인공지능 데이터 구축·활용 가이드라인

- 야외 실제 촬영 한글 이미지 -

인공지능 학습용 데이터 구축	사업 총괄	동양시스템즈(주)
	데이터 설계	KINY.
	원천데이터 수집 및 정제	인포플라
	데이터 가공	인포플라
	데이터 검수	인포플라 동양시스템즈(주)
	크라우드 소싱	인포플라
	저작도구 개발	KINY.
	AI모델 개발	인포플라
	응용 서비스 개발	인포플라
가이드라인 작성	인포플라	최인묵
가이드라인 버전	v3.0 (2021. 2. 16)	

목 차

1. 데이터 명세 정보

1.1 데이터 정보 요약

데이터 이름	야외 실제 촬영 한글 이미지	
활용 분야	웨어러블카메라 및 모바일OCR(책표지 인식통한 온라인 구매) 등 야외에서 한글의 인식이 필요한 분야	
데이터 요약	일상에서 접할 수 있는 실내외 이미지 속 한글의 디지털 텍스트를 다량으로 확보하여, 문서나 필기체 외에 다양한 폰트로 노출되어 있는 한글자원을 효과적으로 활용하기 위한 인공지능 기반 OCR 기술 개발용 학습 데이터	
데이터 출처	간판, 책표지	
데이터 이력	배포버전	v3.0
	개정이력	개정
	작성자/ 배포자	작성 및 배포자 : 최인묵

1.2 데이터 포맷

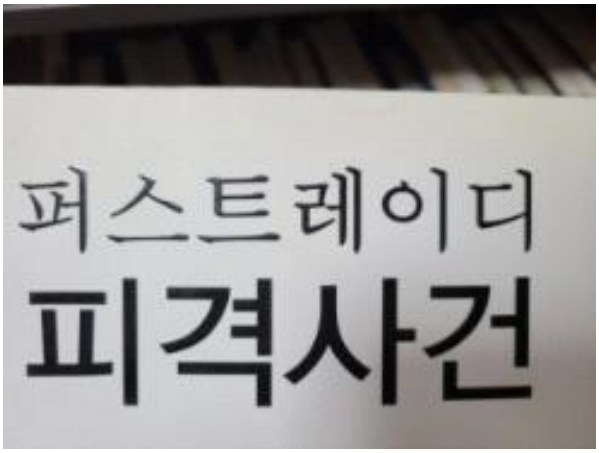
1.2.1 문자 인식 학습 데이터 : 간판 한글 인식(OCR) 학습용 데이터

○ JPEG + JSON 파일 쌍


```
{ "images": [{"id": 1,
               "width": 1600,
               "height": 1200,
               "file_name": "간판_가로형_000101.jpg",
               "date_captured": "2020-10-27 23:34:20"}],
    "annotations": [
        {   "id": 1,
            "image_id": 1,
            "text": "백씨네",
            "bbox": [360, 531, 863, 305]},
        {   "id": 2,
            "image_id": 1,
            "text": "커피가게",
            "bbox": [196, 849, 1224, 281]},
        {   "id": 3,
            "image_id": 1,
            "text": "xxx",
            "bbox": [16, 610, 116, 110]}
    ],
    (하략)
}
```

1.2.2 문자 인식 학습 데이터 : 책표지 한글 인식(OCR) 학습용 데이터

○ JPEG + JSON 파일 쌍




```
{ "images": [{"id": 1,
              "width": 1600,
              "height": 1200,
              "file_name": "책표지_문학_000001.jpg",
              "date_captured": "2020-10-27 23:50:46"}],
    "annotations": [
        {   "id": 1,
            "image_id": 1,
            "text": "퍼스트레이디",
            "bbox": [16, 394, 1426, 271]},
        {   "id": 2,
            "image_id": 1,
            "text": "피격사건",
            "bbox": [40, 699, 1377, 390]}
    ],
    (하략)
}
```


1.3 어노테이션 포맷

No.	항목		길이	타입	필수 여부	비고
	한글명	영문명				
1	**어노테이션정보**	**annotations**		**List**		
1-1	어노테이션 식별자	annotations[].id	10	Number	Y	
1-2	인식문자이미지식별자	annotations[].image_id	10	Number	Y	
1-3	어노테이션 텍스트	annotations[].text	1000	String	Y	
1-4	어노테이션 바운딩박스	annotations[].bbox	4	List		
2	**Crop라벨**	**cropLables**		**List**		
3	**이미지정보**	**images**		**List**		
3-1	이미지식별자	images[].id	10	Number	Y	
3-2	이미지너비	images[].width	4	Number	Y	
3-3	이미지높이	images[].height	4	Number	Y	
3-4	이미지파일명	images[].file_name	100	String	Y	

