테크니컬 리포트

2020년 1차
인공지능
학습용
데이터 구축

헬스케어 영역

구강악 2D·3D 이미지

개요: 구강악 2D, 3D 데이터셋이란?

의료영상 AI 분야 중에서 국내 의료기기 제조업과 시너지를 극대화하여, 즉시 시장 경쟁력을 높일 수 있는 치의영상 기반의 AI 모델 개발에 활용할 수 있는 학습 데이터 셋으로 서울대학교치과병원과 ㈜헬스허브에서 구축했으며, 파노라마영상(2D) 10,000건 (10,000장)과 CBCT영상(3D) 4,000건 (약 2,000,000장)으로 구성되어 있다

치과 의료용 기기는 우리나라가 세계 시장점유 3위인 산업으로, 선진국과 경쟁하고 있으나, 최근 중국 등에서 저가 장비의 공세로 고급화가 절실한 상황이다. 우수한 장비제조 기술과 AI 소프트웨어가 접목되어 시장 경쟁력을 높이고 고급화된 제품의 생산을 통해 선진국과 경쟁에서 우위를 점하기 위해 활발하게 연구가 진행되고 있다

2D 영상은 파노라마 X-ray 영상으로, 일반적으로 치과에서 가장 많이 촬영되는 영상이며, 최근 CT 장비의 가격하락으로 3D 영상도 활발하게 촬영되고 있다. 본 데이터셋은 2D와 3D 영상에서 치아와 하치조신경관을 식별할 수 있는 데이터 이며, 치의영상에서 가장 기본이 되는 해부학적 구조물을 식별할 수 있다.

