



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2023-0158786
(43) 공개일자 2023년11월21일

(51) 국제특허분류(Int. Cl.)
G06V 30/413 (2022.01) G06V 30/412 (2022.01)
(52) CPC특허분류
G06V 30/413 (2022.01)
G06F 40/131 (2020.01)
(21) 출원번호 10-2022-0058336
(22) 출원일자 2022년05월12일
심사청구일자 2022년05월12일

(71) 출원인
창원대학교 산학협력단
경상남도 창원시 의창구 창원대학로 20(퇴촌동)
(72) 발명자
김병욱
경상남도 창원시 의창구 창원대학로 20, 51호관 326호
김민태
세종특별자치시 달빛로 158, 707동 1102호 (아름동, 범지기마을 7단지)
(74) 대리인
김종선, 이형석

전체 청구항 수 : 총 14 항

(54) 발명의 명칭 문서 내 테이블 정보를 추출하는 방법 및 이를 지원하는 테이블 정보 추출 장치

(57) 요약

적어도 하나의 이미지 및 적어도 하나의 테이블을 포함하는 제1 파일 형식의 문서 내에서 제1 영역을 식별하는 테이블 인식부, 상기 식별된 제1 영역에서 행과 열 각각의 선(line) 정보를 획득하는 영상 처리 알고리즘 모듈, 및 상기 획득된 선 정보에 기반하여, 상기 적어도 하나의 테이블에서 복수의 텍스트 데이터들을 지정된 정렬 형태에 따라 추출하는 데이터 가공부를 포함하는 테이블 정보 추출 장치가 개시된다. 이 외에도 본 문서를 통해 파악되는 다양한 실시 예가 가능하다.

대표도 - 도1



(52) CPC특허분류
G06V 30/412 (2022.01)

이 발명을 지원한 국가연구개발사업

과제고유번호	1711153443
과제번호	2022R1A2B5B01001543
부처명	과학기술정보통신부
과제관리(전문)기관명	한국연구재단
연구사업명	개인기초연구(과기정통부)
연구과제명	교통정보 서비스 고도화를 위한 자유공간 광전송 기반 저궤도 위성통신의 빔 정밀조
향 기술 연구	
기 여 율	1/1
과제수행기관명	창원대학교 산학협력단
연구기간	2022.03.01 ~ 2023.02.28

명세서

청구범위

청구항 1

테이블 정보 추출 장치에 있어서,

적어도 하나의 이미지 및 적어도 하나의 테이블을 포함하는 제1 파일 형식의 문서 내에서 제1 영역을 식별하는 테이블 인식부;

상기 식별된 제1 영역에서 행과 열 각각의 선(line) 정보를 획득하는 영상 처리 알고리즘 모듈; 및

상기 획득된 선 정보에 기반하여, 상기 적어도 하나의 테이블에서 복수의 텍스트 데이터들을 지정된 정렬 형태에 따라 추출하는 데이터 가공부를 포함하는, 테이블 정보 추출 장치.

청구항 2

제 1 항에 있어서,

상기 영상 처리 알고리즘 모듈은 상기 행과 열이 직교하는 교차점의 개수에 기반하여, 상기 식별된 제1 영역을 상기 적어도 하나의 테이블이 위치한 제2 영역으로 결정하도록 설정된, 테이블 정보 추출 장치.

청구항 3

제 2 항에 있어서,

상기 영상 처리 알고리즘 모듈은 상기 결정된 제2 영역에 기반하여 상기 적어도 하나의 테이블의 좌표를 결정하도록 설정된, 테이블 정보 추출 장치.

청구항 4

제 1 항에 있어서,

상기 제1 파일 형식은 PDF(portable document format, PDF)인, 테이블 정보 추출 장치.

청구항 5

제 1 항에 있어서,

상기 적어도 하나의 이미지는 설계 대상의 적어도 일부에 해당하는 객체를 포함하는, 테이블 정보 추출 장치.

청구항 6

제 1 항에 있어서,

상기 추출된 복수의 텍스트 데이터들은 상기 제1 파일 형식과 다른 제2 파일 형식의 문서에 저장되는, 테이블 정보 추출 장치.

청구항 7

제 6 항에 있어서,

상기 제2 파일 형식은 CSV(comma-separated values, CSV)인, 테이블 정보 추출 장치.

청구항 8

문서 내 테이블 정보를 추출하는 방법에 있어서,

적어도 하나의 이미지 및 적어도 하나의 테이블을 포함하는 제1 파일 형식의 문서 내에서 제1 영역을 식별하는 단계;

상기 식별된 제1 영역에서 행과 열 각각의 선(line) 정보를 획득하는 단계; 및

상기 획득된 선 정보에 기반하여, 상기 적어도 하나의 테이블에서 복수의 텍스트 데이터들을 지정된 정렬 형태에 따라 추출하는 단계를 포함하는, 문서 내 테이블 정보를 추출하는 방법.

청구항 9

제 8 항에 있어서,

상기 획득하는 단계는 상기 행과 열이 직교하는 교차점의 개수에 기반하여, 상기 식별된 제1 영역을 상기 적어도 하나의 테이블이 위치한 제2 영역으로 결정하는, 문서 내 테이블 정보를 추출하는 방법.

청구항 10

제 9 항에 있어서,

상기 획득하는 단계는 상기 결정된 제2 영역에 기반하여 상기 적어도 하나의 테이블의 좌표를 결정하는, 문서 내 테이블 정보를 추출하는 방법.

청구항 11

제 8 항에 있어서,

상기 제1 파일 형식은 PDF(portable document format, PDF)인, 문서 내 테이블 정보를 추출하는 방법.

청구항 12

제 8 항에 있어서,

상기 적어도 하나의 이미지는 설계 대상의 적어도 일부에 해당하는 객체를 포함하는, 문서 내 테이블 정보를 추출하는 방법.