No.	항목		길이	타입	필수 여부	비고
	한글명	영문명				
3-5	이미지촬영일자	date_created	100	String	Y	
4	**데이터셋정보**	**info**		**Object**		
4-1	데이터셋명	info.name	100	String	Y	
4-2	데이터셋설명	info.description	1000	String		
4-3	데이터셋생성일자	info.date_created	100	String	Y	
5	**메타데이터**	**metadata**		**List**		
5-1	분류	metadata[].class	100	String	Y	ex)실외/실내 간판
5-2	세분류	metadata[].subclass	100	String		ex)가로형/세로형 간판
5-3	지역	metadata[].area	100	String		ex)서울경기권/충청권
5-4	수집장치	metadata[].device	100	String	Y	ex)스마트폰/DSLR
5-5	날씨	metadata[].weather	100	String		ex)맑음/흐림
5-6	조도	metadata[].illuminance	100	String		ex)밝음/중간/어두움
5-7	광원	metadata[].light	100	String		ex)간접광/자체발광/자연광
5-8	외곽선 선명도	metadata[].outline	100	String	Y	ex)상/중/하
5-9	글씨방향	metadata[].wordorientation	100	String	Y	ex)가로/세로
5-10	글씨크기	metadata[].wordsize	100	String	Y	ex)대/소
5-11	글씨폰트	metadata[].wordfont	100	String	Y	ex)인쇄체/캘리그라피
5-12	글자색	metadata[].wordcolor	100	String	Y	ex)단색/다색
5-13	글자연결	metadata[].wordconnection	100	String	Y	ex)띄어쓰기 있음/없음

○ 한글 이미지 다양성 확보를 위한 메타데이터 수집 항목. 다양한 조건의 한글간판이미지 수집을 통해 특정 환경에서만 동작하는 OCR을 지양하고, 향후 학습 및 인식의 난이도 조정 등 데이터 활용성을 높임.

- 지역 : 데이터 활용목적에 부합하는 지역 특성정보를 수집
- 수집장치 : Ground Truth외에 학습에 유용한 수집장치 정보 추가 라벨링
- 날씨 : 실외간판의 경우 날씨에 따른 간판 학습 및 인식의 연속성을 확보
- 조도 : 실내외 간판의 조도에 따른 간판 한글 이미지의 탐지 및 인식율 저하 방지를 위한 다양성 확보

- 광원 : 학습 및 인식에 지대한 영향을 끼치는 빛의 종류 메타데이터 확보
- 외곽선 선명도 : 글씨 구분의 중요 데이터로써 AI 및 non-AI 모두에서 OCR 학습 및 인식의 성능에 영향을 주는 항목 다양성
- 글씨방향 : 한글의 가로방향 및 세로방향 데이터 구분 수집
- 글씨크기 : 학습 및 인식의 정확성에 영향을 주는 대상의 크기 다양성 수집
- 글씨폰트 : 한글 간판의 인쇄체 및 캘리그라피(붓/펜글씨) 다양성 수집
- 글씨색 : 한글 이미지의 grayscale 또는 컬러 색상 다양성 수집
- 글씨연결: 글씨연결 구분을 통해 향후 단어 등 추가분리 학습의 처리가능성 확보

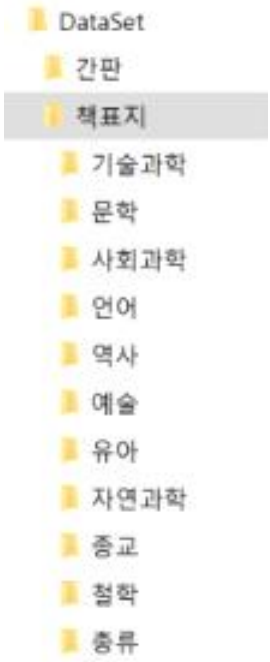
1.4 데이터 구성

1.4.1 간판 한글 인식(OCR) 학습용 데이터

분류	파일명
DataSet 간판 가로형간판 돌출간판 세로형간판 실내간판 실내안내판 지주이용간판 창문이용광고물 현수막	1.간판 이미지 파일규칙 • 간판_"세분류"_"일련번호".jpg • 예시 : 간판_가로형간판_000001.jpg 2.간판 어노테이션 파일규칙 • 간판_"세분류"_"일련번호".json • 예시 : 간판_가로형간판_000001.json

1.4.2 책표지 한글 인식(OCR) 학습용 데이터

분류	파일명

	1.책표지 이미지 파일규칙 • 도서_"KDC구분"_"일련번호".jpg • 예시 : 책표지_문학_000001.jpg 2.책표지 어노테이션 파일규칙 • 도서_"KDC구분"_"일련번호".json • 예시 : 책표지_문학_000001.json

1.5 데이터 통계

1.5.1 데이터 구축 규모

1.5.1.1 간판 인식 학습 데이터

○ 간판 사진은 전국 범위를 대상으로 낮, 밤, 우천, 눈 등 조건하에서 가로/세로형 간판, 돌출형 간판, 창문이용 간판 등을 중심으로 실내외 환경에서 여러 시야각도로 촬영을 목표로 함.

○ 최종 인공지능 데이터 : 간판 한글단어 바운딩박스 45만 건이 포함된 1600*1200 해상도의 한글 이미지들과 한글이 입력되어 있는 해당 이미지별 어노테이션 JSON 파일

- JPEG 이미지 약 450,000건 : 한글단어 45만 건에 해당하는 학습 데이터 구축
- 어노테이션 JSON 파일 약 450,000개 : 한글단어 45만 건 이상에 해당하는 학습 데이터 구축 완료
- 이미지와 JSON 파일 수량 비율은 1:1

1.5.1.2 책표지 인식 학습 데이터

○ 최종 인공지능 데이터 : 책표지 한글단어 바운딩박스 5만 건이 포함된 1600*1200 해상도의 한글 이미지들과 한글이 입력되어 있는 해당 이미지별 어노테이션 JSON 파일

- JPEG 이미지 약 50,000건 : 한글단어 5만 건에 해당하는 학습 데이터 구축
- 어노테이션 JSON 파일 약 50,000개 : 한글단어 5만 건 이상에 해당하는 학습 데이터 구축 완료
- 이미지와 JSON 파일 수량 비율은 1:1

