JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 23, No. 11, pp. 2239-2248, Nov. 2022

Check for updates

한국 고문헌 디지털화를 위한 라인분할 및 서순정렬 알고리즘

이 아 람[1] · 민 기 현[2*] · 김 거 식[3] · 김 정 은[4] · 강 현 서[5]
[1]연구원 [2*]선임연구원 [3]책임연구원 [4]선임연구원 [5]책임연구원 한국전자통신연구원, 광ICT융합연구실

Line Segmentation and Reading Order Detection Algorithm for Digitization of Korean Historical Texts

Aram Lee[1] · Gihyeon Min[2*] · Keo Sik Kim[3] · Jeong Eun Kim[4] · Hyun Seo Kang[5]

[1]Researcher [2*]Senior researcher [3]Principal researcher [4]Senior researcher [5]Principal researcher Optical ICT Convergence Research Section, Electronics and Telecommunications Research Institute (ETRI), Gwangju 61012, Korea

[요 약]

본 연구는 고문헌 원문이미지 내 개별 한자들의 서순정렬을 통해 자동 디지털 텍스트화가 가능한 라인분할 알고리즘을 개발하였다. 우종서 기반의 서순을 따르며 본문과 주석이 혼재되어 직관적인 서순정렬이 어려운 한국 고문헌의 특성을 고려하여 광학문자인식(OCR; optical character recognition)을 통해 정의된 각 한자들의 크기와 좌표정보를 투영 프로파일 방법으로 분석하는 접근을 택하였다. 또한 원문이미지의 종서 수직 정렬을 최적화하는 기울기 보정 알고리즘을 적용하여 투영 프로파일 분석의 정확도를 개선하였으며 최종적으로 소·중·대분류 3종의 계층적 라인분할을 통해 서순이 정렬된 디지털 텍스트 추출을 가능케 하였다. 본 연구에서 개발된 라인분할과 서순정렬 알고리즘은 기존의 OCR과 자동번역 기술의 접점을 마련하여 고문헌 이미지를 다양한 언어의 디지털 텍스트로 변환하는 end-to-end의 전자동 프로세스 구현에 기여할 것으로 예상된다.

[Abstract]

In this study, we developed a line segmentation algorithm which identifies the reading order of individual characters in a captured image of ancient text for its automatic digitization. To overcome the ambiguities in an assignment of reading orders in Korean historical texts, which follows a unique reading order(top-bottom and right-left) and consists of a mixed form of main/annotation texts, a projection profile method was adopted in which the coordinates and size of each character obtained by optical character recognition(OCR) are geometrically analyzed. In addition, the accuracy of projection based analysis was improved by applying a tilt correction algorithm that optimizes the vertical alignment of the columns in an original image. Finally, through a 3-stage hierarchical segmentation, digital texts could be extracted in a recovered reading order. Serving as an interface between existing OCR and auto-translation technologies, the line segmentation and reading order detection algorithm developed in this study is expected to implement an end-to-end translation service, converting a snapshot of a historical text into digital texts in different languages.

색인어 : 디지털텍스트, 고문헌, 광학문자인식, 라인분할, 서순정렬

Keyword : Digital text, Historical text, Optical character recognition, Line segmentation, Reading order detection

http://dx.doi.org/10.9728/dcs.2022.23.11.2239

Received 11 October 2022; **Revised** 28 October 2022
Accepted 08 November 2022

***Corresponding Author; Gihyeon Min**

Tel: +82-62-970-6688
E-mail: ghmin@etri.re.kr

Ⅰ. 서 론

국내 소장 도서 중 통상적으로 1910년 이전에 제작된 간본, 사본, 문서류는 고문헌으로 분류되며 그 보유량은 2021년 기준 국공립 도서관과 연구기관 230만 점, 지역별 박물관과 기타기관 70만 점을 포함하여 총 300만 점에 달한다[1]. 고문헌 중 공적 역사기록을 다룬 사료가 가장 큰 비중을 차지하며 대표적으로 삼국사기, 조선왕조실록, 승정원일기 등이 있다. 1999년 국책사업의 일환으로 한국학 자료의 DB화가 추진되었으며 광학 이미징과 디지털 데이터 압축기술의 발전으로 전체 장서량의 5%에 달하는 원문 이미지들이 현재 부분적으로 제공되고 있다[1]. 이러한 고문헌의 DB화는 정보 공유 뿐 아니라 해당 자료의 훼손이나 유실의 경우에 대비하여 유구한 문화유산을 보존하기 위한 목적도 있다.

고문헌 내 개별 자형까지 디지털화하는 원문텍스트 서비스는 원문이미지 서비스 대비 문헌의 활용도가 높지만, 다양한 서체의 한자가 주를 이루는 고문헌을 대상으로 한 인적 기반의 개별 자형 판별 및 전산 입력 과정에 많은 노동력과 비용이 요구되므로 전체 고문헌 보유량 대비 서비스 구축 규모가 미비한 실정이다[2]. 이러한 제한적 상황을 극복하기 위해 이미지 내부에 포함된 텍스트를 자동으로 검출·인식하는 광학문자인식(OCR; optical character recognition) 기술이 대안으로 부상하고 있다[3]. 2010년대 딥러닝 기술의 획기적인 발전으로 CNN/RNN 등의 알고리즘을 기반으로 한 문자인식 연구가 활성화되었으며 인공지능 컴퓨팅 프로세서의 성능 개선과 학습 DB의 대형화는 OCR 성능을 가속화시켜 영문을 대상으로 한 경우 99% 이상의 인식 정확도가 보고되고 있다[4]. 고서 한자의 경우 영문 알파벳 대비 수백 배 이상으로 다분화된 자형으로 인해 비교적 낮은 80% 대의 인식정확도를 기록하고 있지만 이미지 전처리 기법과 지속적인 DB 증강을 토대로 꾸준한 상승세를 보이고 있다[5, 6].

그림 1과 같이 OCR 알고리즘은 스캔된 도서 내 각 자형에 대한 위치(x, y)와 크기(w, h) 정보를 포함하는 바운딩 박스(bounding box)를 형성(검출)하고 자형의 종류를 추론(인식)한 뒤 레이블 파일에 기록한다. 이러한 결과물은 개별 한자에 대한 독립적인 정보만을 제공할 뿐 그들의 연속적 집합인 어구와 문장을 구성하지 못하며, 후속 단계인 번역의 과정에 직접적으로 사용되기 위해서는 인적 기반 우종서 서순의 수동 배정이 요구된다. 이와 같은 수동적 후처리는 전체 고문헌 디지털화 과정의 병목으로 작용하여 작업시간과 비용을 증가시키며 이로 인해 OCR 기술의 잠재력이 충분히 활용되지 못하고 있는 실정이다. 따라서 본 연구에서는 한문 고서 이미지에 대한 OCR의 검출 결과(x, y, w, h)를 기반으로 라인분할(line segmentation)을 통해 우종서 서순이 자동으로 적용된 원문텍스트를 출력하는 알고리즘을 개발하였다.