청구항 13

제 8 항에 있어서,

상기 추출된 복수의 텍스트 데이터들을 상기 제1 파일 형식과 다른 제2 파일 형식의 문서로 저장하는 단계를 더 포함하는, 문서 내 테이블 정보를 추출하는 방법.

청구항 14

제 13 항에 있어서,

상기 제2 파일 형식은 CSV(comma-separated values, CSV)인, 문서 내 테이블 정보를 추출하는 방법.

발명의 설명

기술 분야

[0001] 본 문서에서 개시되는 실시 예들은 문서 내 테이블 정보를 추출하는 방법 및 이를 지원하는 테이블 정보 추출 장치에 관한 것이다.

배경 기술

[0002] 건축물, 구조물 및 각종 제품 등은 그에 해당하는 설계 도면을 필요로 할 수 있다. 예를 들어, 건물을 건축할 때에는 건물의 외형을 구상하는 것뿐만 아니라 건물의 각층에서 복수의 공간들을 구분하거나 건물의 안전성을 확보하기 위하여 설계 도면이 이용될 수 있다.

[0003] 설계 도면은 설계 대상(예: 건축물, 구조물 또는 각종 제품)을 설계하는 과정에서 수차례에 걸쳐 변경될 수 있다. 설계 도면은 변경된 내용을 포함하여 지정된 파일 형식의 문서로 저장될 수 있다. 예를 들어, 설계 도면은 PDF(portable document format, PDF) 파일의 문서로 저장될 수 있다.

[0004] 지정된 파일 형식의 문서로 저장된 설계 도면에는 설계와 관련된 다양한 텍스트 데이터들이 테이블 형태로 포함

될 수 있다. 예를 들어, 설계 대상인 건물의 특정 층(예: 1층)에 해당하는 설계 도면에는, 상기 특정 층의 구조를 나타내는 평면도가 한쪽에 배치되고, 상기 특정 층의 제1 구역과 관련된 제1 수치 및 상기 특정 층의 제2 구역과 관련된 제2 수치를 포함하는 테이블이 다른 한쪽에 배치될 수 있다.

[0005] 상술한 설계 도면은 설계 도면의 버전 관리 등 필요에 따라 테이블 내 정보들을 제공할 수 있다. 상기 제공과 관련하여, 사용자는 최초 설계 도면부터 변경된 설계 도면에 이르기까지 여러 설계 도면들을 일일이 확인하며 테이블 내 정보들을 추출할 수 있다. 그러나, 사용자의 수작업에 의존하는 것은 상기 테이블 내 정보들을 추출하는 데 빈번한 오류 발생 확률과 더불어 상당한 소요 시간, 그 소요 시간에 따른 비용 증가를 동반할 수밖에 없다. 설계 도면에서 테이블 내 정보를 추출하기 위하여 OCR(optical character reader, OCR) 등의 기능을 이용하는 방법이 존재하지만, 이것은 정형화된 문서, 예컨대 doc 파일 형식의 문서에 포함된 테이블에서 정보를 추출하기 위해 보다 적합한 방법일 수 있다. 따라서, 설계 대상의 이미지를 비롯하여 여러 형태의 테이블을 포함하는 설계 도면으로부터 테이블 정보를 추출하는 새로운 방법이 요구된다.

발명의 내용

해결하려는 과제

[0006] 이에 따라 본 문서에 개시되는 다양한 실시 예에서는, 이미지 및 테이블을 포함하는 설계 도면에서 테이블을 구분하여 텍스트 데이터들을 추출할 수 있는 문서 내 테이블 정보를 추출하는 방법 및 이를 지원하는 테이블 정보 추출 장치를 제공할 수 있다.

과제의 해결 수단

[0007] 일 실시 예에 따르면, 테이블 정보 추출 장치는 적어도 하나의 이미지 및 적어도 하나의 테이블을 포함하는 제1 파일 형식의 문서 내에서 제1 영역을 식별하는 테이블 인식부, 상기 식별된 제1 영역에서 행과 열 각각의 선(line) 정보를 획득하는 영상 처리 알고리즘 모듈, 및 상기 획득된 선 정보에 기반하여, 상기 적어도 하나의 테이블에서 복수의 텍스트 데이터들을 지정된 정렬 형태에 따라 추출하는 데이터 가공부를 포함할 수 있다.

[0008] 일 실시 예에 따르면, 상기 영상 처리 알고리즘 모듈은 상기 행과 열이 직교하는 교차점의 개수에 기반하여, 상기 식별된 제1 영역을 상기 적어도 하나의 테이블이 위치한 제2 영역으로 결정하도록 설정될 수 있다.

[0009] 일 실시 예에 따르면, 상기 영상 처리 알고리즘 모듈은 상기 결정된 제2 영역에 기반하여 상기 적어도 하나의 테이블의 좌표를 결정하도록 설정될 수 있다.

[0010] 일 실시 예에 따르면, 상기 제1 파일 형식은 PDF(portable document format, PDF)일 수 있다.

[0011] 일 실시 예에 따르면, 상기 적어도 하나의 이미지는 설계 대상의 적어도 일부에 해당하는 객체를 포함할 수 있다.

[0012] 일 실시 예에 따르면, 상기 추출된 복수의 텍스트 데이터들은 상기 제1 파일 형식과 다른 제2 파일 형식의 문서에 저장될 수 있다.

[0013] 일 실시 예에 따르면, 상기 제2 파일 형식은 CSV(comma-separated values, CSV)일 수 있다.

[0014] 일 실시 예에 따르면, 문서 내 테이블 정보를 추출하는 방법은 적어도 하나의 이미지 및 적어도 하나의 테이블을 포함하는 제1 파일 형식의 문서 내에서 제1 영역을 식별하는 단계, 상기 식별된 제1 영역에서 행과 열 각각의 선(line) 정보를 획득하는 단계, 및 상기 획득된 선 정보에 기반하여, 상기 적어도 하나의 테이블에서 복수의 텍스트 데이터들을 지정된 정렬 형태에 따라 추출하는 단계를 포함할 수 있다.

