논문 2021-2-16 http://dx.doi.org/10.29056/jsav.2021.12.16

양식 문서 영상에서 도표 구조 분석을 위한 라인 추적 알고리즘

김계경*†

Line Tracking Algorithm for Table Structure Analysis in Form Document Image

Kye-Kyung Kim*†

요 약

도표로 작성된 양식 문서에서 도표의 레이아웃 해석에 필요한 그리드 라인을 추출하기 위해 다양한 필터링 또는 모폴로지 등의 방법을 사용하여 직선 성분을 선명하게 개선시키기 위한 연구들이 많이 진행되고 있다. 도표의 직선 성분을 선명화하더라도 직선 내부에 절단 점들이 존재하거나 기울어진 경우에는 직선 추출이 어렵고 도표 셀들의 레이아웃을 논리적으로 표현하는데 여전히 어려움을 겪을 수 있다.

본 연구에서는 직선에 절단점들이 존재하거나 기울어져도 직선을 검출할 수 있는 라인 추적 알고리즘을 제안하였다. 이를 이용하여 그리드 라인을 추출하고 라인들의 교차점 및 셀 정보들을 찾아 도표의 구조를 분석할 수 있는 알고리즘을 마련하였다. 제안한 알고리즘을 실제 양식 문서 영상을 대상으로 실험한 결과 평균 0.41초 처리시간에 96.4%의 도표 구조를 분석할 수 있음을 확인하였다.

Abstract

To derive grid lines for analyzing a table layout, line image enhancement techniques are studying such as various filtering or morphology methods. In spite of line image enhancement, it is still hard to extract line components and to express table cell's layout logically in which the cutting points are exist on the line or the tables are skewing .

In this paper, we proposed a line tracking algorithm to extract line components under the cutting points on the line or the skewing lines. The table document layout analysis algorithm is prepared by searching grid-lines, line crossing points and gird-cell using line tracking algorithm. Simulation results show that the proposed method derive 96.4% table document analysis result with average 0.41sec processing times.

한글키워드 : 양식 문서, 도표 셀, 도표 해석, 라인 추적, 도표 레이아웃

keywords : form document, table cell, table analysis, line tracking, table layout

* 한국전자통신연구원

† 교신저자: 김계경(email: kyekyung@etri.re.kr)

접수일자: 2021.11.30. 심사완료: 2021.12.08.
게재확정: 2021.12.20.

1. 서 론

관공서, 금융권 또는 기업체 등 다양한 기관에

서 도표로 이루어진 양식 문서를 이용하여 자료 조사, 통지, 계약, 또는 증명을 위한 문서 작업을 수행하고 있다. OCR을 이용하여 이러한 양식 문서의 내용을 인식하는 업무를 처리할 경우 문서 영상을 스캐닝한 다음 문서를 구성하고 있는 도표를 검출하고 도표의 레이아웃을 분석하여 그리드 셀들의 위치를 알아낸 다음 각 셀에 포함된 문자들을 세그멘테이션하는 과정이 필수적으로 요구된다[1,2].

양식 문서를 대상으로 문서 영상 파일을 생성하기 위해 디지털 장치로 스캐닝할 경우 문서가 기울어지거나 그림자가 유입될 수 있고 정보를 포함하는 픽셀들이 소실되어 그리드 라인들의 직선 성분에 절단된 부분이 발생할 수도 있다. 양식 문서를 디지털화 할 때 발생되는 이러한 문제들은 양식 문서 영상의 도표 해석을 힘들게 하는 중요한 요인이 될 수 있다. 양식 문서 영상에서 그리드 라인의 일부가 절단 되거나 그리드 라인들의 교차점 일부가 소실된 경우에도 이를 해석하여 그리드 라인의 교차점을 찾고 그리드 셀들의 위치 및 상호 연결 관계를 해석하기 위해 다양한 필터링 방법이나 영상 모폴로지 방법들이 사용되고 있다[1-4]. 다양한 방법으로 문서 영상을 개선시켜 도표의 직선 성분을 강조하더라도 직선상에 소실된 픽셀들이 여전히 존재할 경우 복잡한 형태의 도표에서 모든 그리드 라인을 정확하게 검출하거나 도표 셀들의 구조를 논리적으로 표현하는 것이 어려운 작업일 수 있다[5-8].