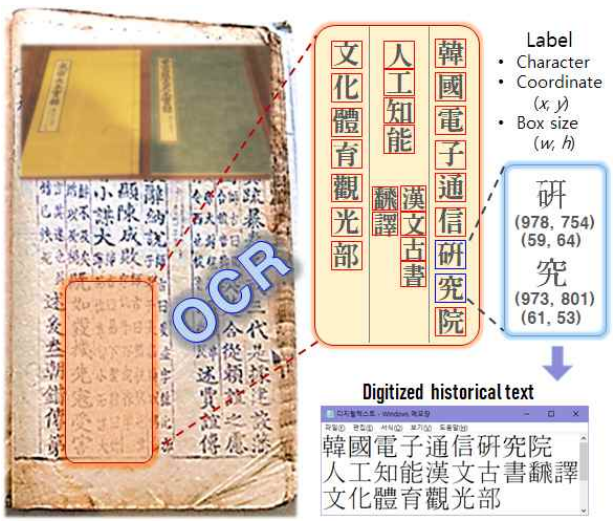

그림 1. 광학문자인식 기반 고문헌 디지털화 프로세스
Fig. 1. Digitization process of historical text based on OCR

Ⅱ. 한국 고문헌의 본문/세주 구조 및 서순 정의

한자문화권의 영향을 받은 한국의 고문헌들은 대부분의 기록이 한자로 남겨졌을 뿐 아니라, 그림 2(좌)와 같이 우측열부터 위에서 아래로 쓰여진 후 좌측으로 열을 옮겨가는 우종서의 형식을 따른다. 우종서는 종이가 보급되기 전 대나무를 엮어 만든 죽간을 기록매체로 사용하던 시기에 왼손으로 두루마리를 펴면서 오른손으로 기록을 진행하기에 적합하여 오랜 기간 문헌 작성의 표준이 되었다고 알려진다. 현재까지 전래된 고문헌들은 대부분 이보다 이후의 시기에 목판, 목활자, 금속활자 등을 이용해 종이 상에 인출되었지만 서순은 여전히 기존의 우종서를 따랐다.

고문헌의 구조 중 세주(annotation)는 그림 2(우)와 같이 본문(body)의 단일 열에서 비교적 작은크기의 글자들이 두 개의 열로 분화되는 부분을 명하며 이는 앞선 본문에 대한 주석의 역할을 한다. 기존의 인력기반 고문헌 디지털화 과정에서는 제한된 시간적/경제적 자원하에 세주를 제외하고 본문만을 번역한 경우도 있다. 하지만 역사학 연구에 있어서 세주가 중요한 정보를 전달하는 경우도 있으므로 이를 포함한 효율적인 완역에 대한 기술적 발전이 필요하다. 대표적인 예로 1454년 편찬된 세종실록 지리지의 우산도(독도)와 무릉도(울릉도)에 대한 기록 중 "二島相去不遠, 風日淸明, 則可望見 (두 섬은 서로 거리가 멀지 않아서 날씨가 맑으면 바라볼 수 있다)" 라는 구절이 세주에서 발견되어 대한민국의 독도 영유권 주장에 강력한 근거로 제시되고 있다[7] (그림 12 노란색 음영참조). 또한 주석서 등에 만재된 세주들은 기존 문헌에 대한 후대 지식인들의 세대를 거듭한 고찰을 함축하므로 역사 전통을 향유하는데 기여한다.

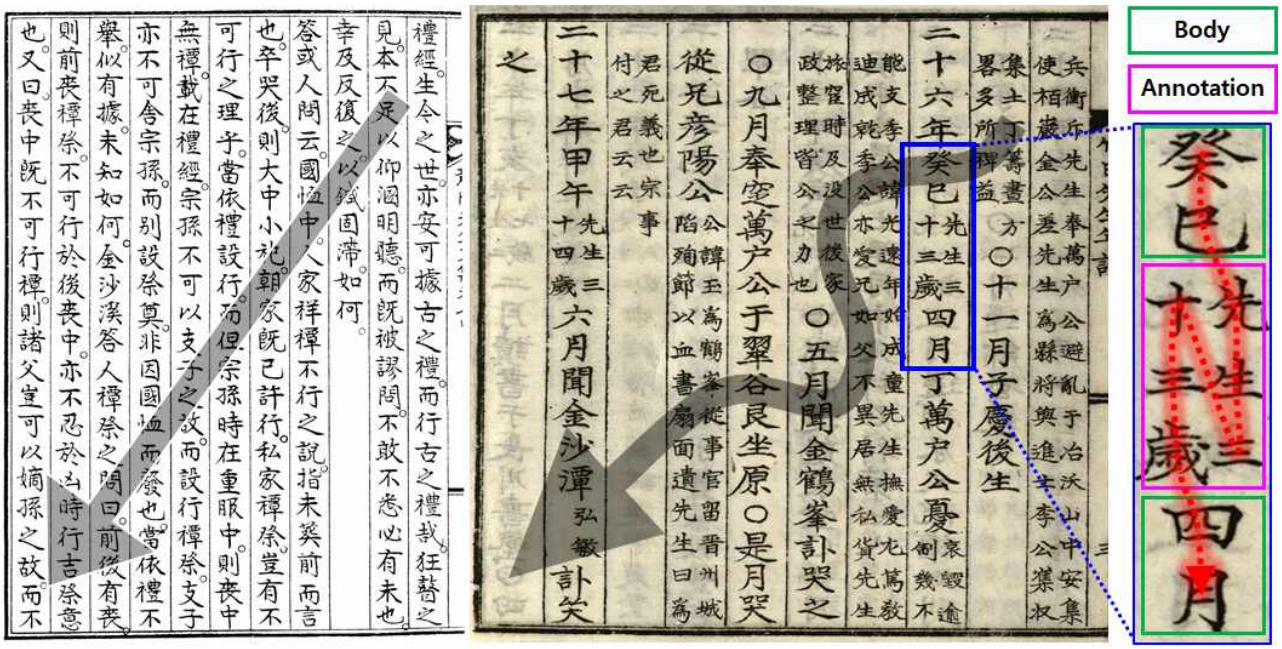

그림 2. 우종서 기반의 고문헌 서순. (좌) 본문만을 포함한 원문 이미지, (우) 본문과 세주가 혼용된 원문 이미지.
Fig. 2. Historical text written in top-bottom and right-left order. Left and right Images w/o and w/ annotation, respectively.

그림 2(좌)의 본문만이 포함된 원문이미지 대비 그림 2(우)의 세주가 혼재된 원문이미지는 빨간 화살표와 같이 서순의 방향이 다소 모호하다. 일반적으로 우종서 서순 진행 중 세주가 출현할 경우 해당 세주 군집(분홍색 네모) 내에서 우측의 세주열(우세주, right annotation)을 우선적으로 읽은 뒤 좌측 세주열(좌세주, left annotation)로 이동해야 하며 상기 예시의 경우 서순이 '癸巳先生三十三歲四月'로 되어야 한다. 하지만 단순히 우종서 원칙을 기반으로 우→좌(x축) 또는 상→하(y축) 순으로 정렬될 경우, 결과는 각각 '先生三癸巳四月十三歲', '癸巳先十生三三歲四月'로 오독이 발생한다. 심지어, 이미지 내 문자들의 행·열 정렬이 완벽하지 않은 경우 그 결과의 예측이 더욱 어려우며 세주가 없는 그림 2(좌)의 경우도 동일한 문제가 발생할 수 있다. 이러한 이유로 원문이미지 내의 행(횡서의 경우) 또는 열(종서의 경우)을 분할하는 라인분할과 각 분할요소 내의 문자 서순을 배정하는 알고리즘에 관한 다음과 같은 연구들이 수행되었다.