[0015] 일 실시 예에 따르면, 상기 획득하는 단계는 상기 행과 열이 직교하는 교차점의 개수에 기반하여, 상기 식별된 제1 영역을 상기 적어도 하나의 테이블이 위치한 제2 영역으로 결정할 수 있다.

[0016] 일 실시 예에 따르면, 상기 획득하는 단계는 상기 결정된 제2 영역에 기반하여 상기 적어도 하나의 테이블의 좌표를 결정할 수 있다.

[0017] 일 실시 예에 따르면, 상기 제1 파일 형식은 PDF(portable document format, PDF)일 수 있다.

[0018] 일 실시 예에 따르면, 상기 적어도 하나의 이미지는 설계 대상의 적어도 일부에 해당하는 객체를 포함할 수 있다.

[0019] 일 실시 예에 따르면, 상기 추출된 복수의 텍스트 데이터들을 상기 제1 파일 형식과 다른 제2 파일 형식의 문서로 저장하는 단계를 더 포함할 수 있다.

[0020] 일 실시 예에 따르면, 상기 제2 파일 형식은 CSV(comma-separated values, CSV)일 수 있다.

발명의 효과

[0021] 본 문서에 개시되는 다양한 실시 예에 따른 문서 내 테이블 정보를 추출하는 방법 및 이를 지원하는 테이블 정보 추출 장치는 이미지 및 테이블을 포함하는 설계 도면에서 테이블을 구분하여 텍스트 데이터들을 추출함으로써, 설계 도면의 변경에 따른 변경 사항을 사용자로 하여금 쉽게 파악하도록 할 수 있다.

[0022] 이 외에도 본 문서를 통해 직접적 또는 간접적으로 파악되는 다양한 효과가 제공될 수 있다.

도면의 간단한 설명

[0023] 도 1은 일 실시 예에 따른 테이블 정보 추출 장치의 블록도이다.

도 2는 일 실시 예에 따른 문서 내 테이블 정보를 추출하는 방법을 도시한 흐름도이다.

도 3은 일 실시 예에 따른 문서 내 테이블이 위치한 영역의 좌표에 따라 복수의 텍스트 데이터들을 추출하여 다른 문서로 저장하는 방법을 도시한 흐름도이다.

도 4a는 일 실시 예에 따른 이미지 및 테이블을 포함하는 문서를 예시적으로 도시한 도면이다.

도 4b는 도 4a에 도시된 문서를 지정된 정규화 과정에 따라 변환한 도면을 예시적으로 도시한 도면이다.

도 4c는 도 4b에 도시된 도면에서 테이블이 위치한 것으로 예측되는 영역을 예시적으로 도시한 도면이다.

도 4d는 도 4c에 도시된 도면에서 테이블이 위치한 것으로 예측되는 영역의 행과 열을 예시적으로 도시한 도면이다.

도 4e는 도 4a에 도시된 문서에서 추출 대상인 복수의 텍스트들을 포함하는 테이블을 상기 문서에 예시적으로 도시한 도면이다.

도면의 설명과 관련하여, 동일 또는 대응되는 구성요소에 대해서는 동일한 참조 번호가 부여될 수 있다.

발명을 실시하기 위한 구체적인 내용

[0024] 이하, 첨부된 도면을 참조하여 본 발명의 바람직한 실시 예를 상세히 설명한다. 본 발명의 이점 및 특징, 그리고 그것들을 달성하는 방법은 첨부되는 도면과 함께 상세하게 후술되어 있는 실시 예들을 참조하면 명확해질 것이다. 그러나 본 발명은 이하에서 개시되는 실시 예들에 한정되는 것이 아니라 서로 다른 다양한 형태로 구현될 것이며, 단지 본 실시 예들은 본 발명의 개시가 완전하도록 하며, 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자에게 발명의 범주를 완전하게 알려주기 위해 제공되는 것이며, 본 발명은 청구항의 범주에 의해 정의될 뿐이다. 이하에서 동일한 참조 부호는 동일 구성요소를 지칭한다.

[0025] 비록 제1, 제2 등이 다양한 소자, 구성요소 및/또는 섹션들을 서술하기 위해서 사용되나, 이들 소자, 구성요소 및/또는 섹션들은 이들 용어에 의해 제한되지 않음은 물론이다. 이들 용어들은 단지 하나의 소자, 구성요소 또는 섹션들을 다른 소자, 구성요소 또는 섹션들과 구별하기 위하여 사용하는 것이다. 따라서, 이하에서 언급되는 제1 소자, 제1 구성요소 또는 제1 섹션은 본 발명의 기술적 사상 내에서 제2 소자, 제2 구성요소 또는 제2 섹션일 수도 있음은 물론이다.

[0026] 본 명세서에서 사용된 용어는 실시 예들을 설명하기 위한 것이며 본 발명을 제한하고자 하는 것은 아니다. 본 명세서에서, 단수형은 문구에서 특별히 언급하지 않는 한 복수형도 포함한다. 명세서에서 사용되는 "포함한다(comprises)" 및/또는 "이루어지다(made of)"는 언급된 구성요소, 단계, 동작 및/또는 소자는 하나 이상의 다른 구성요소, 단계, 동작 및/또는 소자의 존재 또는 추가를 배제하지 않는다.

[0027] 다른 정의가 없다면, 본 명세서에서 사용되는 모든 용어(기술 및 과학적 용어를 포함)는 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자에게 공통적으로 이해될 수 있는 의미로 사용될 수 있을 것이다. 또한, 일반적으로 사용되는 사전에 정의되어 있는 용어들은 명백하게 특별히 정의되어 있지 않는 한 이상적으로 또는 과도하게 해석되지 않는다.