1.5.2 데이터 분포

○ 간판 : 외부전문가 4인의 자문 및 옥외광고물 가이드라인 기준을 참고하여 실내외 간판으로 분류하고, 가로형 간판, 세로형 간판 등 8개의 세분류로 구분. 동일 간판에 대해 날씨, 조도, 시야각도 등의 다양성 수집 허용(최대 5장까지). 서로 다른 간판 이미지 중 사전에 기록된 명사 텍스트의 경우 중복을 허용하며, 사전에 비기재된 동일 고유명사 텍스트의 경우 중복 4건 이하만 허용

분류	세분류	촬영수량	비율
실외간판	가로형 간판	140,000	31.0%
	세로형 간판	30,000	6.7%
	돌출 간판	55,000	12.2%
	지주이용 간판	30,000	6.7%
	현수막	30,000	6.7%
	창문이용광고물	30,000	6.7%
	소계	315,000	70%
실내간판	실내간판	90,000	20%
	실내안내판	45,000	10%
	소계	135,000	30%
합계		450,000	100%

○ 책표지 : 한국십진분류표 (KDC)에 따라 분류간 유사비율로 촬영하며, 서로 다른 책표지 이미지 중 사전에 기록된 명사 텍스트의 경우 중복을 허용하며, 사전에 비기재된 동일 고유명사 텍스트의 경우 중복 4건 이하만 허용

분류(KDC)	촬영수량	비율
총류	2,500	5%
철학	3,500	7%
종교	4,000	8%
사회과학	8,000	16%
자연과학	2,500	5%
기술과학	8,000	16%
예술	2,500	5%
언어	4,000	8%
문학	8,000	16%
역사	3,000	6%
기타(유아)	4,000	8%
합 계	50,000	100%

1.5.3 기타 활용 통계

해당 없음

1.6 원시데이터 특성

1.6.1 대상분류

실제에 해당

1.6.2 제약조건

제약없음에 해당

1.6.3 속성

종류	내용	비고
자료형태	▪ 디지털 이미지	
사진이미지	▪ 간판 : 90% 비율로 수집(45만건 분량) ▪ 책표지 : 10% 비율로 수집(5만건 분량)	
원본형태	▪ 야외 텍스트	
파일포멧	▪ jpg, jpeg	
이미지해상도	▪ 1600*1200 이상	
이미지색상	▪ 컬러	
규모	▪ 약 500,000건	이미지 1건당 한글 단어 3건이하
가공여부	▪ 직접 촬영하여 후보정 작업을 거치지 않은 이미지 수집	무늬, 얼룩처럼 원본 속의 다양한 문자인식 방해요소가 포함되어 있는 학습용 데이터를 구축하여 한글 인식(OCR) 모델의 성능 강화 목적
중요성	▪ 자유도 최대의 한글 글씨체들로서 한글문자 인식 학습데이터 구축에 최적의 자료	
확보방안	▪ 직접 촬영하여 확보 계획	크라우드소싱 활용
정제방안	▪ 관리효율 향상을 위한 파일명 보완 및 분류, 해상도 확인 후 본격적인 학습용 데이터 구축 작업 착수 ▪ 촛점이 맞지 않아 문자 자체의 식별이 불가능한 이미지 제외	육안 검토
저작권	▪ 간판 : 저작권 자유 ▪ 책표지 : 저작권 자유	저작권 관련 사항 검토

1.7 기타정보

1.7.1 포괄성

○ 간판 : 8개 세분류 간판을 포괄하며, 지역, 수집장치, 날씨, 조도, 광원, 외곽선 선명도 등 활용목적에 부합하는 다양한 특성정보를 포함

○ 책표지 : KDC 분류의 책을 고르게 촬영하여 다양한 주제를 나타내는 한글이 포함되도록 함

1.7.2 독립성

○ 데이터가 타 법령 등에 대해 독립적임

○ 공개되는 학습용 데이터는 원시데이터에 대한 의존성 없음

1.7.3 유의사항

○ 데이터 배포시 부정적인 문제가 발생하지 않을 것으로 판단됨

○ 기대효과

▪ 야외 환경의 한글 디지털 텍스트라는 인공지능 학습데이터와 학습모델의 공유를 통해, 그동안 실내 환경의 한글이용에 머물러왔던 관련 솔루션과 서비스들의 확장을 촉진하고, 웨어러블 컴퓨팅, 모바일 인식 등 한글

을 이용하는 다양한 분야에서의 부가가치를 창출하여 관련 산업의 활성화를 통해 일자리 창출 도모

1.7.4 관련 연구

○ 2019년 text in the wild 조사예가 있음 (https://aihub.or.kr/aidata/133)

○ ICDAR, International conference Document Analysis and Recognition (http://icdar2021.org)

2. 데이터 구축 가이드

2.1 데이터 구축 개요

문서(font)나 필기체 외에 일상에서 접할 수 있는 다양한 한글 이미지를 이용하여 각종 비정형 한글을 인식함으로써 각종 솔루션에 사용될 수 있는 한글 이미지 학습 데이터 구축