Ⅲ. 라인분할 및 서순정렬 알고리즘 비교

원문이미지의 라인분할은 크게 top-down과 bottom-up 방식으로 나눌 수 있으며 전자의 경우 이미지 전체를 대상으로 구획을 설정하는 추론을, 후자의 경우 사전에 검출된 개별 문자들의 위치정보를 기반으로 군집화를 실행한다.

초창기의 top-down 라인분할 방식인 projection profile 기법은 이진화 된 이미지 내 픽셀들이 한 축으로 누적 투영된 히스토그램을 분석하여 픽셀 밀도가 문턱값보다 낮은 부분에 경계선을 형성한다[8]. 이러한 기법은 이웃한 행·열간 거리가 가깝거나 겹치는 경우 문턱값을 높게 설정하여야 하며 라인분할 정확도가 하락한다. 추후 딥러닝의 융합으로 라인분할 사전학습 DB를 통한 모델 개발이 진행되었으며, 한문 고서를 대상으로 한 대표적인 성공사례로 layout analysis가 있다[9]. 해당 기술은 Resnet 기반의 backbone을 사용하여 feature를 추출한 뒤 문자단위의 검출·인식을 위한 character branch와 라인영역 검출을 위한 layout branch의 병렬적 조합을 통해 본문열과 세주열 영역을 분리하였다. 딥러닝의 추가적인 적용 사례로 라인분할 정확도 상승을 위해 라인 전체가 아닌 문자의 테두리나 x-height만을 대상으로 경계를 형성하는 연구도 보고되었다[10]. 상기 지도학습 기반 모델의 구축에 필요한 학습 DB 라벨링의 비효율성을 우회하기 위해 비지도 학습을 통한 라인분할 또한 제시되었지만 원문의 정형도(행·열의 수직·수평 정렬)에 따라 정확도가 크게 변화하는 단점이 있다[11].

Bottom-up 라인분할의 대표적인 사례로, 국내에서는 OCR로 검출된 각 문자들의 중심 좌표를 대상으로 문서 좌하단으로부터의 거리를 비교하여 서순을 정하는 시도가 있었지만 세주 등의 비정형 구조에는 적용할 수 없다[12]. 이와 같은 bottom-up 기반의 서순배정은 사람의 인지에는 직관적일 수 있으나, 수학적 연산처리에서는 각 문자들의 x, y 두 위치정보에 대해 종합적 우선순위를 판별(degree of freedom = 2)하는 것이 어렵다. 따라서 본 연구에서는 OCR로 검출된 문자 좌표(bottom-up)를 수직 방향으로 x축에 투영한 뒤 projection profile 분석(top-down)을 적용한다. 이와 같은 hybrid 방식의 라인분할을 통해 형성된 각 라인요소들은 x축 위치정보가 배제된 y축의 정보(degree of freedom = 1)만

을 대상으로 내부 서순을 명료하게 배정할 수 있다.

그림 3은 기존의 문자픽셀 투영(pixel projection, 회색 화살표)과 본 연구의 OCR 검출박스 중심좌표 투영(point projection, 파란색 화살표)을 통한 라인분할 과정을 단적으로 비교한다. Pixel projection 기법은 4개 대표열 사이 3개 구간과 세주만으로 구성된 3열의 중앙에 대해 확연한 픽셀 저밀도 구간(< 밀도 문턱값, 회색체크마크)을 탐지한다. 하지만 2, 4열과 같이 본문과 세주가 혼합된 경우 각 세부열(본문, 우세주, 좌세주) 요소의 투영이 중첩되며 이를 분리하기 어렵다. 반면 본 연구의 point projection은 획기적으로 증가된 문자박스좌표 투영체의 해상도를 통해 명확한 군집화가 확인되며 군집간의 거리가 특정 문턱값 이상으로 벌어지는 곳(파란색 체크마크)에 라인분할을 수행할 수 있다.

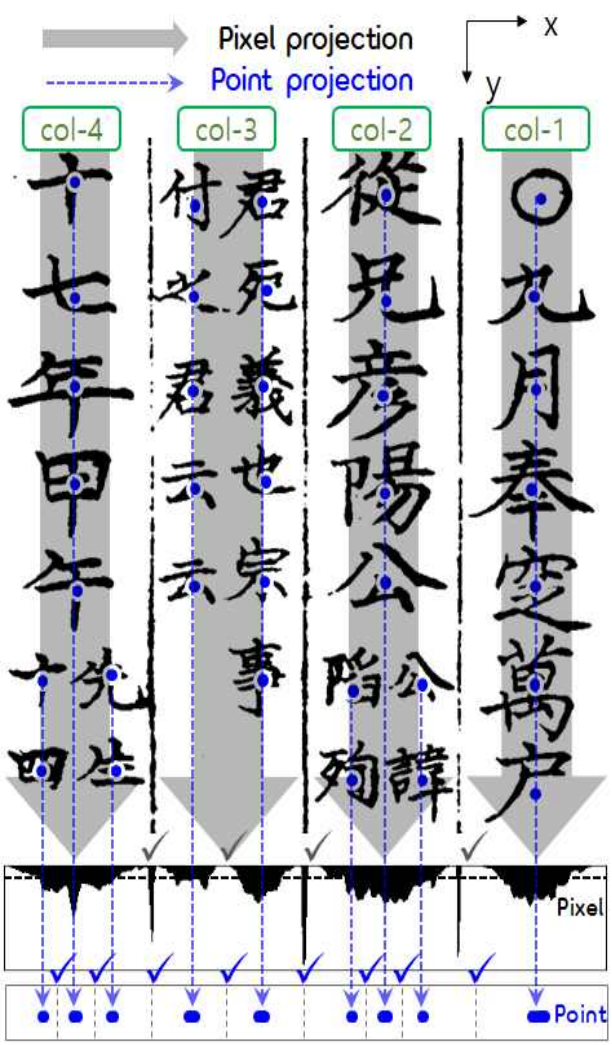

그림 3. 픽셀(회색)/포인트(파란색) 프로젝션 프로파일을 통한 라인분할 성능비교

Fig. 3. Comparison of line segmentation methods based on pixel(grey) and point(blue) projection profiles

Ⅳ. 한국 고문헌 라인분할 및 서순정렬 알고리즘

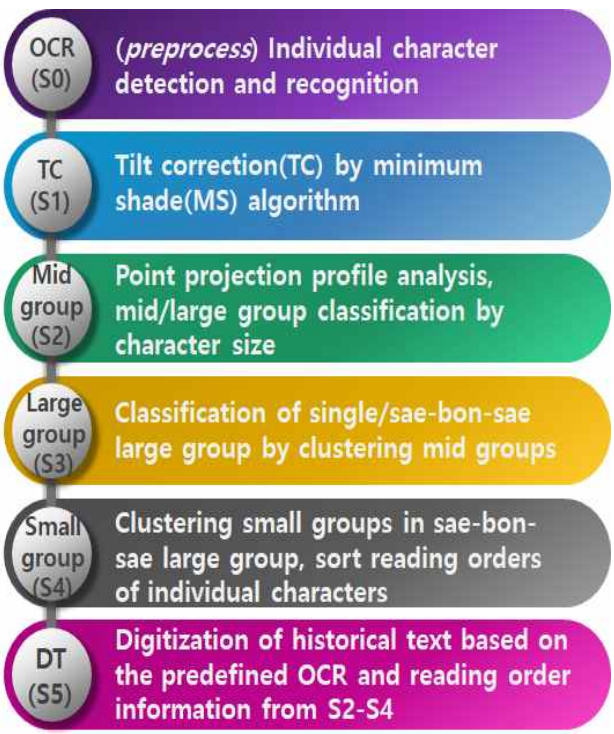

그림 4. 고문헌 라인분할 및 서순정렬 프로세스

Fig. 4. Process of line segmentation and reading order sorting in historical text