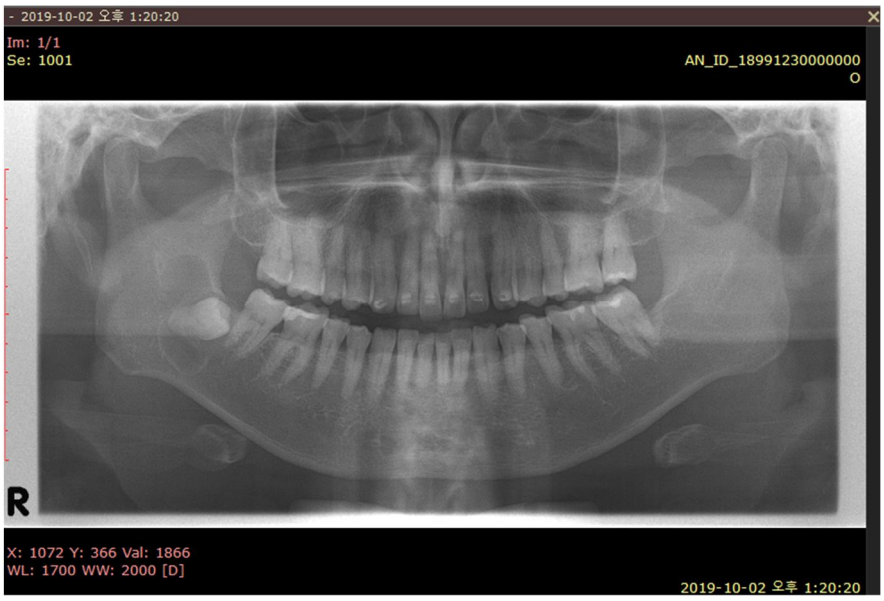

그림 1. 2D 파노라마 X-ray 영상

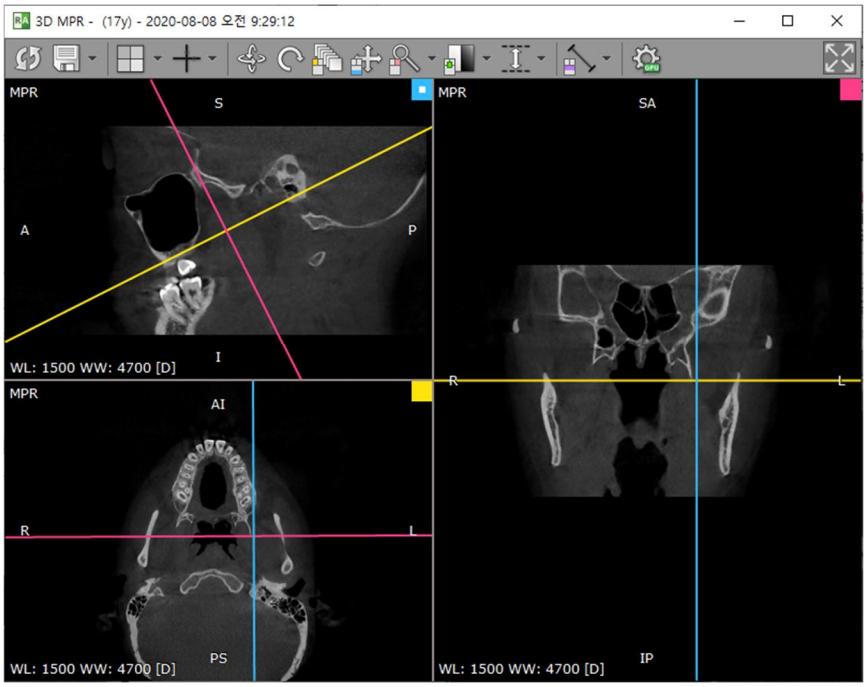

그림 2. 3D CBCT 영상

데이터셋의 구성

본 데이터셋은 파노라마 X-ray(2D) 10,000건 (10,000장)과 CBCT(3D) 4,000건 (약 2,000,000장)으로 구성되어 있다.

파노라마 X-ray(2D)의 경우에는 5,000건은 어노테이션이 포함되어 있지 않은 데이터이며, 5,000건은 치아 세그멘테이션 어노테이션이 포함되어 있다. 치아 세그멘테이션은 치아번호를 포함하고 있으며, FDI World Dental Federation notation 방식으로 표기되어 있다.

CBCT(3D)의 경우에는 1건이 약 500장의 영상으로 구성되어 있으며, 총 4,000건 중 2,000건은 치아 세그멘테이션 어노테이션이, 1,000건은 하치조신경관 세그멘테이션이 포함되어 있다. 마찬가지로 1,000건은 어노테이션이 포함되지 않은 데이터이다.

어노테이션이 포함되지 않은 데이터는 함께 제공되는 어노테이션도구를 통해 AI 개발에 필요한 어노테이션을 직접 수행하여 사용할 수 있다.

제공되는 모든 영상자료는 비식별화가 완료된 정보이며, 개인정보는 생년과 성별 이외에는 모두 제거되어 있다.

데이터 종류	**포함 내용**		**제공 방식**
파노라마 X-ray(2D) 영상	Raw data	5,000건 / 5,000장	DCM 포맷파일

(10,000건/10,000장)	치아 세그멘테이션	5,000건 / 5,000장	DCM, JPG, BIN[1)], JSON 포맷 파일
CBCT(3D) 영상 (4,000건/2,000,000장)	Raw data	1,000건 / 500,000장	DCM 포맷파일
	치아 세그멘테이션	2,000건 / 1,000,000장	DCM, JPG, BIN, JSON 포맷 파일
	하치조신경관 세그멘테이션	1,000건 / 500,000장	DCM, JPG, BIN, JSON 포맷 파일

1) BIN 파일은 어노테이션을 영상으로 확인할 수 있는 바이너리파일이며, 전용 뷰어를 별도로 제공함

데이터셋의 설계 기준과 분포

영상데이터의 편향성을 줄이기 위해 서울대학교치과병원과 강릉원주대학교치과병원 두 곳에서 영상을 수집하였다. 이는 촬영장비, 촬영방법 및 촬영환경의 다양성을 제공하기 위함이다. 두 기관에서 50%씩 수집하였다.

데이터 구조

데이터셋에 따른 항목과 해당 값은 아래 테이블과 같다.