본 논문에서는 양식 문서에 포함된 도표를 해석하기 위해 그리드 라인을 추적하는 방식으로 직선 성분을 찾아내고 그리드 셀의 레이아웃 정보를 검출할 수 있는 알고리즘을 제안하였다. 수평 또는 수직 형태의 그리드 라인이 절단되거나 기울어져 있어도 그리드 라인의 시작점에서 끝점까지 픽셀들을 연속해서 추적할 수 있는 알고리즘을 마련하여 그리드 라인의 위치를 정확히 찾아낼 수 있도록 하였다. 추적한 그리드 라인들의 교차점을 해석하여 그리드 셀의 위치와 레이아웃을 표현하는 방법도 마련하여 도표 해석 결과 각 그리드 셀 별로 문자들을 세그멘테이션하여 OCR 등에 적용할 수 있도록 하였다.

제안한 방법을 이용하여 관공서 및 은행 등에서 사용되는 양식 문서 영상을 대상으로 실험용 데이터베이스를 구축하고 도표 해석 실험을 수행한 다음 그 결과를 분석하였다.

2. 양식 문서 도표 구조 분석 개요

문서 영상에 포함된 도표 레이아웃을 해석하여 그리드 셀 정보를 찾아내는 알고리즘의 흐름도를 그림 1에 도시하였다.

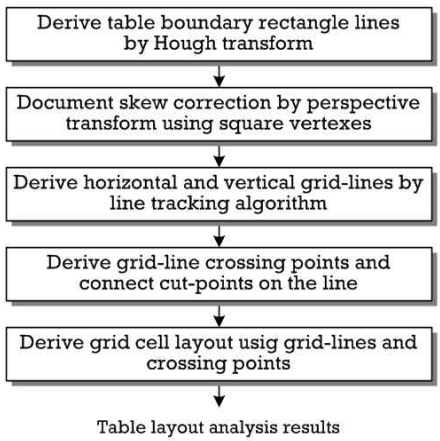


그림 1. 문서 영상에서 도표 구조 분석 알고리즘 흐름도

Fig. 1. The flow-chart of table structure analysis algorithm in document image

그림 1의 알고리즘에서는 먼저 문서를 스캐닝하여 문서 영상을 생성했을 때 문서의 형태가 기울어질 수 있으므로 이를 보정하는 작업을 수행한다. 다음 단계에서 라인 추적 알고리즘을 이용

하여 수평 및 수직 그리드 라인들의 시작점의 위치를 탐색하고 절단된 부분을 연결하면서 라인의 종료점까지 추적한다. 그리드 라인들을 추출한 후 교차점들을 찾고 도표 그리드 라인 규칙을 적용하여 잡음으로 추출된 라인들을 제거한다. 마지막으로 교차점들의 위치 특성, 즉 시작점 중간점 종료점 등을 해석하고 이를 이용하여 그리드 셀의 레이아웃 정보를 추출한다. 도표 구조 분석을 위한 알고리즘 각 단계별로 생성한 문서 영상의 예를 그림 2에 도시하였다. 문서 영상의 기울기를 보정하고 그리드 라인과 교차점들을 추출하는 과정과 그리드 셀의 레이아웃을 해석하고 문자들을 세그멘테이션한 예제를 도시하였다.