고문헌 원문이미지의 디지털 텍스트화를 위해 본 연구에서 개발된 알고리즘은 그림 4와 같이 사전에 개발된 YOLOv5 기반 OCR을 통한 개별한자 검출·인식 데이터 사전 준비(S0, OCR) 이후의 라인분할과 서순정렬을 포함한 총 5개의 과정으로 진행된다. 첫 번째로 원문이미지의 기울기 보정 전처리(S1, TC, tilt correction)를 통해 각 문자들의 중심좌표 정렬을 최적화한다. 각 보정좌표들은 그림 3에서 설명된 point projection을 통해 세부열별로 분리되며 글자 크기에 따라 본문열과 세주열로 구분된다 (S2, 중분류, mid group). 중분류들은 이웃간의 배치순서와 거리에 따라 단일 본문/세주 대분류(single large group, 그림 3의 1, 3열)로 구분되거나 우세주-본문-좌세주(세본세)의 대분류(sae-bon-sae large group, 그림 3의 2, 4열)로 군집화 된다 (S3, 대분류, large group). 세본세 대분류의 경우 그림 2의 최우측과 같이 수직방향을 따라 본문부와 세주부의 군집으로 재분할 된다 (S4, 소분류, small group). 상기 언급된 분할 요소들의 개략적인 구조는 그림 5와 같이 요약되며 각각의 분할 과정에 대한 세부 프로세스는 다음 챕터에서 설명된다. 마지막 단계인 (S5, DT, digital text)에서는 S2-S4에서 정의된 라인분할 체계에 따라 서순이 정렬된 한자들을 별도의 폰트/문장부호를 통해 본문/세주로 구분하여 디지털 텍스트화한다.

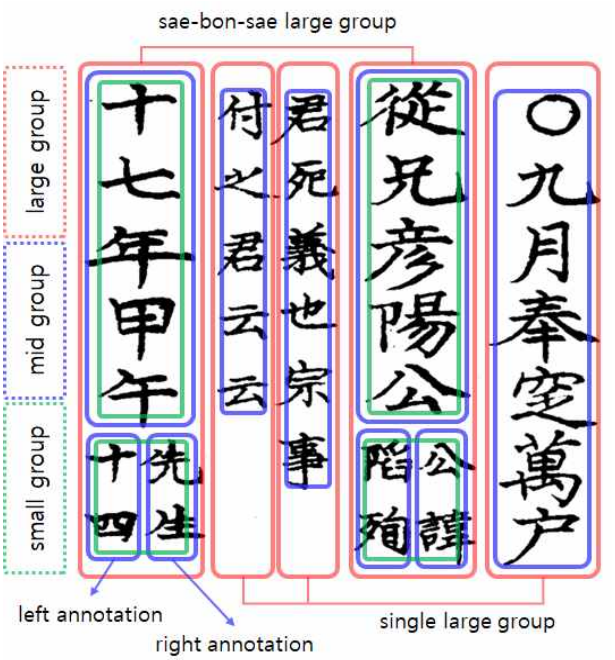

그림 5. 라인 분할 요소의 명칭 정의
Fig. 5. Definition of line segmentation components

4-1 (S1) : 고문헌 원문이미지 기울기 보정

본 연구에서 제안된 라인분할의 기본 알고리즘인 projection profile은 우종서 텍스트의 중심좌표들을 x축에 투영한 패턴을 분석하므로 이웃한 열 간의 겹침을 최소화하기 위해 종서의 정렬 방향이 x축 수직에 가까워야 한다. 하지만 고문헌은 그 제작과정에 있어서 필사되거나 사람의 손에 의해 판본되므로 현대의 전자문서와는 달리 기울어짐이 필연적으로 발생한다. 또한 이미지 원문화 과정에서 사용된 카메라나 스캐너에 대한 원본체의 오정렬 또한 상기 오차를 가중화 시킬 수 있다. 이로부터 파생되는 디지털텍스트화 프로세스의 오류를 감소시키기 위해 본 연구에서는 원문이미지의 기울기를 보정하는 전처리 알고리즘(TC; tilt correction)을 도입하였다.

TC 알고리즘은 그림 6과 같이 이미지 상단에서 하단을 향해 평행광(청록색 화살표)이 비춰지며 각 문자들이 피사체로 작용하여 이미지 하단에 그림자를 생성하는 시나리오를 가정한다. 이 경우 이미지의 기울어짐(θ)을 조정함에 따라 바닥면의 그림자 너비가 변하며 각 열의 정렬이 최적화 된 경우 그림 6(우)와 같이 최소 너비의 그림자(MS; minimum shade)가 생성된다. 본 알고리즘에서 피사체의 설정은 문자 전체를 구성하는 픽셀들이 아닌 그림 6의 빨간색 화살표와 같이 OCR로 검출된 문자 중심좌표를 좌우로 소폭(<문자 너비) 확장하는 형태를 사용한다. 확장이 없는 point projection만으로는 각 문자들의 투사체 겹침이 미미하며, 확장이 과도한 경우는 기존의 pixel projection 기반 기울기 보정[13]과 같이 이웃한 열과의 겹침이 분석의 정확성을 하락시킬 수 있다.

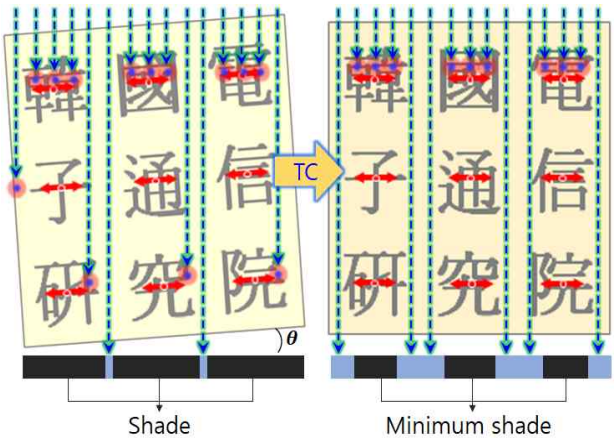

그림 6. 이미지 기울기 보정(TC) 알고리즘
Fig. 6. Image tilt correction(TC) algorithm

TC 대상 이미지는 수식 (1) 의 rotation matrix를 통해 중심좌표를 기준으로 회전되며 누적된 각 회전각도(θ)별 그림자 너비 그래프의 다항식 curve fitting을 통해 MS에서의 최적 회전보정각(θ_{opt})이 결정된다. 해당 회전보정각에 따라 OCR에서 검출된 문자 중심 좌표들도 모두 회전변환된 후 라인분할 알고리즘(S2)의 입력값으로 사용된다.

$$\begin{bmatrix}\cos\theta & -\sin\theta \\ \sin\theta & \cos\theta\end{bmatrix}\begin{bmatrix}x \\ y\end{bmatrix}=\begin{bmatrix}x' \\ y'\end{bmatrix} \quad (1)$$