- [0028] 이하, 첨부된 도면을 참조하여 본 발명의 구성에 대하여 상세히 설명한다.
- [0029] 도 1은 일 실시 예에 따른 테이블 정보 추출 장치의 블록도이다.
- [0030] 도 1을 참조하면, 일 실시 예에 따른 테이블 정보 추출 장치(100)는 테이블 인식부(110), 영상 처리 알고리즘 모듈(120), 및 데이터 가공부(130)를 포함할 수 있다. 다양한 실시 예에서, 테이블 정보 추출 장치(100)는 상기 구성요소들 중 적어도 하나를 생략하거나 다른 구성요소를 추가적으로 포함할 수 있다. 예를 들어, 테이블 정보 추출 장치(100)는 적어도 하나의 메모리를 더 포함할 수 있다.
- [0031] 일 실시 예에 따르면, 테이블 인식부(110)는 적어도 하나의 이미지 및 적어도 하나의 테이블을 포함하는 제1 파일 형식의 문서 내에서 제1 영역을 식별할 수 있다. 예를 들어, 테이블 인식부(110)는 설계 대상의 적어도 일부에 해당하는 객체 및 상기 객체와 관련된 수치를 포함하는 테이블 각각이 서로 다른 영역에 배치된 설계 도면(예: PDF 문서) 상에서 상기 테이블이 위치한 것으로 예측되는 영역을 제1 영역으로 결정할 수 있다.
- [0032] 일 실시 예에 따르면, 영상 처리 알고리즘 모듈(120)은 테이블 인식부(110)에 의하여 식별된 제1 영역에서 행과 열 각각의 선(line) 정보를 획득할 수 있다. 예를 들어, 영상 처리 알고리즘 모듈(120)은 설계 도면 내 테이블이 위치한 것으로 예측되는 제1 영역에서 가로선과 세로선 각각의 배치에 대응하는 선들을 선 정보로써 획득할 수 있다.
- [0033] 일 실시 예에 따르면, 영상 처리 알고리즘 모듈(120)은 획득된 선 정보에 기반하여 문서 내 제1 영역을 테이블이 위치한 제2 영역으로 결정할 수 있다. 예를 들어, 영상 처리 알고리즘 모듈(120)은 제1 영역에서 직교하는 행과 열의 교차점의 개수가 지정된 개수 이상인지 여부에 따라 제1 영역을 제2 영역으로 결정할 수 있다. 상기 지정된 개수는 테이블 형태를 구성하기 위한 교차점의 개수(예: 4개)에 대응할 수 있다.
- [0034] 일 실시 예에 따르면, 영상 처리 알고리즘 모듈(120)은 문서 내 제1 영역을 테이블이 위치한 제2 영역으로 결정하면, 결정된 제2 영역의 좌표를 결정할 수 있다. 상기 결정된 좌표는 문서 내 테이블에 포함된 테이블 정보(예: 설계 대상과 관련된 수치) 및/또는 상기 테이블 내 배치된 또 다른 테이블에 포함된 테이블 정보를 추출하기 위하여 이용될 수 있다.
- [0035] 일 실시 예에 따르면, 데이터 가공부(130)는 영상 처리 알고리즘 모듈(120)에 의하여 획득된 선 정보에 기반하여, 문서 내 테이블에서 테이블 정보를 지정된 정렬 형태에 따라 추출할 수 있다. 예를 들어, 데이터 가공부(130)는 설계 도면 내 테이블이 위치한 것으로 결정된 제2 영역에서 설계 대상과 관련된 수치를 제2 파일 형식(예: CSV 파일)의 문서에 이용될 수 있는 형태로 추출할 수 있다.
- [0036] 다양한 실시 예에 따르면, 상술한 테이블 정보 추출 장치(100) 내 적어도 하나의 구성요소들의 동작(또는 기능)은 적어도 하나의 프로세서(미도시)에 의하여 수행될 수 있다.
- [0037] 다양한 실시 예에 따르면, 적어도 하나의 프로세서는 메모리(미도시)로부터 명령(command) 또는 인스트럭션들(instructions)을 수신하고, 수신된 명령 또는 인스트럭션들에 따라 각 구성요소들을 제어하여, 다양한 기능들을 수행할 수 있다. 적어도 하나의 프로세서는 중앙 처리 장치(central processing unit, CPU), 마이크로 컨트롤 유닛(micro control unit, MCU), 마이크로 프로세서 유닛(micro processor unit, MPU) 등으로 구현될 수 있다.
- [0038] 도 2는 일 실시 예에 따른 문서 내 테이블 정보를 추출하는 방법을 도시한 흐름도이다.
- [0039] 단계 S210을 참조하면, 일 실시 예에 따른 테이블 정보 추출 장치(100)는 적어도 하나의 이미지 및 적어도 하나의 테이블을 포함하는 제1 파일 형식의 문서 내에서 제1 영역을 식별할 수 있다. 예를 들어, 테이블 인식부(110)는 설계 대상의 적어도 일부에 해당하는 객체 및 상기 객체와 관련된 수치를 포함하는 테이블 각각이 서로 다른 영역에 배치된 설계 도면(예: PDF 문서) 상에서 상기 테이블이 위치한 것으로 예측되는 영역을 제1 영역으로 결정할 수 있다.
- [0040] 단계 S220을 참조하면, 일 실시 예에 따른 테이블 정보 추출 장치(100)는 테이블 인식부(110)에 의하여 식별된 제1 영역에서 행과 열 각각의 선(line) 정보를 획득할 수 있다. 예를 들어, 영상 처리 알고리즘 모듈(120)은 설계 도면 내 테이블이 위치한 것으로 예측되는 제1 영역에서 가로선과 세로선 각각의 배치에 대응하는 선들을 선 정보로써 획득할 수 있다.
- [0041] 단계 S230을 참조하면, 일 실시 예에 따른 테이블 정보 추출 장치(100)는 영상 처리 알고리즘 모듈(120)에 의하여 획득된 선 정보에 기반하여, 문서 내 테이블에서 테이블 정보를 지정된 정렬 형태에 따라 추출할 수 있다.