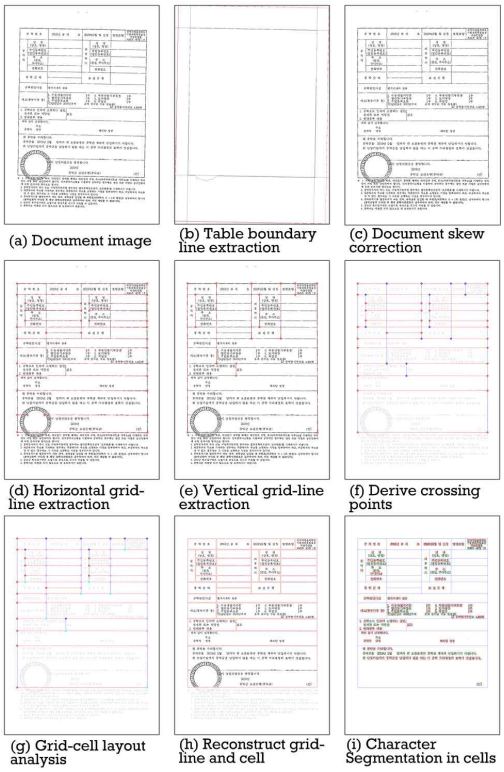


그림 2. 도표 구조 분석 각 단계별로 생성한 문서 영상

Fig. 2. Document image samples for each step of table structure analysis

3. 양식 문서 도표 구조 분석 알고리즘

3.1 양식 문서 회전 보정

양식 문서 영상에서 도표의 형태가 기울어져 있을 경우 그리드 라인 및 그리드 셀을 탐색하는데 오차가 발생할 수 있으므로 문서 영상 회전 보정 작업을 수행한다. 그림 3에 기울어진 양식 문서를 보정하는 단계별 영상들을 예시하였다.

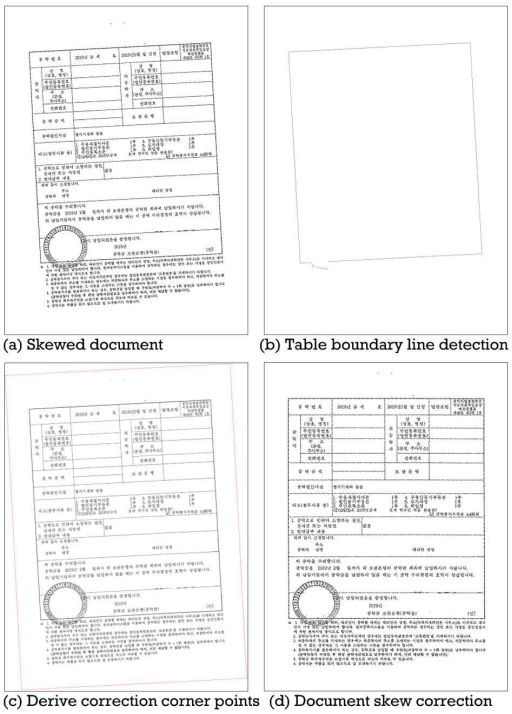


그림 3. 문서 기울기 보정 단계별 영상 예시

Fig. 3. Example images of document skew correction steps.

그림 3에서 예시한 것처럼 양식 문서의 회전 보정을 위해 허프 변환[9,10]을 이용해서 도표를 구성하는 4개의 외접 직선들을 구하고 이를 문서의 가장자리 끝 부분까지 확장하여 회전 보정에 적용할 기울어진 형태를 표현하는 보정용 4개 꼭지점, A, B, C, D들의 좌표(x, y)를 구한다. 보정용 4

개의 꼭지점 좌표와 보정결과 생성할 문서 영상의 꼭지점, A', B', C', D'들과 꼭지점의 좌표 (x', y')를 Perspective Transform 알고리즘[6]에 적용하면 회전 보정된 영상을 구할 수 있다.

$$x' = ax + by + cxy + d \quad (1)$$
$$y' = ex + fy + gxy + h \quad (2)$$

식 (1) 및 (2)에서 회전 보정용 영상의 4개의 꼭지점과 보정결과 영상의 4개 꼭지점들을 각각 수식에 대입해서 도출한 8개 방정식을 풀면 상수 a, b, c, d 및 e, f, g, h 의 값을 구할 수 있다. 이를 이용하여 보정 결과 영상의 각 픽셀이 보정 대상 영상에서 어느 위치인지를 구하는 역방향 변환으로 회전 보정된 영상을 생성할 수 있다.

3.2 그리드 라인 추적 알고리즘

수평 또는 수직 형태 그리드 라인들의 시작점과 종료점을 정확히 찾고 이를 이용하여 그리드 라인의 교차점과 그리드 셀의 레이아웃을 해석한다. 수평 그리드 라인 추적 예를 그림 4에 도시하였다.