구축단계	세부절차	필수여부	세부설명
수집	수집 대상 선정	필수	• 수집 데이터 정의 - 간판/도서 분야 - 각 분야별 분류 설정
	수집 Guideline 작성	필수	• 유효한 데이터 수집을 위한 가이드 제시 • 샘플 이미지 촬영 및 가공을 통한 사전 문제점 파악
	수집 담당자/기관 선정	필수	• 수집전문 인력 채용 및 크라우드 소싱 확보
	데이터 수집	필수	• 실내, 실외의 한글실제 이미지 데이터 촬영
정제	데이터 정제	필수	• 촬영된 이미지 데이터의 유효성 자체 검증
가공	1차 가공(영역 검출)	필수	• 이미지 데이터 내 한글 글자 영역 Bounding Box 표시
	2차 가공(한글 입력)	필수	• 표시된 영역 내 Label (한글텍스트) 입력
검수	교차 검사	필수	• 크라우드 소싱을 통한 검수 인력 채용 • 가공검수 체크리스트에 따른 이미지의 유효성 검수
	최종검수	필수	• 수행기관 및 검수업체를 통한 품질 인증

2.2 문제정의

2.2.1 임무 정의

○ 기존 OCR분야는 최적의 인식을 위해 노이즈 없는 인쇄체나 필기체를 학습 및 인식의 대상으로 집중해 왔으며, 따라서 실내 사무분야 등에서는 인공지능 인식이 성과를 보이고 있음. 그러나 실외 간판이나 책표지와 같은 비정형 폰트형태의 한글을 많이 사용하는 분야는 산업계의 관심 및 투자 부족과 기술개발의 난이도 등으로 해당분야 한글 이미지의 인식율이 매우 낮았음.

○ 이에 따라 본 과제에서는 인공지능 OCR엔진이 실내외에 위치한 간판, 책표지의 한글데이터를 인식할 수 있도록 학습데이터를 구축하고 학습모델을 개발하여 비정형 한글데이터의 인식율 향상을 노력하고자 함.

○ 1600*1200 해상도의 jpeg 이미지를 촬영하며, 야외(LSVT) 45만장, 책표지 5만장 총 50만장의 이미지 및 라벨링 json 구축 목표

○ 본 사업은 야외 간판 한글데이터 수집을 주요 목적으로 하고 있으나, 좀더 다양한 유형의 한글데이

터 구축을 위하여 비정형 한글 폰트를 사용하는 책표지의 한글데이터를 일부 포함하여 구축하기로 함.

2.2.2 데이터 구축 유의사항

○ 여러 수집자가 사진 이미지를 촬영하는 이유로 동일 또는 유사한 대상 이미지가 수집될 수 있음

○ 동일한 한글 텍스트가 추출되었으나, 서로 다른 이미지 중 사전에 기록된 명사 텍스트의 경우 중복을 허용하며, 사전에 비기재된 동일 고유명사 텍스트의 경우 중복 4건 이하만 허용. 간판 이미지의 경우 낮/밤, 조도, 광원, 시야각도 등의 다양성 수집의 경우에는 동일 텍스트에 대해 동일 간판 이미지 5장까지 허용함.

○ 저작권과 관련해서 간판 및 책표지 이미지는 외부 법률 전문가를 통해 문의한 바 문제가 없는 것으로 검토됨.

2.3 획득·정제

2.3.1 원시데이터 선정

○ 외부전문가 4인의 자문 및 옥외광고물 가이드라인 기준을 참고하여 실내외 간판으로 분류하고, 가로형 간판, 세로형 간판 등 8개의 세분류로 구분

분류	세분류	촬영수량	비율
실외간판	가로형 간판	140,000	31.0%
	세로형 간판	30,000	6.7%
	돌출 간판	55,000	12.2%
	지주이용 간판	30,000	6.7%
	현수막	30,000	6.7%
	창문이용광고물	30,000	6.7%
	소계	315,000	70%
실내간판	실내간판	90,000	20%
	실내안내판	45,000	10%
	소계	135,000	30%
합 계		450,000	100%