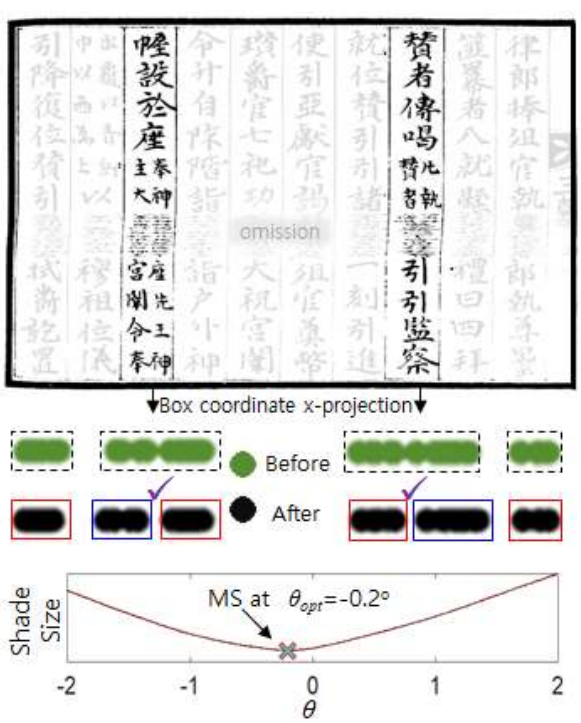

그림 7. 이미지 기울기 보정(TC)을 통한 열별 겹침 해소
Fig. 7. Resolving column overlap by image tilt correction

그림 7은 원문 이미지 기울기 보정 알고리즘(TC)의 적용 예시이다. 보정 전 세본세 대분류 내 개별 문자 박스좌표의 x 축 투영(연두색 점)은 본문열과 세주열 군집이 다소 밀접해있다. 물론 작은 문턱값을 사용할 경우 분리가 가능할 수도 있지만 동일한 열에 속한 문자들이 다른 열로 인식할 가능성 또한 높아진다. 반면 TC가 적용된 후의 x축 투영(검정색 점)은 보라색 체크마크 부분과 같이 세주열(빨간색 네모)과 본문열(파란색 네모)의 확연한 분리가 보인다. 그림 7에서 사용된 원문이미지의 경우 MS가 θ_{opt} = − 0.2°에서 나타났으며 이와 같은 작은 각도의 기울기보정 전처리도 projection profile 기반 라인분할에 유의미한 영향을 줄 수 있음이 확인되었다.

4-2 (S2) : 본문열/세주열 중분류 설정

기울기 보정을 거친 원문이미지 내의 검출박스들은 텍스트의 우종서 읽기 용이성을 위해 각각의 열로 분리되어야 한다. 정형성이 높은 고문헌의 경우 그림 3과 같이 대표열이 세로줄로 확연히 표시되어 있지만 보존상태가 좋지 않아 구분이 희미하거나 원 제작 시 경계가 표기되지 않은 경우도 있다. 그러므로 본 연구에서는 수직 구분선에 의존하지 않고 텍스트의 배치 양상만으로 열을 분리하는 과정을 포함한다. 본 챕터부터 예시로 보여지는 고문헌 이미지는 한국지능정보사회진흥원에 의해 AIHub에 공개된 “고서 한자 인식 (OCR)” DB의 KSAC_I_A08000260_001_020.jpg 파일을 2100 × 3000 픽셀로 리사이즈 후 기울기 보정(S1)을 거쳐 사용하였다.

그림 8 상단의 원문이미지 내 문자별 검출 좌표들은 x축에 프로젝션되어 파란색 점으로 표기된다. 이웃한 두 점간 간격이 기 설정된 중분류 문턱값 보다 클 경우 그 중간지점에서 열을 분리한다 (빨간 점선). 해당 문턱값은 모든 검출박스크기($\sqrt{w \times h}$)에 대한 평균값의 1/8배로 설정되었다. 각각의 분리된 군집은 본문이나 세주만을 포함한 중분류로 구분되며 그림 8의 상단과 같이 우측으로부터 번호가 배정된다. (해당 중분류 번호는 서순을 의미하지 않음). 각각의 중분류에 포함된 검출박스들의 크기 평균을 구한 뒤 그림 8의 좌하단과 같이 분포를 이분화하면 본문/세주 중분류의 구분이 가능하다. 본문열과 세주열은 각각 7, 16개로 판별되었으며 군집 테두리에 분홍색과 연두색으로 표기되었다. 중분류별 평균이 아니라 단순히 텍스트 전체의 검출박스 크기에 대한 본문/세주 이분화 또한 가능하지만, 이 경우 평균치를 다소 벗어난 소수 불특정 박스에 대해 오판이 발생할 가능성이 높다.

상기 본문/세주의 분류는 그림 8의 우하단과 같이 본문만으로 이루어진 이미지에 대한 예외처리가 필요하다. 이를 위해 중분류를 이분화하기 이전에 각 중분류별 박스크기 평균의 상대표준편차(RSD)가 분포 문턱값(본 연구의 경우 0.05로 설정)보다 작을 경우 이분화과정 없이 모든 텍스트를 본문 중분류로 설정한다.

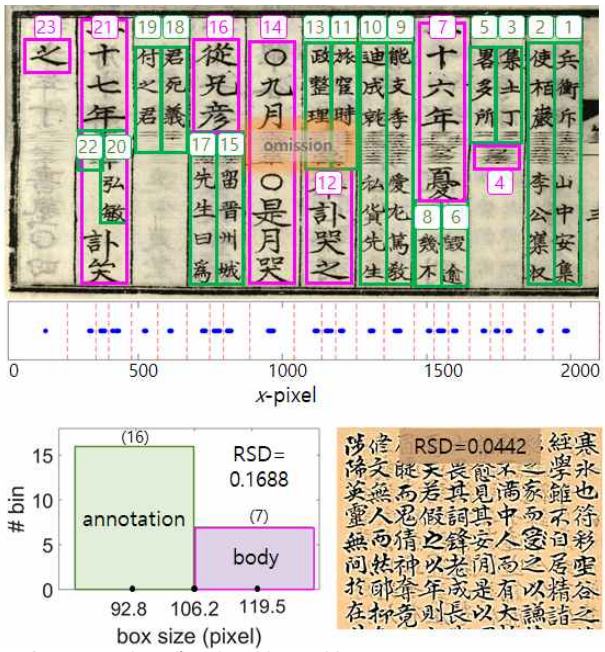

그림 8. 본문열 / 세주열 중분류 구분
Fig. 8. Mid-group classification of main/annotation columns

4-3 (S3) : 중분류 군집화를 통한 대분류 설정

S2에서 배정된 중분류들의 종류(본문열/세주열)를 색으로 표현하여 각 중분류의 중심좌표에 따라 x축에 투영 시 그림 9 하단과 같은 결과를 확인할 수 있다. 이러한 중분류들의 나열은 본문/세주만으로 포함된 단일 대분류와 우세주-본문-좌세주의 조합으로 구성된 세본세 대분류로 군집화 될 수 있다. 세본세 대분류 군집화의 조건은 다음과 같다. 조건 1) 본문중분류 좌우 모든 방향에 세주중분류들이 이웃함, 조건 2) 본문중분류 중심으로부터 세본세 문턱값 이내의 거리(그림 9 하단, 분홍색 화살표)에 좌우측 세주중분류가 모두 포함됨. 본 예제에서는 조건 2의 문턱값으로 이미지 내 본문에 해당하는 모든 검출박스 크기의 평균을 사용하였으며 5개의 세본세 대분류(파란색 실선)들을 정상적으로 군집화 함이 확인되었다. 그림 9의 하단에 파란색 점선으로 표기된 군집은 세본세 조건 1을 만족시키지만 조건 2에 의해 세본세 분류에서 제외되었으며 상단 그림에서도 해당 점선 내 중앙 본문(8번)이 단일 본문 대분류로 판정되었음이 확인된다.