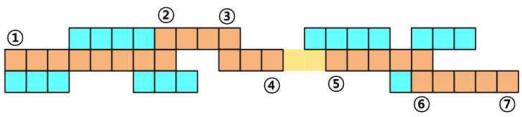


그림 4. 절단된 수평 그리드 라인 추적 예
Fig. 4. A separated horizontal grid line tracking example

그림 4에 도시한 수평 그리드 라인 추적에서는 먼저 좌측 픽셀들을 스캐닝해서 좌측 시작점(①)을 찾는다. 좌측 시작점과 동일한 위치에 위쪽 또는 아래쪽에 픽셀이 존재할 경우 우측으로 가장 긴 연결선을 갖는 픽셀을 시작점으로 선택한다. 우측으로 스캐닝하면서 연결이 끊어진 픽셀을 만나면 위쪽과 아래쪽 픽셀을 찾아 더 긴 연결선을 갖는 픽셀을 새로운 시작점(②,③,⑥)으로 선택하고 우측으로 계속 스캐닝한다. 더 이상 연결 선을 찾을 수 없을 경우(④), 미리 정한 절단 허용 범위 내에 연결 픽셀이 존재하는지 탐색하고 만일 범위 내에 픽셀(⑤)이 존재할 경우 라인 추적을 이어간다. 만일 절단 허용 범위 내에 픽셀이 존재하지 않을 경우(⑦) 해당 픽셀을 수평 라인의 종료점(⑦)으로 판단하고 라인 추적을 끝낸다.

수평 그리드 라인 H_m과 수직 그리드 라인 V_n을 형성하는 픽셀 $pel(x,y)$들의 집합을 얻기 위한 그리드 라인 추적 알고리즘을 다음과 같이 정의할 수 있다.

$$H_m = \bigcup_{i=x_s}^{x_e} \{ pel(i,y) | i \equiv S_h, y \equiv T_h \} \quad (3)$$

where,

$$S_h \in \begin{cases} pel(i+1,y) \\ \max(pel(i,y-1)|_{szie}, pel(i,y+1)|_{size}) \\ pel(i+dist,y) \end{cases} \quad (4)$$

$$T_h \in \begin{cases} pel(i+1,y) \\ pel(i,y-1) \\ pel(i,y+1) \end{cases} \quad (5)$$

$$V_n = \bigcup_{j=y_s}^{y_e} \{ pel(x,j) | x \equiv S_v, j \equiv T_v \} \quad (6)$$

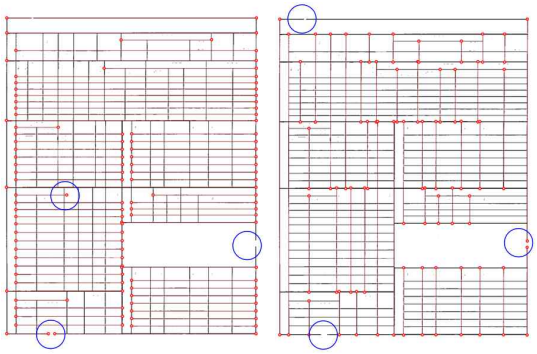
where,

$$S_v \in \begin{cases} pel(x,j+1) \\ pel(x-1,j) \\ pel(x+1,j) \end{cases} \quad (7)$$

$$T_v \in \begin{cases} pel(x,j+1) \\ \max(pel(x-1,j)|_{szie}, pel(x+1,j)|_{size}) \\ pel(x,j+dist) \end{cases} \quad (8)$$

식 (3) 및 (6)은 그리드 라인 추적 알고리즘으로 찾은 m번째 수평 그리드 라인 H_m 및 n번째 수직 그리드 라인 V_n을 이루는 각각의 픽셀 집합을 표현한 것이다. 식 (4) 및 (5)는 수평 그리

드 라인의 우측 및 상하 우선순위별 픽셀 탐색 규칙을 표현한 것이고 식 (7) 및 (8)은 수직 그리드 라인의 아래 및 좌우 우선순위별 픽셀 탐색 규칙을 표현한 것이다. 식에서 $dist$는 수평 또는 수직 라인의 최대 탐색 허용 절단 픽셀 개수를 의미한다. 일반적으로 도표의 최소 셀 크기보다 작은 값으로 설정한다. $pel(x,y)|_{size}$는 우측으로 연결된 최대 픽셀 개수이다. 전술한 그리드 라인 추적 알고리즘을 이용하여 수평 및 수직 그리드 라인들을 검출한 예를 그림 5에 도시하였다.