항목			다중	필수	타입	설명
데이터그룹			1	Y		구축해야 할 데이터 종류: 2D-P-T, 3D-C-T, 3D-C-MC
	그룹ID		1		고유번호	"2D-P-T"로 고정
	명칭		1		문자	"2D Panoramic radiolographic for Teeth detection and labelling"로 고정
케이스	ID		1	Y	고유번호	하나의검사에대한하나의어노테이션ID
	참고		1		문자	Study Instance UID나Folder path 등이후해당case에대한실제dicom 등관련파일을access할수있는key값들을저장하기위함.
	상태		1		코드	COLLECTED / IN PROGRESS / IN VALIDATION / COMPLETED 이annotation결과물의진행상태를의미함.
	변경이력		n			
		사용자	1		고유번호	
		일자	1		년월일시	"YYYYMMDDHHmmss"
		설명	1		문자	
출처			1	Y		해당영상이수집된기관으로서울대치과병원, 강릉원주대치과병원등으로기록.
	출처ID		1		고유번호	

환자			1	Y		dicom tag
	환자ID		1		고유번호	가명화하여Subject ID를사용
	생년		1		연도	"YYYY"
	성별		1		코드	M / F / O: transgender 등M과F로결정되지않는경우O로기입
연계			1	Y		dicom header 정보를 이용하여 fill
	이름		1		문자	dicom tag의 study description
	프로토콜		1		문자	Optional 정보이며 dicom tag에 위치 등을 파악해야 함
	일자		1		일자	
	검사장비모델		1		문자	검사 장비 정보
어노테이션			1	Y		
	치아		n			
		번호	1		숫자	치아번호
		상태	1		코드	NORMAL / LOST
	조합		1		문자	x1, y1/.../xn, yn 방법이나binary 정보(0,0,0,1,1,0,0,0,3,3,0,4,4,0) ==> w*h

데이터 예시

어노테이션 정보를 제공하는 JSON 정보는 아래와 같다. 2D의 경우, cordinate을 2개 단위로 3D의 경우 3개를 한단위로 끊어 읽으면 된다.

.


```
{"data-group": { "id": "2D-C-T",
        "name": "2D Cone beam computed tomography for Teeth detection and labeling"},
        "case": { "id": 1599051329870,
        {"ref": "1.3.6.1.4.1.25403.172056116518522.46876.20200720125053.119",
                "status": "IN PROGRESS"},
        "facility": {"id": 1,},
        "patient": { "id": "00641366","birth_date": "1997","sex": "F "},
        "study": {
                "name": "standard panoramic",
                "protocol": "Normal",
                "date": "20200720",
                "acquisition_model": "RAYSCAN Alpha "},
        "meta": {"width": 2988,"height": 1468},
        "annotation": {"tooth":
        { "1": { "status": "normal","coordinate":
                [ 834.0,770.0,835.0, ........,1020.0]},
        "2": {"status": "lost","coordinate": null },
        ..............
```

"28": { "status": "lost", "coordinate": null}}}}

데이터 구축 과정

데이터 구축은 서울대학교치과병원과 강릉원주대학교치과병원에서 촬영된 파노라마 X-ray와 CBCT 영상을 반반씩 수집하였다.

수집된 데이터는 1차 비전문가 어노테이션, 2차 전문가 어노테이션 및 피드백, 3차 최종 승인 단계로 진행하였으며, 비전문가 어노테이션 그룹도 현 치의대 학생들이 참여하여 기본적으로 치과 관련 지식이 있는 사람들로 구성하여 그 전문성을 높혔다.

또한 어노테이션 정확성을 높이기 위해 마우스가 아닌 애플 팬슬을 활용하여 최대한 근접한 세그멘테이션이 가능하도록 작성하였다.