세본세 군집화에 포함되지 않은 나머지 모든 중분류들은 단일 대분류로 배정된다. 대분류간의 서순은 그림 9 상단의 순번(빨간색 숫자)과 같이 우→좌의 순서를 따르며 서로간의 내부 서순에 영향을 주지 않는 독립적인 개체이다. 대분류의 설정은 앞서 원문이미지 내 수직선으로 분리된 대표열(그림 3)의 설정과 비슷하지만 그림 3의 3열과 같이 세주열 2개만으로 구성된 대표열의 경우 세주열 각각을 분리된 대분류로 설정함에서 차이가 있다.

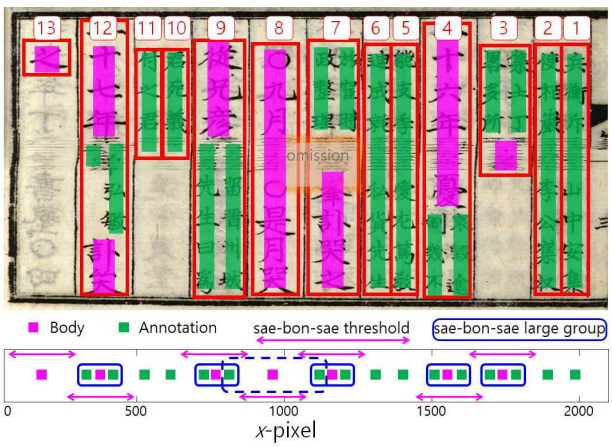

그림 9. 중분류 군집화를 통한 대분류 설정
Fig. 9. Large-group assignment by mid-group clustering

4-4 (S4) : 소분류 설정 및 대분류 내 서순배정

본문 또는 세주 요소만을 포함한 단일 대분류 내의 서순은 단순히 y좌표에 따라 상→하 순서로 배정할 수 있다. 하지만 세본세 대분류 내의 서순배정은 그림 2의 우측에서 표현된 것처럼 추가적인 소분류 구분의 과정이 필요하다.

그림 10의 예제는 소분류 구분 과정의 설명을 위해 원문이미지 내 세본세 대분류만을 다룬다. 소분류의 경계는 각 대분류 내에서 모든 문자별 검출 좌표들의 y축 투영(그림 10의 각 대분류 우측 점으로 표기) 후 본문→세주 또는 세주→본문의 전환이 이루어지는 곳에 설정(연두색 화살표)되며 위아래로 분리된 구획들을 각각의 소분류로 배정한다.

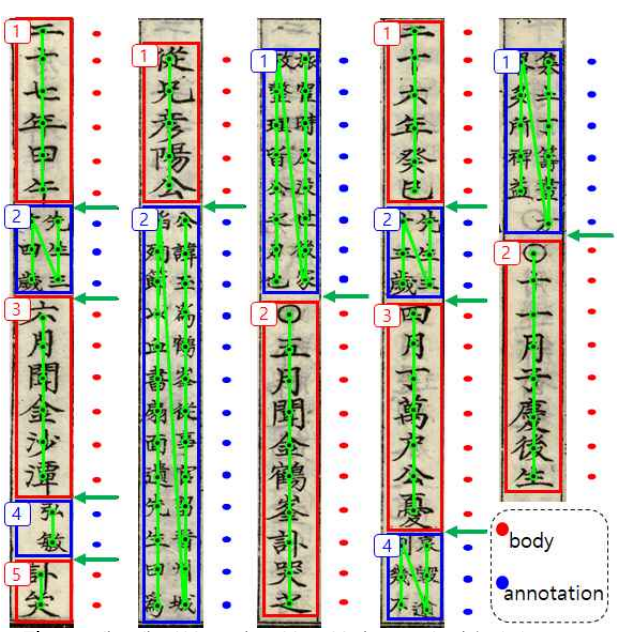

그림 10. 세본세 대분류 내 소분류 분할을 통한 서순배정
Fig. 10. Reading order sorting in sae-bon-sae large-group by small-group division

각 대분류 내 소분류들의 서순은 그림 10에 표기된 순번들처럼 상→하의 순서를 따른다. 파란색 네모로 표시된 세주 소분류들의 경우 연두색 실선으로 표기되었듯이 우세주 내부를 상→하의 순서로 먼저 읽은 후 좌세주로 이동한다.

4-5 (S5) : 원문 이미지 전체에 대한 서순배정 및 결과 확인

OCR 데이터를 기반으로 S1-S4의 라인분할 과정에 걸쳐 소·중·대분류 군집화를 통해 고문헌 원문이미지 개별 한자들의 서순 배정이 완료되었다. 한문 고서의 텍스트 서순배정 법칙은 다음과 같이 요약된다.

1) 대분류는 우→좌의 순서로 배정
2) 세본세 대분류 내 소분류는 상→하의 순서로 배정
3) 세주 소분류는 우세주→좌세주 순서로 배정
4) 이하 명시되지 않은 개별 하위 문자들의 서순은 우종서의 기본 법칙을 따름

그림 11은 원문 이미지에 적용된 본 알고리즘의 실사용 예를 보여준다. 그림 11(좌)의 원문 이미지는 라인분할에 사용된 S0 단계에서의 OCR 검출 결과를 나타낸다. S2단계에서 설정된 본문/세주 중분류 정보를 기반으로 본문은 빨간색, 세주는 파란색 박스로 표기되었으며, OCR에서 미검출 된 자형은 박스로 구분되지 않았다. 그림 11의 우상단은 디지털 텍스트화(DT)를 거쳐 최종적으로 출력된 원문텍스트 결과이다. 서순에 따라 원문이미지에서 추출된 자형들은 본문일 경우 빨간색, 세주일 경우 괄호내 파란색 폰트로 표기되었으며 각 대분류는 행분리로 구분되었다. OCR 모델 학습에서 라벨링 되지 않은 자형으로 인식된 경우 '?'로 처리 되었으며 OCR의 제한적 한자 인식 정확도로 인해 소수의 자형에 대한 오인식을 포함한다. 미검출 및 오인식 자형에 대한 OCR 자체의 오류는 후속 연구 과정에서 DB 추가 확보 및 인공지능 모델의 개선, API 기반 감수 시스템을 통해 개선될 예정이다.

그림 11의 예제에서 생성된 디지털 텍스트 결과물은 정답지(그림 11 우하단)와 비교를 거쳐 OCR 검출 정확도 (검정+분홍)/(검정+분홍+연두)×100%=93.8%, 인식 정확도 (검정)/(검정+분홍)×100%=84.3%로 평가되었다. 미검출과 오인식 사례는 정답지 내에 각각 연두색과 분홍색으로 표기되었다. 또한 디지털 텍스트 결과물의 서순은 2장에서 정의된 우종서 기반 본문-세주 혼합 구조의 서순을 정확히 반영하여 정렬됨이 정성적으로 확인되었다.