(a) 수평 그리드 라인 (b) 수직 그리드 라인

그림 5. 도표 영상에서 절단된 수평 및 수직 그리드 라인을 검출한 예

Fig. 5. Examples of the separated horizontal and vertical grid-line extraction in table image

그림 5에서 테이블의 그리드 셀을 구성하는 모든 그리드 라인들이 검출된 예를 볼 수 있다. 그리드 라인이 절단 되었을 때 $dist$ 이내의 길이는 자동으로 연결되고 $dist$ 이상의 길이는 종료점으로 표현되었다. 그리드 라인이 절단되어 분할된 경우에는 그리드 라인 교차점 해석 과정에 도표의 특성을 반영하여 하나의 라인으로 결합 되도록 하였다.

3.3 교차점 추출을 통한 셀 레이아웃 분석

그리드 라인 추적 알고리즘을 이용하여 추출한 수평 및 수직 그리드 라인들의 시작점과 종료점들을 이용하여 테이블을 구성하고 있는 그리드 라인들의 교차점을 모두 구하고 수평 수직 라인의 시작점, 중간점 및 종료점 등의 위치 특징을 그림 6과 같이 9개의 종류로 분류한다.

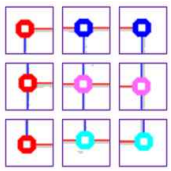


그림 6. 수평 및 수직 그리드 라인의 교차점

Fig. 6. Crossing points of horizontal and vertical grid-lines

도표 영상에서 그리드 라인의 교차점들을 추출하고 수평 수직 라인의 특징들을 분류한 다음 라인 내부의 분할된 점들을 연결한 결과를 그림 7에 도시하였다.

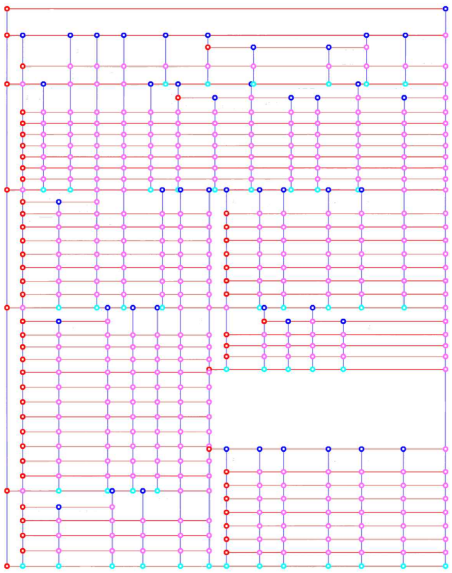


그림 7. 그리드 라인 교차점을 추출하고 교차점의 종류를 분류한 예

Fig. 7. Grid-line crossing point extraction and crossing point type classification

그리드 라인의 교차점들을 찾아서 그 종류를 확인한 다음 도표의 레이아웃을 분석하여 각 셀들을 인덱싱한다. 그림 8에 셀 인덱싱 규칙을 적용하여 도표 레이아웃 분석 결과를 표시한 예를 도시하였다.

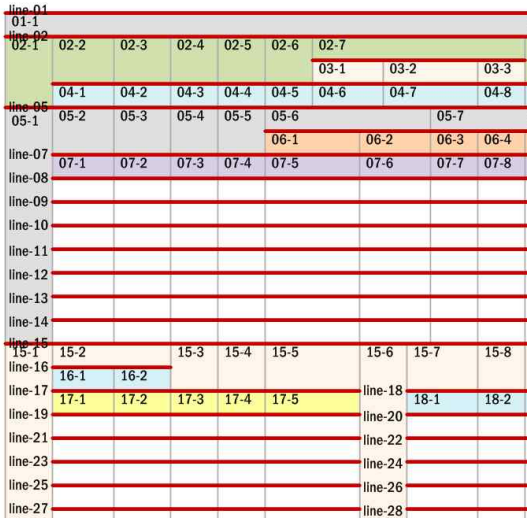


그림 8. 도표 레이아웃 분석 결과에 따른 셀 인덱싱 예

Fig. 8. Cell indexing example according to table layout analysis result

셀들을 인덱싱할 때 상단 셀 우선 - 좌측 셀 우선 규칙에 따르면 일관성 있게 도표 레이아웃 해석 결과를 표현할 수 있다. 상단 셀 우선 규칙에 따라 수평 그리드 라인들을 문서 상단에서 하단으로 위치에 따라 차례로 인덱싱한 다음 각 수평 라인마다 별도로 좌측에서 우측으로 교차점 위치에 따라 셀들을 차례로 인덱싱한다.