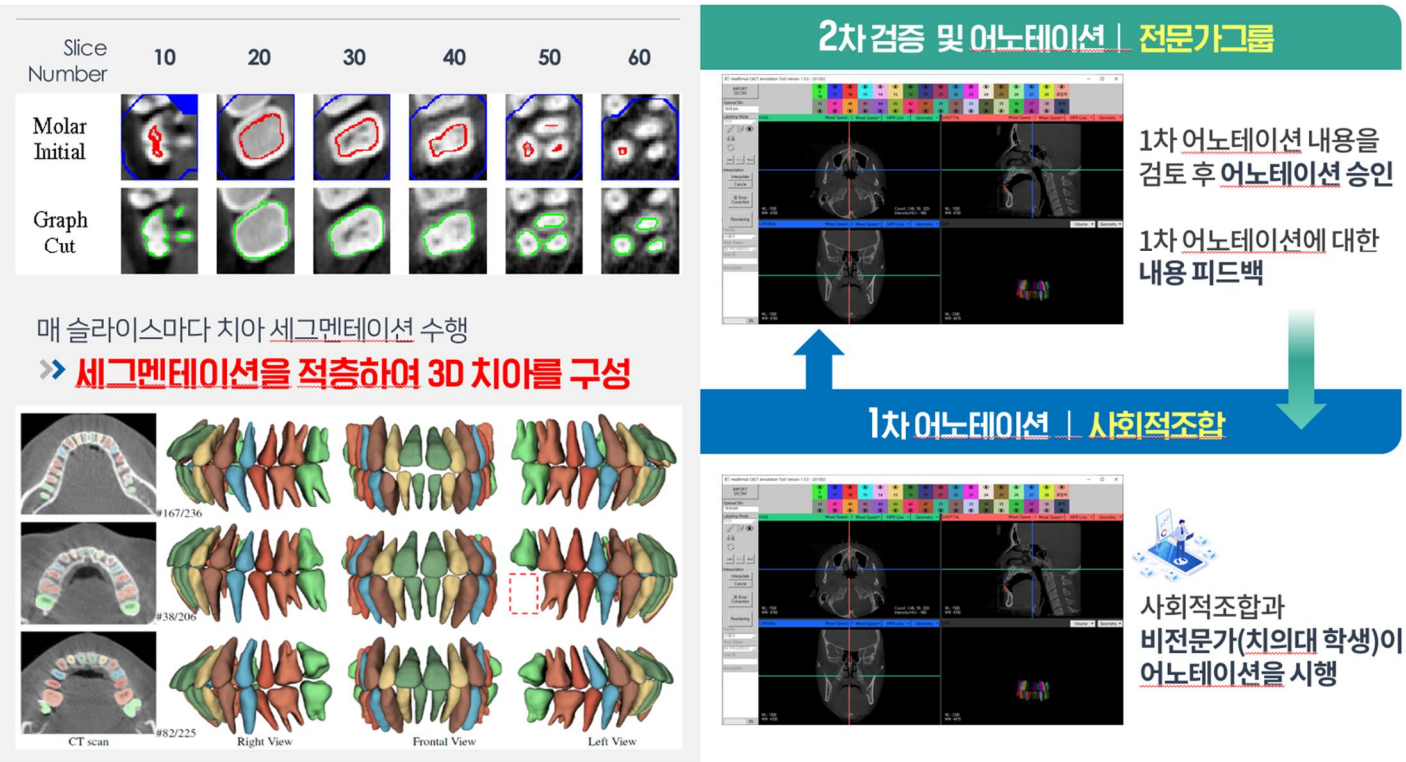

그림. 3D 어노테이션 방법 및 검증과정

검수와 품질 확보

의료영상의 특수성으로인해 크라우드 소싱인력으로 어노테이션을 진행하지 않았으며, 모두 치의영상학 전문의들로 구성된 검수단이 2차 어노테이션 및 검증을 진행하였다. 전문의가 검수한 영상은 외부 검수기관을 통해 3차 검증을 진행하였으며, 생산된 데이터를 바탕으로 자체 AI 모델 개발 및 데이터손을 통해 AI 모델 개발 및 성능에 문제가 없음을 확인하였다.

데이터 구축 담당자

수행기관(주관) : ㈜헬스허브 (전화: 02-511-3601), 이메일: hh-all@healthhub.kr