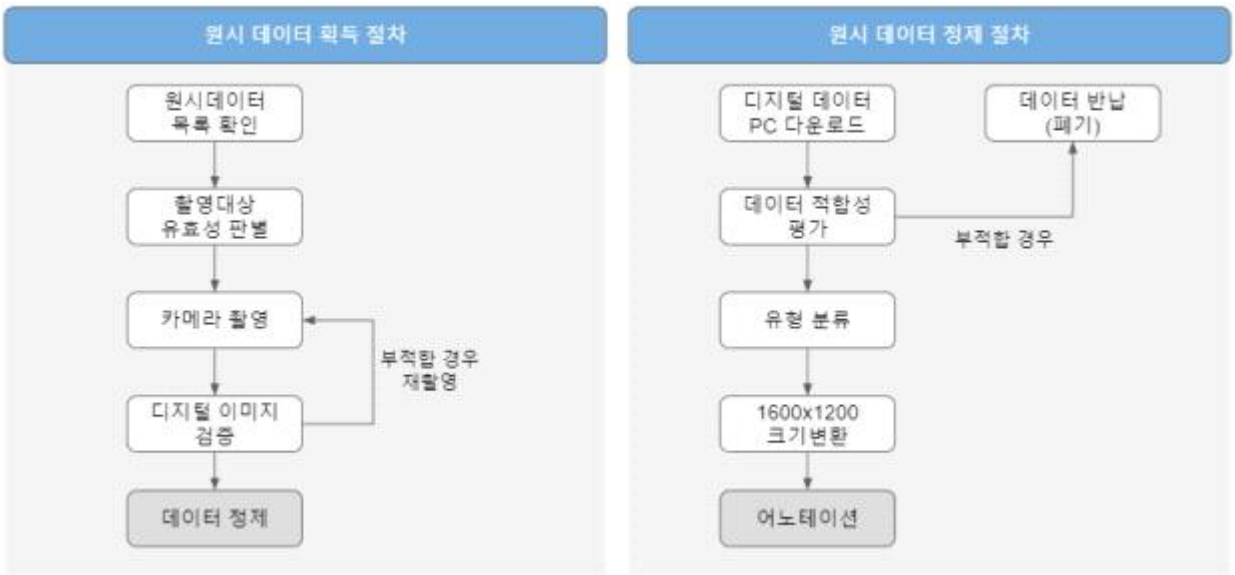
○ 책표지 : 한국십진분류표(KDC)에 따라 분류간 동등비율로 촬영

분류(KDC)	촬영수량	비율
총류	2,500	5%
철학	3,500	7%
종교	4,000	8%
사회과학	8,000	16%
자연과학	2,500	5%
기술과학	8,000	16%
예술	2,500	5%
언어	4,000	8%

문학	8,000	16%
역사	3,000	6%
기타(유아)	4,000	8%
합 계	50,000	100%

○ 동일 간판에 대해 날씨, 조도, 시야각도 등의 다양성 수집 허용(최대 5장까지)
○ 서로 다른 간판 이미지 중 사전에 기록된 명사 텍스트의 경우 중복을 허용하며, 사전에 비기재된 동일 고유명사 텍스트의 경우 중복 4건 이하만 허용

2.3.2 획득·정제 절차



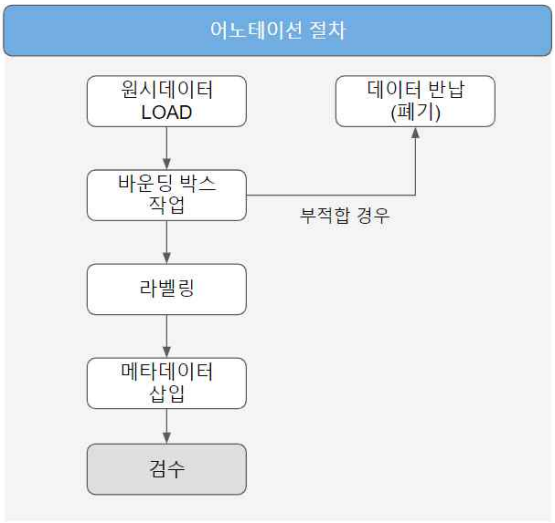
○ 사진 이미지 전체중 상하 또는 좌우 약 50%를 차지할 수 있도록 대상 이미지 촬영
○ DSLR 경우 사진 해상도 1600*1200으로 촬영하며, 스마트폰인 경우 해당 해상도의 설정이 불가능하므로 4:3 비율설정 및 1600*1200 이상의 해상도 설정 후 촬영
○ 스마트폰의 촬영 이미지는 PC로 옮긴 후 무료툴을 사용하여 1600*1200 해상도로 변환
○ 기존 등록된 이미지들과 중복검사

2.3.3 획득·정제 기준

○ 사진 촬영시 1600*1200 해상도로 촬영하며, 정면에서 상하좌우 30도 이내에서 촬영.
○ 거리 촬영시 안전확보를 위해 보도의 차도편 가장자리 수준에서 촬영하며, 이 경우는 1층의 간판 위주로 촬영. 충분한 거리확보가 가능할 경우 2층 등 고층의 간판촬영. 줌 촬영시 흔들림 및 해상도 고려함.
○ 이미지는 jpg/jpeg으로 저장하며, 파일 크기는 200KB~5MB.
○ 맑은 날과 흐린 날이 혼합되도록 촬영.
○ 동일 간판 사전 미등록 고유대명사 텍스트에 대해 서로 다른 이미지 4장까지만 중복을 허용.
○ 활용목적에 부합하는 환경, 지역, 시간 등 다양한 특성정보에 대한 다양성을 고려한 간판 원시데이터 수집. 동일 간판에 대해 날씨(맑음/흐림/비/눈), 조도(밝음/보통/어두움), 시야각도(정면/좌/우) 등의 다양성 수집 허용(최대 5장까지).

2.4 어노테이션/라벨링

2.4.1 어노테이션/라벨링 절차



2.4.1.1 바운딩박스

○ 한글이 포함된 이미지를 화면에 열어 이미지 속의 한글에 한 단어씩 또는 여러 단어에 박스를 그리는 절차

○ 어노테이션 툴(키니 제작)에서 사용자가 직접 박스 및 한글 텍스트를 생성

○ 이미지 내의 모든 한글에 바운딩박스를 쳐야 함

○ 바운딩박스는 해당 한글을 포함하는 가장 작은 박스를 그리는 것이 원칙임.

○ 바운딩박스 안에 한글, 숫자, 영어, 기호 등 목적하지 않은 다른 문자가 오면 안됨

○ 정상 사례 - 하단 "바운딩박스 정상 사례" 이미지에서

- "백씨네" : 유효한 한글 바운딩박스
- "커피가게" : 유효한 한글 바운딩박스
- "간" : 글씨가 너무 작아서 유효하지 않은 바운딩박스 (don't care 처리 -> xxx 표기)