예를 들어, 데이터 가공부(130)는 설계 도면 내 테이블이 위치한 것으로 결정된 제2 영역에서 설계 대상과 관련된 수치를 제2 파일 형식(예: CSV 파일)의 문서에 배치할 수 있는 형태로 추출할 수 있다.

- [0042] 도 3은 일 실시 예에 따른 문서 내 테이블이 위치한 영역의 좌표에 따라 복수의 텍스트 데이터들을 추출하여 다른 문서로 저장하는 방법을 도시한 흐름도이다. 도 3의 단계 S310 내지 단계 S350은 도 2의 단계 S220 이후 수행될 수 있다.
- [0043] 단계 S310을 참조하면, 일 실시 예에 따른 테이블 정보 추출 장치(100)는 제1 파일 형식의 문서 내 테이블이 위치한 것으로 예측되는 제1 영역에서 직교하는 행과 열의 교차점의 개수가 지정된 개수 이상인지 여부를 확인할 수 있다. 예를 들어, 영상 처리 알고리즘 모듈(120)은 제1 영역에서 직교하는 행과 열의 교차점의 개수가 테이블 형태를 구성하기 위한 교차점의 개수(예: 4개)에 대응하는지 여부를 확인할 수 있다.
- [0044] 단계 S320을 참조하면, 일 실시 예에 따른 테이블 정보 추출 장치(100)는 단계 S310에서 교차점의 개수가 지정된 개수 이상인 것으로 확인되는 것에 응답하여, 제1 영역을 테이블이 위치한 제2 영역으로 결정할 수 있다.
- [0045] 단계 S330을 참조하면, 일 실시 예에 따른 테이블 정보 추출 장치(100)는 문서 내 제1 영역을 테이블이 위치한 제2 영역으로 결정하면, 결정된 제2 영역의 좌표를 결정할 수 있다. 상기 결정된 좌표는 문서 내 테이블에 포함된 테이블 정보(예: 설계 대상과 관련된 수치) 및/또는 상기 테이블 내 배치된 또 다른 테이블에 포함된 테이블 정보를 추출하기 위하여 이용될 수 있다.
- [0046] 단계 S340을 참조하면 일 실시 예에 따른 테이블 정보 추출 장치(100)는 문서 내 테이블에서 테이블 정보를 지정된 정렬 형태에 따라 추출할 수 있다. 예를 들어, 데이터 가공부(130)는 설계 도면 내 테이블이 위치한 것으로 결정된 제2 영역에서 설계 대상과 관련된 수치를 제2 파일 형식(예: CSV 파일)의 문서에 이용될 수 있는 형태로 추출할 수 있다.
- [0047] 단계 S350을 참조하면, 일 실시 예에 따른 테이블 정보 추출 장치(100)는 단계 S340에서 추출된 복수의 텍스트 데이터들을 제2 파일 형식의 문서로 저장할 수 있다.
- [0048] 도 4a는 일 실시 예에 따른 이미지 및 테이블을 포함하는 문서를 예시적으로 도시한 도면이다.
- [0049] 도 4a를 참조하면, 일 실시 예에 따른 테이블 정보 추출 장치(100)는 설계 도면(400a)을 테이블 정보 추출 대상으로서 활성화할 수 있다. 설계 도면(400a)은 제1 파일 형식에 해당하는 PDF 문서로써, 문서 내에서 검색 가능한(searchable) 기능을 지원할 수 있다. 설계 도면(400a) 내에는 설계 대상인 객체(410)(예: 평면도)가 일부 영역에 배치되고, 테이블로 예측될 수 있는 주변 객체들(420a, 420b, 420c, 420d)이 다른 일부 영역에 배치될 수 있다.
- [0050] 다양한 실시 예에서, 상기 설계 도면(400a)은 본 발명의 이해를 돕기 위하여 예시적으로 도시한 도면이며, 건축물의 설계 도면뿐만 아니라 다양한 종류의 설계 도면으로 대체될 수 있다.
- [0051] 도 4b는 도 4a에 도시된 문서를 지정된 정규화 과정에 따라 변환한 도면을 예시적으로 도시한 도면이다. 도 4c는 도 4b에 도시된 도면에서 테이블이 위치한 것으로 예측되는 영역을 예시적으로 도시한 도면이다.
- [0052] 도 4b 및 도 4c를 참조하면, 일 실시 예에 따른 테이블 정보 추출 장치(100)는 도 4a에 도시된 설계 도면(400a)을 정규화 도면(400b)으로 변환한 후, 상기 정규화 도면(400b)을 예측 도면(400c)으로 변환할 수 있다. 먼저, 정규화 도면(400b)에서는 설계 도면(400a)의 크기가 지정된 크기(예: 896x896)로 조정될 수 있다. 예측 도면(400c)에서는 정규화 도면(400b) 상에서 흑백 처리된 객체들 중 테이블로 예측될 수 있는 주변 객체들(420a', 420b', 420c', 420d')이 지정된 색상(예: 주변 영역과 다른 명도의 색상)으로 변환될 수 있다. 다양한 실시 예에서, 테이블 정보 추출 장치(100)는 지정된 프로그램(예: TableNet)을 이용하여 상기 정규화 도면(400b) 및 상기 예측 도면(400c)을 생성할 수 있다. 또한, 테이블 정보 추출 장치(100)는 지정된 형식의 이미지(예: bmp 이미지)들에 기반하여 테이블이 위치한 영역을 예측할 수 있는 데이터베이스를 구축할 수 있다.
- [0053] 다양한 실시 예에 따르면, 설계 도면(400a) 상에는 복수의 테이블들뿐만 아니라 구조 또는 기계 장치와 관련된 이미지들이 포함될 수 있다. 상기 이미지들은 복수의 테이블들과 마찬가지로 가로 방향 및/또는 세로 방향의 직선들을 포함하므로, 상기 복수의 테이블들과 오인식될 수 있다. 따라서, 테이블 정보 추출 장치(100)는 상술한 정규화 도면(400b) 및 예측 도면(400c)을 거쳐 테이블로 예측될 수 있는 영역들을 개략적으로 검출하고, 지정된 형태(예: 사각형)로 마스킹한 이후, 상기 지정된 형태에 대응하는 영역의 좌표를 검출할 수 있다.
- [0054] 도 4d는 도 4c에 도시된 도면에서 테이블이 위치한 것으로 예측되는 영역의 행과 열을 예시적으로 도시한 도면

이다.