기존 서순정렬 기술과의 차별성은 그림 11의 디지털 텍스트 4번 행(연두색 네모, 笑已先生三十三歲四月)에서 확인할 수 있다. 대표적인 최신 한문고서 DB인 AIHub의 "고서 한자 인식(OCR), 2022"에서는 본문열과 세주열의 구분이 불가능하여 [12]의 기술과 같이 모든 개별 문자에 동등하게 우종서 서순을 적용하므로 상기 동일구문이 先生三...癸巳四月...十三歲의 서순으로 오정렬되어 있다. 이러한 서순오류 사례는 해당 DB에서 매 본문-세주 혼합부마다 발견되며 본 연구의 본문/세주 분리 분석 알고리즘은 해당 건의 대안을 제시할 수 있다.

그림 11. 고문헌 원문 이미지 디지털 텍스트화 최종 결과 예시
Fig. 11. Digitization example of scanned historical text image

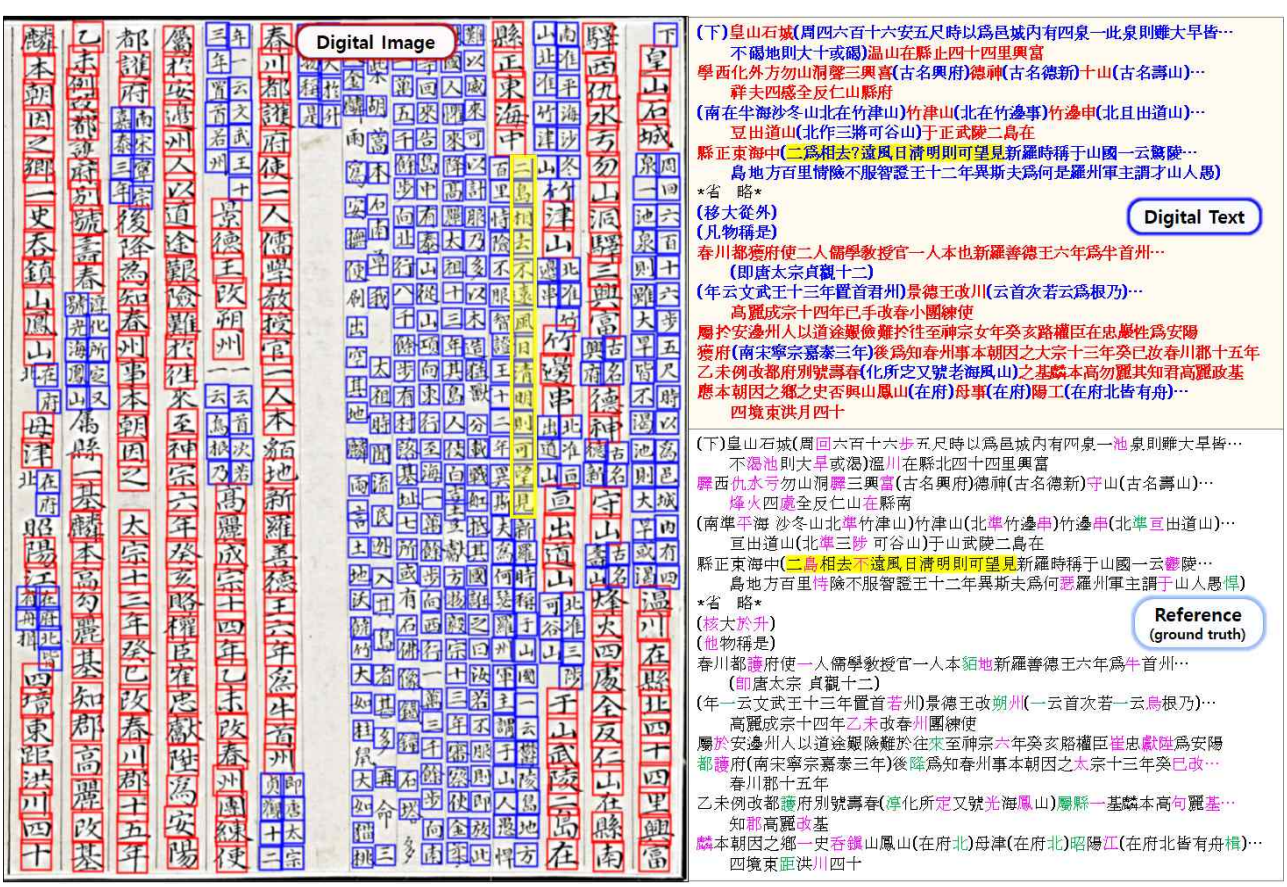

그림 12. 고문헌 원문 이미지 디지털 텍스트화 활용 예시 (세종실록 지리지), (노란색 음영은 독도 영유권 주장 근거 표기)
Fig. 12. Application example of historical text digitization (Geography Section of the Annals of King Sejong)

정량적 서순정렬 성능평가에 관해서 layout analysis 기반의 라인분할 성능평가 방식이 [14]에서 제안되었지만 세주 구조로 인한 열 분리가 빈번한 한문고서를 대상으로는 적용될 수 없었다. 또한 디지털 텍스트와 정답지 내 자형들을 일렬로 나열하여 매칭 여부를 수량화하는 방법이 있지만, 이는 사전과정인 OCR의 자체 성능에 큰 영향을 받으므로 서순정렬의 단독적 평가지표로 보기 어렵다 [15]. 따라서 한문 고서 디지털 텍스트 생성에 관한 후속 연구들의 확장과 성능비교를 위해서는 새로운 평가 전략이 필요한 실정이다.

본 연구결과의 활용성은 다양한 고서에 대해 검증되었으며, 대표적인 예로 그림 12는 세종실록 지리지(153권 011a면, 강원도 삼척도호부 울진현, 국사편찬위원회)에 대한 활용사례를 보여준다. 앞선 그림 11의 결과와 마찬가지로 원문이미지 내 모든 자형들의 본문/세주 구분이 이루어졌으며 서순 또한 정확히 배정됨이 확인되었다. 해당 원문이미지에 대한 OCR의 검출/인식 정확도는 각각 95.5%, 81.4%로 나타났다.

V. 결론

대다수의 기존 라인분할 알고리즘들은 원문이미지 전체를 대상으로 분할 전처리를 적용한 후, 각각의 영역들에 대해 OCR을 적용하는 top-down 기반의 디지털 텍스트화를 선보였다. 반면 본 연구에서는 이미지 내 전체 텍스트에 OCR을 선적용(bottom-up) 한 뒤 검출박스들의 좌표를 대상으로 projection profile을 분석(top-down)하여 군집화 및 서순을 배정하였다. 이러한 hybrid 기반의 라인분할은 기존 OCR에서 제공되는 문자 검출 및 인식 정보의 잠재력을 한층 더 활용하면서 고문헌 원문텍스트 제작의 효율성을 증가시킬 수 있다.

본 연구에서 제안된 라인분할 및 텍스트 서순배정의 정확도는 원문이미지 내 자형들의 수직방향 정렬에 의존하며, 전처리 과정에 추가된 이미지 기울기 보정(TC) 알고리즘은 열별 공간 확보를 통해 전체 디지털 텍스트화의 신뢰도를 증가시켰다. 후속 연구에서는 해당 전처리 과정에서 디지털 이미지 생성 중 발생하는 warping까지 대응하도록 알고리즘을 확장할 예정이다.