○ 잘못된 사례 - 하단 "바운딩박스 잘못된 사례" 이미지에서

- "커피하우스" : 두줄이 하나의 박스로 묶여져 있는 잘못된 사례

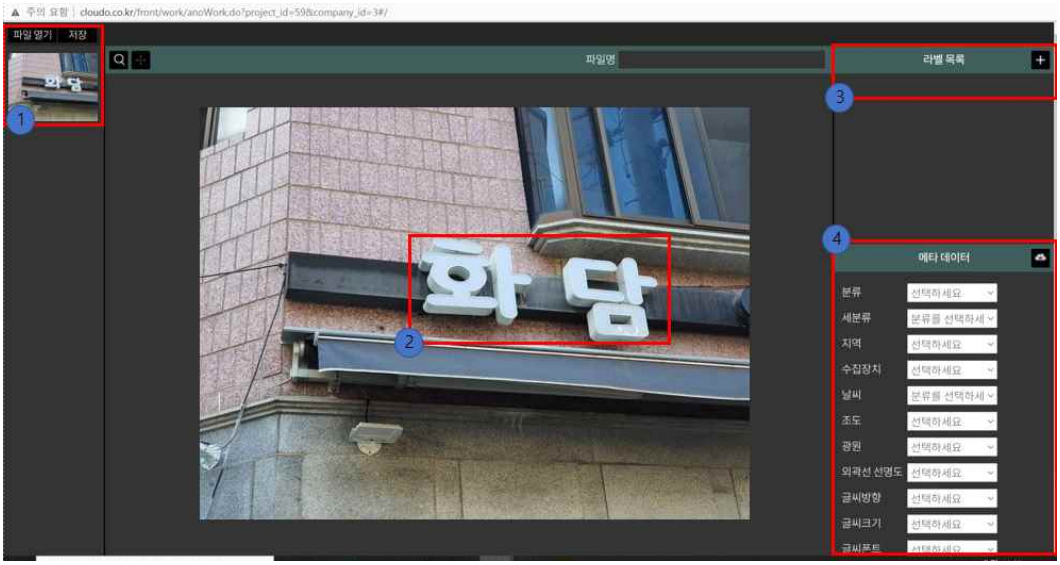
<바운딩박스 정상 사례>

<바운딩박스 잘못된 사례(여러줄 한박스)>

2.4.1.2 입력

○ 바운딩 박스에 해당하는 한글 텍스트를 입력하는 절차
○ 해당 바운딩 박스의 한글을 텍스트 입력창에 입력해줌 (사진속 3번 항목)
○ 그 후에 1번 위의 저장버튼을 누르면 어노테이션 과정 완료
○ 텍스트 입력을 통해 UTF-8 JSON 데이터 생산
○ 작업순서
- 작업순서 1 : 파일 읽기
- 작업순서 2 : Bounding Box 작업
- 작업순서 3 : 텍스트 입력
- 작업순서 4 : 메타데이터 입력

<저작도구 입력>



2.4.2 어노테이션/라벨링 기준

○ 이미지의 한글을 입력하여 JSON 생산
○ 이미지에 여러 한글이 있는 경우 가장 잘 보이는 한글 위주로 최대 3개까지 라벨링 작업실시
○ 이미지에 있는 모든 한글은 바운딩 박스가 그려져야 함
○ 활용목적에 알맞은 라벨링 유형 선택 및 적용의 관점에서 한글을 인지할 수 있는 한도내에서 글씨가 바운딩 박스에 닿아도 한글 학습에 영향 없으므로 이를 허용
○ 하나의 라벨링은 한글 10자 이하로 포함
○ 인식대상에 제외하는 한글의 경우 바운딩 박스 그린 후 don't care 처리 (xxx 표기)
○ 타이포그래피(그림문자) 역시 don't care 처리 (xxx 표기)

2.4.3 어노테이션/라벨링 조직

2.4.3.1 바운딩박스 조직

○ 한글 이미지에서 한글을 읽고, 마우스로 드래깅이 가능한 자
○ 한국어로 관리자의 피드백과 가이드 숙지가 가능한 자
○ 기본적으로 마우스 조작이 능숙한 자가 작업하는 것이 작업속도 확보에 유리함
○ 검수 단계에서 반려 처리시 촬영과 가공 (어노테이션) 작업 중 어느 작업에서 수정작업을 해야 하는지 관리가 어려우므로 수집, 정제, 가공 작업을 동일인이 하도록 유도

2.4.3.2 입력 조직

○ 한글 이미지에서 한글을 읽고, 타이핑이 가능한 자
○ 한국어로 관리자의 피드백과 가이드 숙지가 가능한 자
○ 수집, 정제, 가공 작업을 동일인이 하도록 유도

2.4.4 데이터 수집/정제/가공 작업 규칙

2.4.4.1 수집

○ 4:3 비율 1600x1200 이상의 고품질 사진을 스마트폰 또는 DSLR 카메라 등으로 촬영한다.
○ 전체한글이 사진의 상하 또는 좌우 넓이의 50% 이상이 되도록 촬영하여야 한다.
○ 동일피사체에 대해 저녁/밤, 비/눈/흐림, 정면/좌/우 촬영각도 등이 다른 다양성을 위한 사진이 허용된다.

2.4.4.2 정제

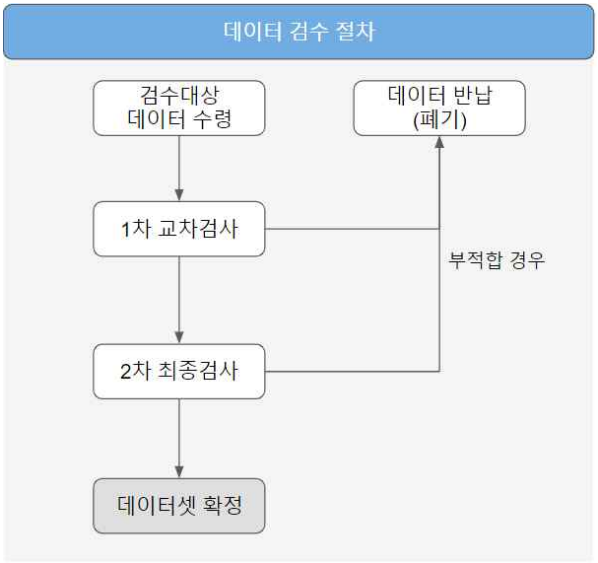
○ 1600x1200 사이즈로 변환되어야 한다.
○ 세분류 별로 분류되어야 한다.