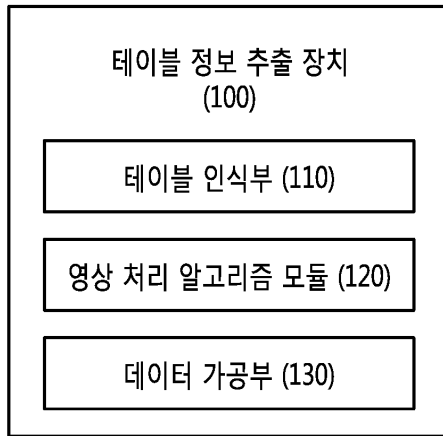
- [0055] 도 4d를 참조하면, 일 실시 예에 따른 테이블 정보 추출 장치(100)는 도 4c의 주변 객체들(420a', 420b', 420c', 420d')에 대응하는 선 정보들을 분석할 수 있다.
- [0056] 제1 상태(401)를 참조하면, 영상 처리 알고리즘 모듈(120)은 제1 주변 객체(420a')에서 세로선을 포함하는 선 정보(401a)를 확인할 수 있다.
- [0057] 제2 상태(402)를 참조하면, 영상 처리 알고리즘 모듈(120)은 제2 주변 객체(420b')에서 가로선, 세로선 및 1개의 교차점을 포함하는 선 정보(401b)를 확인할 수 있다.
- [0058] 제3 상태(403)를 참조하면, 영상 처리 알고리즘 모듈(120)은 제3 주변 객체(420c')에서 가로선을 포함하는 선 정보(401c)를 확인할 수 있다.
- [0059] 제4 상태(404)를 참조하면, 영상 처리 알고리즘 모듈(120)은 제4 주변 객체(420d')에서 가로선, 세로선 및 지정된 개수(예: 4개) 이상의 교차점을 포함하는 선 정보(401d)를 확인할 수 있다.
- [0060] 영상 처리 알고리즘 모듈(120)은 상술한 제1 내지 4 상태(401~404) 중 제4 상태(404)에서 테이블에 해당하는 선 정보(401d)를 확인하였으므로, 제4 주변 객체(420d')를 테이블이 위치한 영역(예: 제2 영역)으로 결정할 수 있다. 다양한 실시 예에서, 테이블 정보 추출 장치(100)는 지정된 프로그램(예: Camelot)을 이용하여 상기 제1 내지 제4 상태(401~404)를 확인할 수 있다.
- [0061] 복수의 테이블들 각각은 서로 직교하는 선들로 연결될 수 있다. 테이블 정보 추출 장치(100)는 상기 직교하는 선들의 교차점의 개수에 기반하여, 상기 복수의 테이블들 각각을 개별적인 테이블로 구분한 이후, 상기 구분된 테이블을 지정된 파일 형식(예: SCV)으로 저장할 수 있다.
- [0062] 도 4e는 도 4a에 도시된 문서에서 추출 대상인 복수의 텍스트들을 포함하는 테이블을 상기 문서에 예시적으로 도시한 도면이다.
- [0063] 도 4e를 참조하면, 일 실시 예에 따른 테이블 정보 추출 장치(100)는 설계 도면(400e) 상에 테이블이 위치한 영역(420)을 표시할 수 있다. 데이터 가공부(130)는 테이블이 위치한 영역(420)으로부터 복수의 텍스트 데이터들을 CSV 파일의 문서에 배치할 수 있는 형태로 추출할 수 있다.
- [0064] 이상과 같이 본 발명의 도시된 실시 예를 참고하여 설명하고 있으나, 이는 예시적인 것들에 불과하며, 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자라면 본 발명의 요지 및 범위에 벗어나지 않으면서도 다양한 변형, 변경 및 균등한 다양한 타 실시 예들이 가능하다는 것을 명백하게 알 수 있을 것이다. 따라서 본 발명의 진정한 기술적 보호 범위는 첨부된 청구범위의 기술적인 사상에 의해 정해져야 할 것이다.

부호의 설명

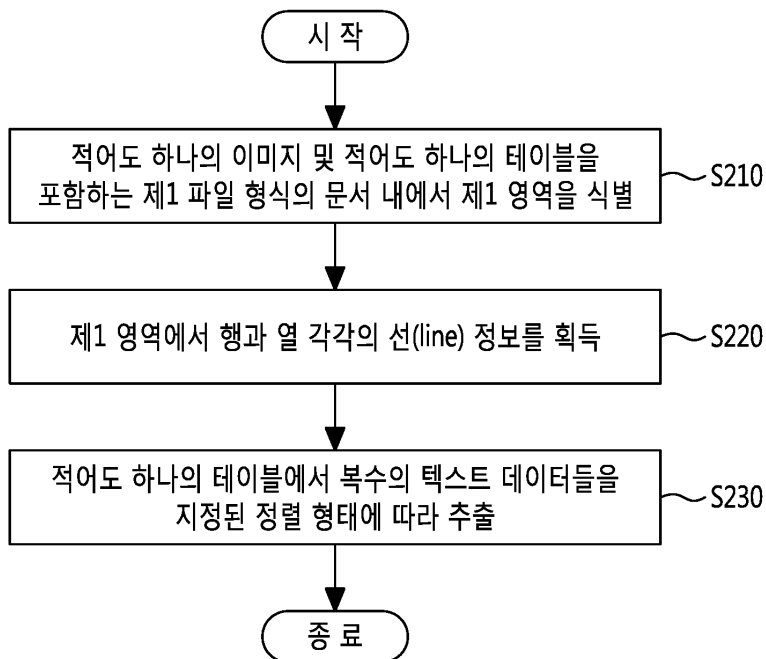
- [0065] 100: 테이블 정보 추출 장치
- 110: 테이블 인식부
- 120: 영상 처리 알고리즘 모듈
- 130: 데이터 가공부