도표 레이아웃을 셀 인덱싱 결과로 표현한 다음 각 셀마다 내부에 포함된 문자들을 별도로 세그멘테이션하고 OCR 문자 인식기를 이용하여 문자 인식을 수행하면 도표 문서를 논리적으로 재구성하여 워드프로세서 편집기에 작업할 수 있는 문서 형태로 표현할 수 있다. 그림 9에 도표 문서 분석 결과를 표현한 셀 인덱싱 결과에 따라 도표를 재구성하고 내부의 문자들을 세그멘테이션 한 결과를 도시하였다.

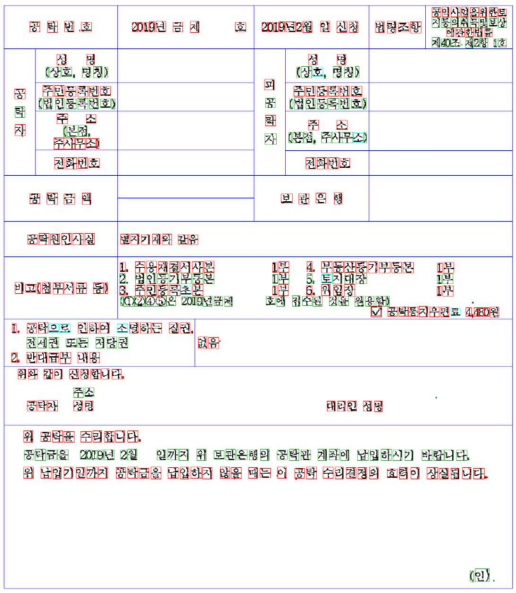


그림 9. 도표 레이아웃 분석 결과에 따라 도표를 재구성하고 문자들을 세그멘테이션한 예

Fig. 9. Table reconstruction and character segmentation example using layout analysis

양식 문서에 포함된 테이블의 구조를 분석하고 그 결과를 셀 인덱스로 표현하는 알고리즘은 다양한 양식 문서들을 고속으로 디지털화 하는 작업에 효과적으로 이용될 수 있다.

4. 시뮬레이션 및 결과 고찰

제안한 양식 문서 도표 구조 분석 알고리즘을 Visual Studio 2019를 이용하여 구현하고 다양한 양식 문서들을 대상으로 구조 분석 결과를 도출한 후 그 성능을 분석하였다. 실험을 위해 관공서 및 은행에서 사용되는 양식 문서 58장을

1,653 x 2,338 픽셀 크기로 스캐닝해서 데이터베이스를 구축하였다. 실험에 사용된 이미지에는 세로 방향 또는 가로 방향으로 작성된 도표들이 포함되어 있다. 제안한 알고리즘의 도표 구조 분석 성능을 평가하기 위해 그림 10에 도시한 것처럼 그리드 라인 추출, 교차점 추출 및 구분 그리고 도표 재구성 결과를 출력하였다.

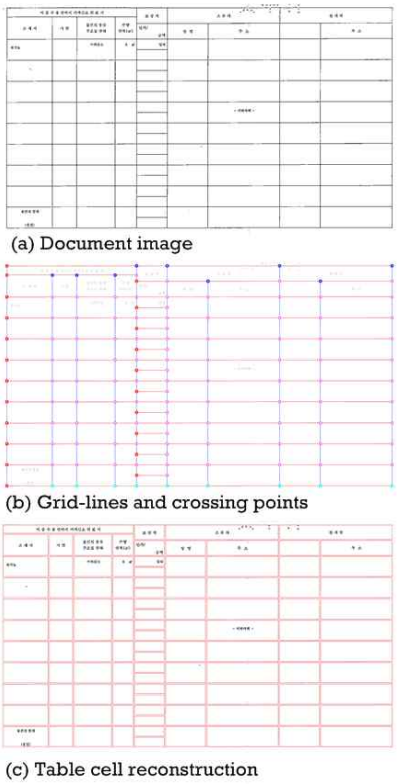


그림 10. 도표 레이아웃 분석 실험을 위한 문서 출력

Fig. 10. Document image generation for table layout analysis experiment

도표로 구성된 58개의 실험용 양식 문서 영상을 대상으로 수평 및 수직 그리드 라인 추출, 교차점 추출 및 도표 셀을 재구성하고 그 결과를 분석한 내용을 표 1에 도시하였다.