본 연구개발 최종 검증단계에서 디지털 텍스트화 결과, 본문과 세주가 혼재된 원문 이미지 내에서도 성공적으로 라인을 분할하고 개별 자형의 본문/세주 분류 및 서순의 배정이 가능함이 확인되었다. 소수의 누락 또는 오인식된 자형이 존재하지만 이는 본 연구 외적인 OCR의 검출과 인식 오류에 기인한 것으로, 지속적인 연구를 통해 성능이 개선될 것으로 전망된다.

상기 개발된 고문헌 원문 이미지로부터의 디지털 텍스트 추출 기술은 고문헌 관리과정 중 교감대조/이본대조 등의 자동화를 지원할 수 있으며, 궁극적으로 코퍼스 DB 구축 과정에 힘입어 원문의 다국어 자동번역까지 전자동화가 가능한 시스템 구축으로 기술 확장이 가능하다.

감사의 글

본 연구는 문화체육관광부 및 한국콘텐츠진흥원의 연구개발지원사업으로 수행되었음. (과제번호 R202104267, 인공지능 기반 개방형 한문고서 번역 및 해석 지원 기술 개발)

참고문헌

[1] S.-S. Kim, A Study on the Mid- to Long-term Development of the Management of Korean Historical Texts, *National Library of Korea*, Seoul, 11-1371029-000171-01, pp. 1-170, Oct. 2018.

[2] K. Yang and D. Shin, “Analysis of Korean Traditional Records Information System,” *J. Kor. Lib. and Inf. Sci.*, Vol. 47, No. 4, pp. 191, Dec. 2016. https://doi.org/10.16981/kliss.47.4.201612.191

[3] J. Memon, M. Sami, R. A. Khan, and M. Uddin, “Handwritten Optical Character Recognition (OCR): A Comprehensive Systematic Literature Review (SLR),” *IEEE Access*, Vol. 8, pp. 142642, July 2020. https://doi.org/10.1109/ACCESS.2020.3012542

[4] A. Baldominos, Y. Saez, and P. Isasi, “A Survey of Handwritten Character Recognition with MNIST and EMNIST,” *Appl. Sci.*, Vol. 9, No. 15, pp. 3169, Aug. 2019. https://doi.org/10.3390/app9153169

[5] G. Min, A. Lee, and H. S. Kang, “A Study on preprocessing for AI based Chinese character segmentation in historical documents,” *2021 Korean Institute of Commucations and Information Sciences (KICS)*, Yeosu, pp. 597, Nov. 2021.

[6] A. Jalali and M. Lee, “High cursive traditional Asian character recognition using integrated adaptive constraints in ensemble of DenseNet and Inception models,” *Pattern Recognit. Lett.*, Vol. 131, pp. 172, Aug. 2020. https://doi.org/10.1016/j.patrec.2020.01.013

[7] I. H. Kim, “The Department of State Exclusive Economic Zone in the Dokdo Island,” *Association of Japanology in East Asia*, Vol. 21, pp. 35, Jan. 2007. http://doi.org/10.18075/jcs..21.200701.35

[8] R. Ptak, B. Zygadlo, and O. Unold, “Projection-based text line segmentation with a variable threshold,” *Int. J. Appl. Math. Comput. Sci.*, Vol. 27, pp. 195, Oct. 2016. http://dx.doi.org/10.1515/amcs-2017-0014

[9] W. Ma, H. Zhang, L. Jin, S. Wu, and J. Wang, “Joint Layout Analysis, Character Detection and Recognition for Historical Document Digitization,” *2020 ICFHR*, pp. 31, Sep. 2020. https://doi.org/10.1109/ICFHR2020.2020.00017

[10] O. Mechi, M. Mehri, R. Ingold, and N. E. B. Amara, "A Text Line Extraction Method for Archival Document Transcription," *17th International Multi-Conference on Systems, Signals & Devices (SSD)*, Amman, pp. 479, July 2020. https://doi.org/10.1109/SSD49366.2020.9364163

[11] B. K. Barakat, et al., "Unsupervised deep learning for text line segmentation," *25th International Conference on Pattern Recognition (ICPR)*, Milan, pp. 2304, Jan. 2021. http://dx.doi.org/10.1109/ICPR48806.2021.9413308

[12] J. Ryu, et al., "Numbering of detection results for translation and detection of chinese cursive character in ancient documents," *Institute of Control, Robotics and Systems*, Gyeongju, pp. 140, May 2019.

[13] S. Li, Q. Shen, and J. Sun, "Skew detection using wavelet decomposition and projection profile analysis," *Pattern Recognit. Lett.*, Vol. 28, pp. 555, Apr. 2007. https://doi.org/10.1016/j.patrec.2006.10.002

[14] C. Clausner, S. Pletchacher, and A. Antonacopoulos, "Scenario Driven In-Depth Performance Evaluation of Document Layout Analysis Methods," *12th International Conference on Document Analysis and Recognition (ICDAR)*, Beijing, pp. 1404, Sep 2011. https://doi.org/10.1109/ICDAR.2011.282

[15] C. Clausner, S. Pletschacher, and A. Antonacopoulos, "Flexible character accuracy measure for reading-order-independent evaluation," *Pattern Recognit. Lett.*, Vol. 131, pp. 390, Feb. 2020. https://doi.org/10.1016/j.patrec.2020.02.003

이아람(Aram Lee)

2010년 : 미시시피주립대학교 (전자공학과, 공학사)
2013년 : 버지니아공대 (전자공학과, 공학석사)
2016년 : 버지니아공대 (전자공학과, 공학박사)

2021년～현 재: 한국전자통신연구원, 연구원
※관심분야 : 광학문자인식, 패턴인식, 영상신호처리

민기현(Gihyeon Min)

2001년 : 충북대학교 (물리학과, 이학사)
2003년 : 충북대학교 (물리학과, 이학석사)
2014년 : 광주과학기술원 (정보통신공학과, 공학박사)

2014년～현 재: 한국전자통신연구원, 선임연구원
※관심분야 : 이미지신호처리, 지능형 영상분석, 광학문자인식

김거식(Keo Sik Kim)

2004년 : 전북대학교 (전자공학과, 공학사)
2006년 : 전북대학교 (전자공학과, 공학석사)
2011년 : 전북대학교 (전자공학과, 공학박사)

2011년～현 재: 한국전자통신연구원, 책임연구원
※관심분야 : 머신러닝, 광이미징, 영상 신호처리

김정은(Jeong Eun Kim)

2002년 : 영남대학교 (물리학과, 이학사)
2004년 : 한국과학기술원 (물리학과, 이학석사)
2011년 : 베를린공과대학교 (물리학과, 이학박사)

2012년～현 재: 한국전자통신연구원, 선임연구원
※관심분야 : 사물인식 딥러닝, 영상 신호처리

강현서(Hyun Seo Kang)

1994년 : 성균관대학교 (전자공학과, 공학사)
1996년 : 성균관대학교 (전자공학과, 공학석사)
2000년 : 성균관대학교 (전자공학과, 공학박사)

2001년～현 재: 한국전자통신연구원, 책임연구원(실장)
※관심분야 : AI 센서, 비전인식, 문자인식, 사물인터넷