도면

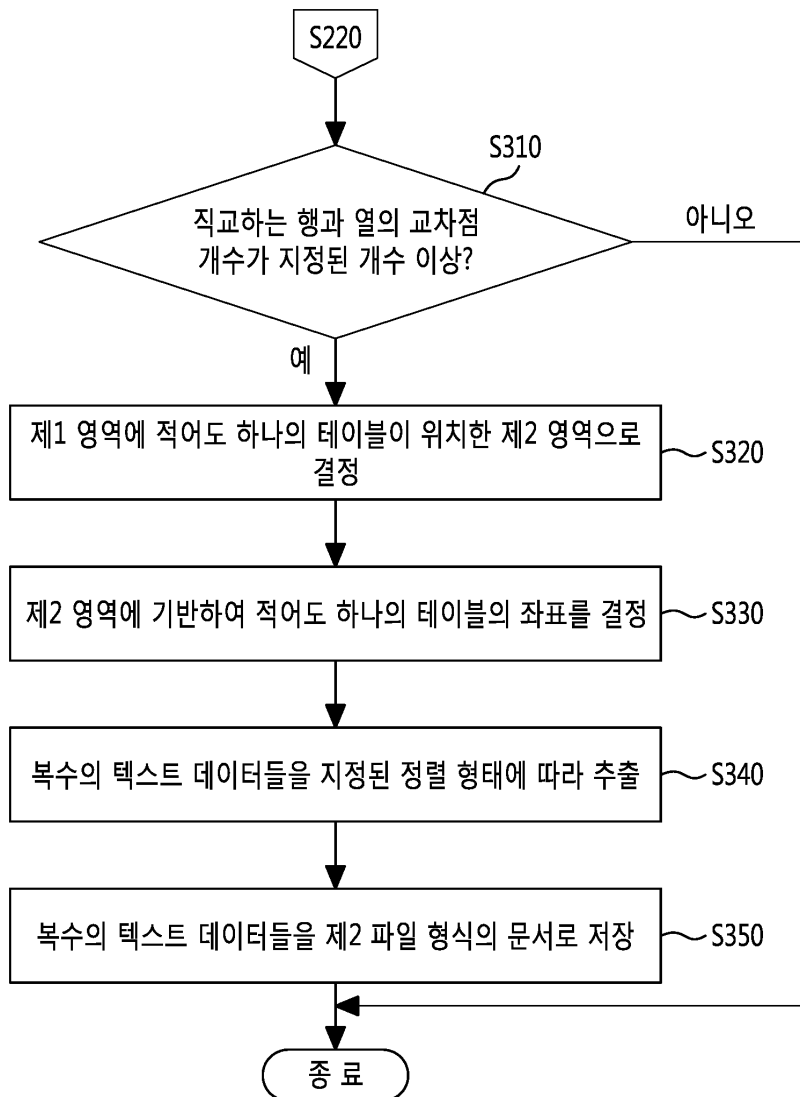
도면1



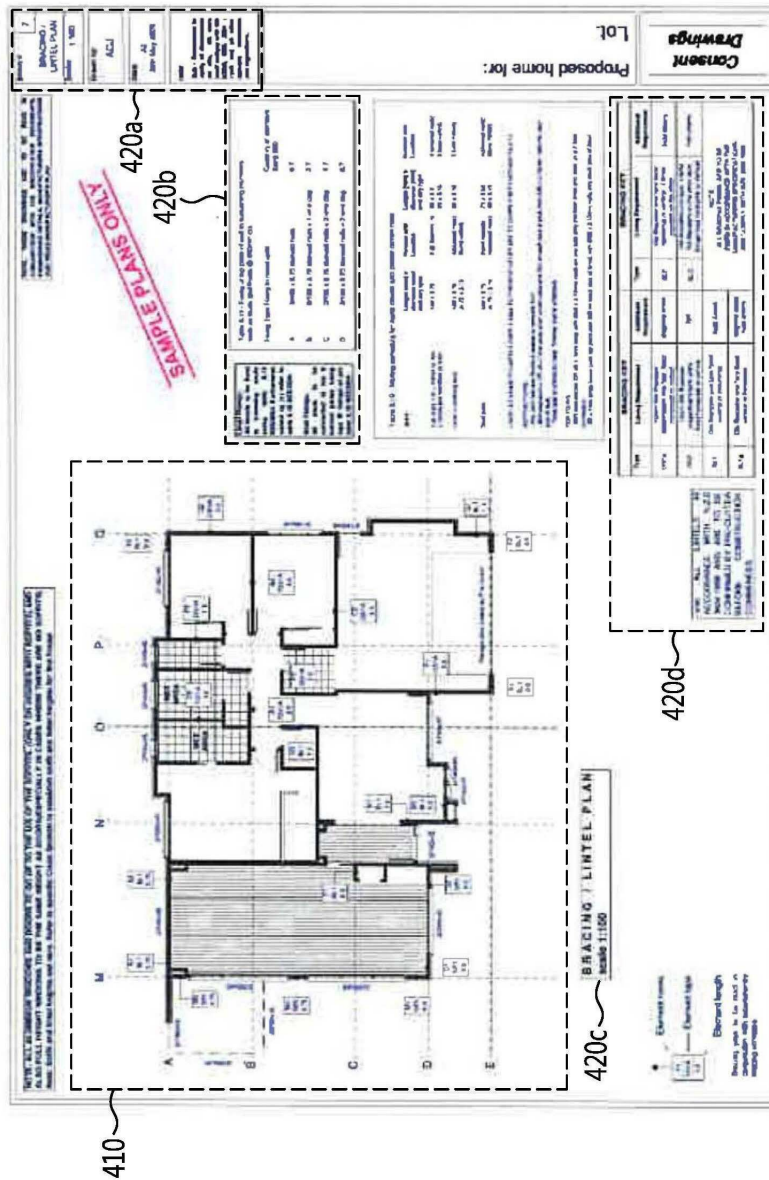
도면2



도면3

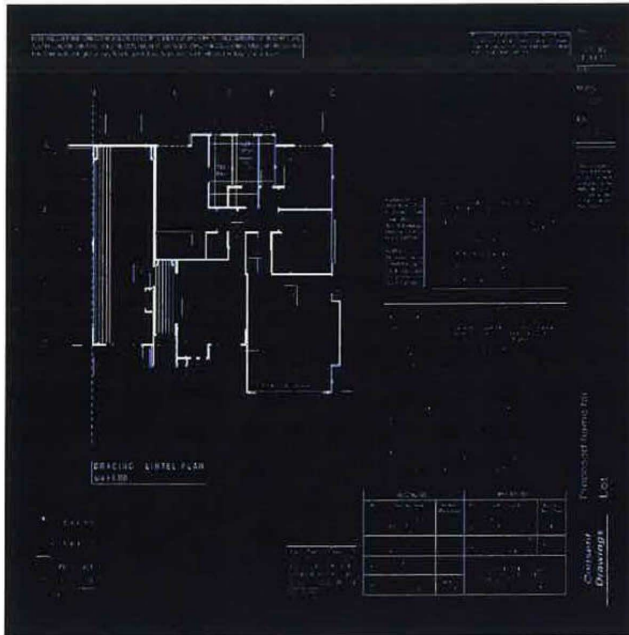