2.4.4.3 가공

○ 사진 내의 모든 한글에 바운딩 박스를 그려야 한다.(유효박스 또는 don't care 박스)
○ 유효박스에는 사진의 주요 한글만 포함하여야 하며, 동일 한글을 라벨링하여야 한다.
○ 유효박스는 기호, 영어, 숫자, 다른 박스의 한글 일부라도 포함할 수 없으며, 한글의 모양을 크게 훼손하는 그림문자도 포함하지 않는다.
○ 유효박스는 가로 또는 세로 한줄 단위로 그려야 한다.
○ 유효박스는 한글의 형태가 유지되는 범위에서 내부의 한글에 닿거나 일부를 절단할 수 있다.
○ 유효박스는 최대 3개까지만 그리며(1개도 가능), 각 박스당 글자는 최대 10자 이내여야 합니다.
○ 유효박스 안의 한글크기가 50% 이상 크게 차이가 나는 경우 크기가 다른 한글을 분리하여 박스를 그린다.
○ don't care 박스는 한글이 너무 작거나, 잘 안보이거나, 한글이 유리에 비치는 경우 그리고 유효박스가 3개를 초과하는 등 유효박스 조건에 벗어나는 모든 한글 및 기호, 영어, 숫자를 포함할 수 있다.
○ don't care 박스는 영어 소문자 xxx로 라벨링한다.
○ don't care 박스는 가로 세로 여러줄을 하나의 바운딩박스로 그릴 수 있으며, 다른 박스의 일부를 포함할 수 있다.
○ 사진 속 주요 유효박스에 해당하는 메타데이터를 기술한다.

2.5 검수

2.5.1 검수 절차



○ 품질검수는 구축 데이터인 야외 촬영 한글 인식 학습데이터(간판, 책표지) 50만 장에 대하여 ①촬영 ②정제 ③입력의 3단계 공정을 모두 마친 후 실시함

○ 품질검수 수량 : 구축 및 공개 목표량인 50만 건의 전부

○ 검수는 전수교차검수, 최종검수 등 총 2차에 걸쳐 수행.

○ 일반 검수자가 데이터를 복수로 검증하여 중복 데이터 및 반려대상 데이터(바운딩박스 오류, 텍스트 오류)를 검수

○ 최종검수에서 샘플링 데이터(30%)에 대해 수행기관과 검수기관이 검수를 수행

○ 품질검수 절차

A. 촬영 이미지 검사 : 사진의 초점이 흔들리지 않아 촬영 피사체가 잘 보이는지 육안 식별하여 검사

B. 바운딩박스 유효성 검사 : 바운딩박스가 한글을 정상적으로 포함하고 있는지, 작업되지 않은 바운딩박스 대상이 없는지 육안 식별하여 검사

C. 텍스트 검사 : 바운딩박스안의 한글과 텍스트안의 한글이 동일한지 육안 식별하여 검사

2.5.2 검수 기준

○ 중복 데이터, 바운딩 박스 및 텍스트 이상 육안 확인

○ 전수교차검수, 최종검수를 통해 바운딩 박스 위치정보 이상, 이미지 포맷 불량 등 확인

#	데이터	구분		측정지표	검수기준	자동화	검사율
1	바운딩박스 데이터	정확도	구조 및 형식	▪ 바운딩박스 정확도 ▪ 단위 : 자료이미지 상의 한글 단어 또는 문장 수와 바운딩박스 수의 쌍 ▪ 방법 : 자료이미지 상의 한글 단어 또는 문장 수 대비 그려진 바운딩 박스 수의 비율 측정	▪ 정확도 99% 이상	수동 (육안)	100% 전수
2	문자인식 데이터	정확도	구조 및 형식	▪ 문자입력 정확도 ▪ 단위 : 이미지 상에서 바운딩박스의 한글 단어 또는 문장 1개와 입력된 디지털 텍스트 1개의 쌍 ▪ 방법 : 자료이미지 상의 총 문자수 대비 정확하게 입력된 문자수의 비	▪ 정확도 99.9% 이상	수동 (육안)	100% 전수