표 1. 도표 검출 및 재구성 성능 분석

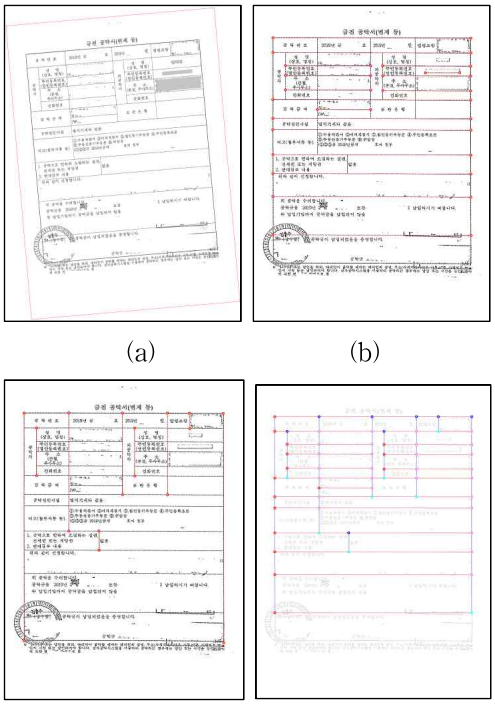
Table 1. Evaluation of table extraction and reconstruction

구 분		데이터베이스		실험 결과	
문서 영상 개수		58			
그리드 라인	수평	1720	2848	1689(98.2%)	2804(98.5%)
	수직	1128		1115(98.8%)	
교차점		929		903(97.2%)	
셀		727		701(96.4%)	

도표 구조 분석 알고리즘의 성능을 평가하기 위해 수집한 58개 양식 문서 영상의 도표에는 수평 그리드 라인 및 수직 그리드 라인이 각각 1,720개 및 1,128개가 포함되어 있어서 전체 2,848개의 그리드 라인이 존재한다. 제안한 알고리즘을 이용하여 그리드 라인을 검출한 결과 수평 및 수직 그리드 라인을 각각 98.2% 및 98.8%를 검출할 수 있어서 전체 그리드 라인 검출율은 98.5%를 보였다.

교차점을 검출하고 결과를 분석해서 셀을 추출하는 성능 실험에서는 각각 97.2% 및 96.4%의 검출율을 보였다. 교차점 검출에서 에러가 발생하는 경우는 그리드 라인의 절단 부분 길이가 허용 임계치를 벗어났을 때였으며 셀 추출에서 에러가 발생하는 경우는 교차점 미검출로 인해 셀 배치 규칙에 벗어난 도표에서 발생하였다. 그림 11에서 기울어진 문서 영상(a)에 대한 수평 라인 추출(b) 및 수직 라인 추출(c) 정보를 활용해서 도표 레이아웃을 분석한 결과(d)에서 수평 잡음 라인은 제거되었지만 도표의 하단부 마지막 셀을 찾지 못한 것을 볼 수 있다.

크기가 1,653 x 2,338인 실험용 문서 영상에 대해 제안한 라인 추적 알고리즘을 기반으로 그리드 라인을 검출하고 도표 레이아웃을 분석한 다음 도표를 재구성하는데 평균 0.41초 정도의 처리 시간이 소요되어 필터링을 통한 테이블 라인 영상 개선 방법들에 비해 비교적 빠른 속도로 우수한 결과를 얻을 수 있음을 확인하였다.



(a) (b)

(c) (d)

그림 11. 기울어진 도표 레이아웃 분석 실험에서 발생한 셀검출 오류 예
Fig. 11. A cell detection error example od a skewed document image.

