라서, 청구범위에 기재된 본 발명의 기술적 사상을 벗어나지 않는 범위 내에서 당 기술분야의 통상의 지식을 가진 자에 의해 다양한 형태의 치환, 변형 및 변경이 가능할 것이며, 이 또한 본 발명의 범위에 속한다고 할 것이다.

부호의 설명

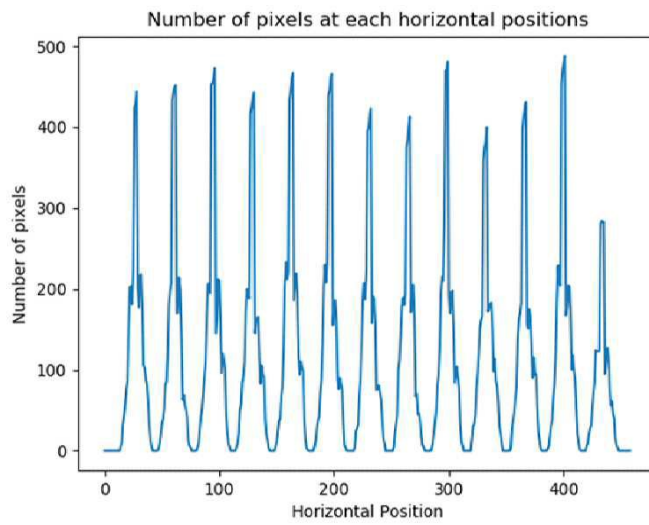
- [0060] 10: 만주어 문서 데이터

도면

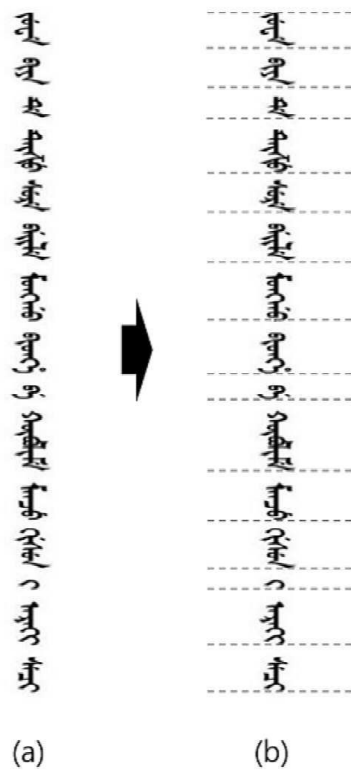
도면1



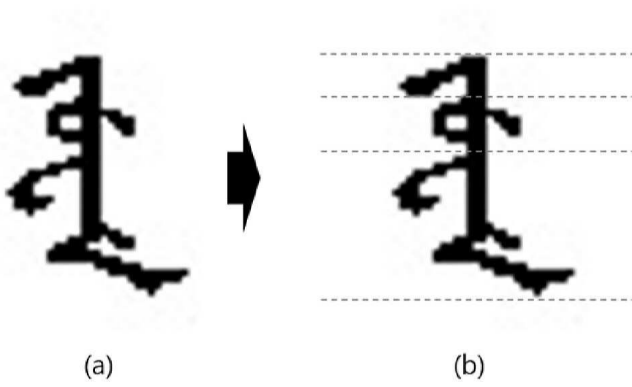
도면2



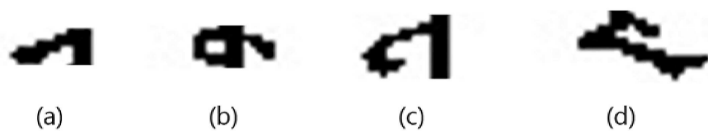
도면3



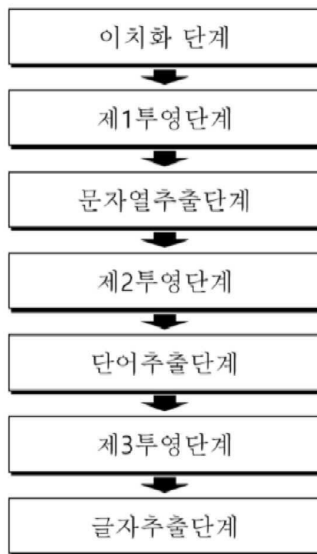
도면4



도면5



도면6



도면7