#	데이터	구분		측정지표	검수기준	자동화	검사율
				율 측정			

2.5.3 검수 조직

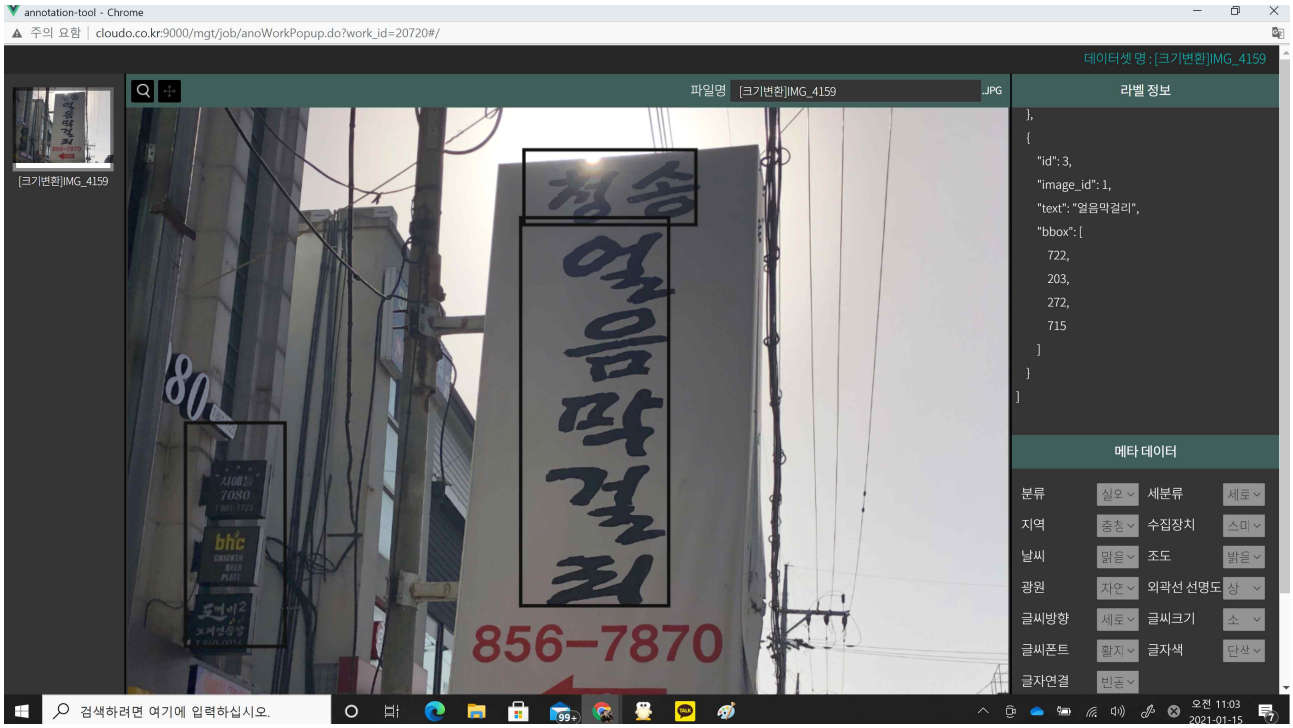
○ 1차 전수 교차검수는 크라우드 소싱을 통해 일반인들로 구성하며, 별도의 교육훈련 실시

○ 2차 최종 전수 검수는 자체 교육된 수행기관 인력과 검수담당 참여기관으로 구성

2.5.4 검수 도구(필요 시 작성)

○ 데이터 구축에 사용된 동일 컨소시엄의 참여사 키니앤파트너스의 어노테이션 저작툴 (Public AnnotationTool)을 사용

< 검수 화면 >



2.5.5 기타 품질관리 활동(필요 시 작성)

○ 야외 실제 촬영 한글 인식 학습데이터 구축 작업은 ①수집 ②정제 ③가공 ④검수의 4단계 절차를 거치게 된다.

○ 각 단계가 진행되면서 앞 단계의 결과물을 다음 단계에서 검사하는 전수검사 과정을 거치게 된다.

- 정제 : 앞 단계인 수집에서 촬영된 이미지의 상태와 해상도를 살피고, 필요한 경우 해상도를 변환한다. 이 과정에서 잘못 촬영된 이미지를 발견할 수 있다.
- 가공 : 앞 단계인 수집과 정제를 거치면서 생성된 이미지에 대해 바운딩 박스 및 텍스트 입력을 진행한다. 이 과정에서 잘못 촬영된 이미지와 잘못된 해상도 등 수집과 정제가 잘못된 이미지를 발견할 수 있다.
- 검수 : 앞 단계에서 생성된 json 한글 데이터셋을 살피고, 해당 한글과 이미지의 바운딩박스가

1:1로 올바르게 매칭되는지 살펴본다. 이 과정에서 잘못 가공된 데이터셋을 발견할 수 있다.

2.6 활용

2.6.1 활용 모델

2.6.1.1 모델 학습

○ 구글이 개발한 오픈소스 AI-OCR엔진 Tesseract를 통해 간판 한글인식 모델을 학습예정
○ Training from Scratch 방식을 사용예정이며, RNN의 일종인 LSTM 알고리즘을 사용할 계획
○ 학습데이터 60%, 검증데이터 10%, 테스트데이터 30% 분배계획

2.6.1.2 서비스 활용 시나리오

간판 촬영 및 인식	번역	결과 화면	사용후기
	 Translating...	 Source: 삼순이네(TTS) English : Samsun(TTS)	

○ 한국을 방문한 외국인이 간판인식 모바일 앱서비스에 로그인
○ 간단한 서비스 이용 안내문을 숙지
○ 한글 간판이 포함된 이미지를 모바일로 촬영하여 업로드
○ 서버에서 이미지속의 한글 간판이 인식되어 한글 텍스트로 변환된 후, 번역기를 통해 영문으로 변역된 내용을 앱 화면에서 확인. 한글 및 영어발음을 TTS로 읽어줌.
○ 외국인은 해당 간판을 영어로 확인
○ 해당 상점 이용후기를 작성가능

2.6.2 데이터 제공

○ 본 과제에서 생산되는 간판, 책표지 데이터는 별도의 자격 검증이 불필요
○ 데이터 공개 플랫폼에서 제시하는 기본 절차만 거치면 다운로드와 활용 가능