5. 결 론

본 논문에서는 양식 문서에 포함된 도표를 해석하기 위해 그리드 라인을 추적하는 방식으로 수평 및 수직 직선 성분을 찾아내고 교차점을 해석하여 그리드 셀의 레이아웃 정보를 추출할 수 있는 알고리즘을 제안하였다.

관공서 및 은행에서 사용된 양식 문서 58장을 대상으로 실험을 수행한 결과 라인 추적 알고리즘으로 그리드라인의 98.5%을 검출할 수 있었으며 교차점과 셀 정보를 각각 97.2% 및 96.4% 추출할 수 있었다. 제안한 방법으로 문서 영상을 처리하는데 평균 0.41초 정도가 소요되어 비교적 빠른 시간 내에 문서영상에 포함된 도표의 구조를 분석할 수 있음을 확인하였다.

감사의 글

본 연구는 산업통상자원부와 한국산업기술진흥원의 “지역혁신클러스터육성(R&D, P0015279_산업 현장 내 맞춤형 AI 서비스를 위한 지능형 플랫폼 개발 및 실증운영)” 사업의 지원을 받아 수행된 연구결과임

참 고 문 헌

[1] X. Liang, A. Cheddad and J. Hall, “Comparative Study of Layout Analysis of Tabulated Historical Documents”, Big Data Research, pp.1-13, Vol. 24, 2021. DOI: https://doi.org/10.1016/j.bdr.2021.100195

[2] Y. Yoon, K. Ban, H. Yoon and J. Kim, “Automatic Container Code Recognition from Multiple Views”, ETRI Journal, pp. 767-775, Vol. 38, Issue 4, 2016. https://www.koreascience.or.kr/article/JAKO201671261180459.page

[3] B. Couasnon and A. Lemaitre, “Recognition of Tables and Forms”, pp.646-699, Handbook of Document Image Processing and Recognition, 2014. https://hal.inria.fr/hal-01087230

[4] P. Rege, C. Chanchal and A. Chandrakar, “Text-Image Separation in Document Images using Boundary/Perimeter Detection”, J. of Signal Image Processing, pp.29-35, Vol. 4, 2013. https://ia600609.us.archive.org/1/items/indexing_theides_1044/1044.pdf

[5] D. Nazir, K. Hashmi, A. Pagani, M. Liwicki, D. Striker and M. Afzal, “HybridTabNet: Towards Better Table

Detection in Scanned Document Images", Applied Science, pp.1-22, 2021. DOI: https://doi.org/10.3390/app11188396
[6] D. Burdick, M. Danilevsky, A. Evmievski, Y. Katsis, and N. Wang,"Table Extraction and Understanding for Scientific and Enterprise Application", PVLDB, Vol. 13, No, 12, pp.3433-3436, 2020.
[7] T. Nguyen, A. Doucet and M. Coustaty, "Enhancing Table of Contents Extraction by System Aggregation", The 14th IAPR International Conference on Document Analysis and Recognition, pp.242-247, 2017.
[8] D. Tran, T. Tran, A. Oh, S. Kim and I. Na, "Table Detection from Document Image using Vertical Arrangement of Text Blocks", International Journal of Contents, pp.77-85, Vol. 11, No. 4, 2015.
[9] S. Perantonis, B. Gatos and N. Papamarkos, "Block decomposition and segmentation for fast Hough transform evaluation", pp.811-824, Vol. 32, 1999. https://users.iit.demokritos.gr/~bgat/BHT.pdf
[10] L. Tong, H. Zhao, Q. Peng, G. Zhan and Y. LI, "Document Image Skew Correction Method based on Characteristic Sample Point Detection and Hough Transform", Journal of Convergence Information Technology, pp.576-584, Vol. 7, No. 22, 2012. https://www.semanticscholar.org/paper/Document-Image-Skew-Correction-Method-based-on-and-Tong-Zhao/d8f03ec797f551660d741bfb479adfcbc341dbac

저 자 소 개

김계경(Kye-Kyung Kim)

1989.2 경북대학교 전자공학과 졸업
1992.2 경북대학교 전자공학과 석사
1997.2 경북대학교 전자공학과 박사
1998.8-2001.2 CENPARMI 방문과학자
2001.3-현재 : 한국전자통신연구원 근무
<주관심분야> 패턴인식, 컴퓨터비전, 로봇비전, 인공지능