도면4a



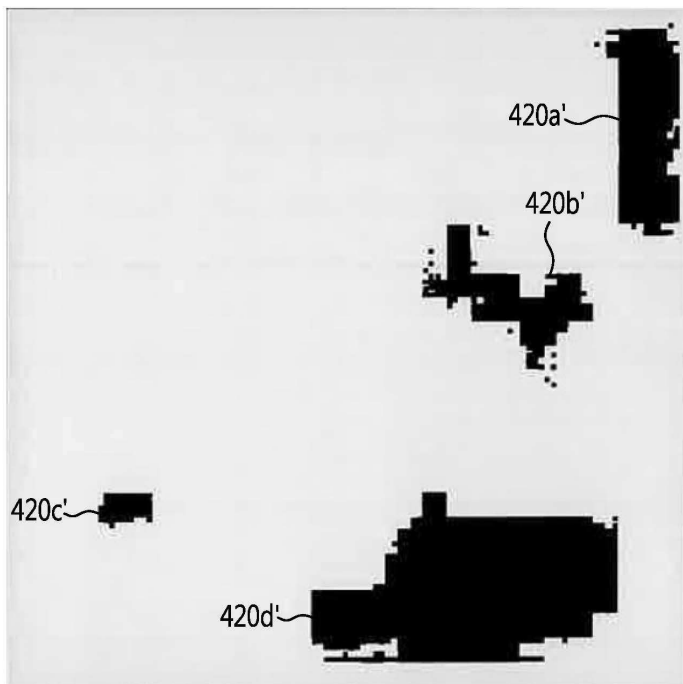
도면4b

400b

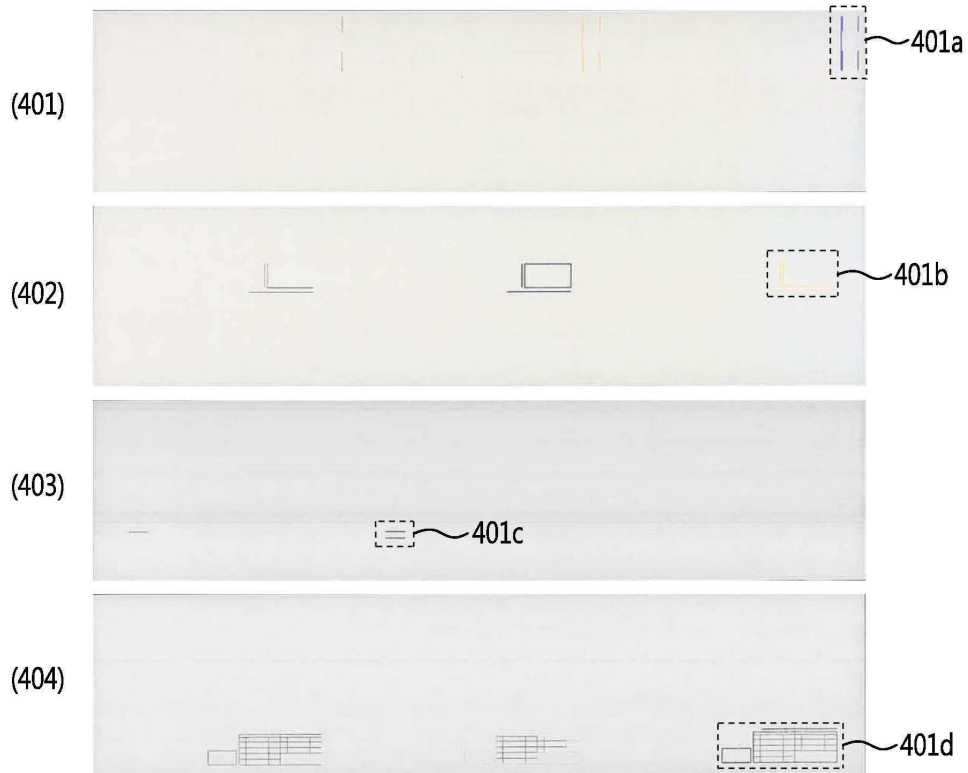


도면4c

400c



도면4d



도면4e

400e

