●○ OCR 과제

야외 실제 촬영 한글 이미지

●○ 개요: 야외 실제 촬영 한글 이미지데이터셋 소개

이 과제는 야외 이미지의 텍스트(LSVT, Large scale Street View Text) 읽기에 중점을 두며, 정해진 폰트나 필기체 외에도 일상에서 접할 수 있는 다양한 한글 이미지를 이용하여 간판, 메뉴판, 책표지, 상품명 등을 인식함으로써 다양한 서비스에 사용될 수 있는 이미지 데이터를 구축하는 것을 목표로 하며 시각장애인 보조도구, 모바일OCR, 웨어러블 카메라 등에서 활용될 수 있다

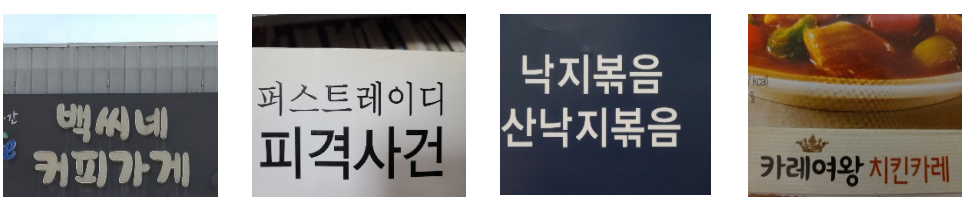


그림1 | 간판, 책표지, 메뉴판, 상품명 이미지 사례

●○ 데이터셋의 구성

본 데이터셋은 간판, 메뉴판, 책표지, 상품명 이미지들이 있고 그 이미지와 쌍을 이루는 JSON파일이 있다. 간판의 경우 다양성을 고려해 세분류 기준을 지역으로 정하였다. 건물외벽간판, 실외상가입구 간판, 돌출간판, 입간판, 현수막, 실내 간판, 시설안내 간판 등 다양한 종류와 글씨의 간판으로 구성되어 있고 총 45만장이 있다. 그 외 책 표지 4만8천장, 메뉴판 1천장, 상품 1천장, 총합 50만장으로 구성되어 있으며, 각 이미지마다 JSON파일이 쌍을 이루어 50만개의 파일이 구성되어 있다.

●○ 데이터셋의 구축 규모

1) 간판 인식 학습 데이터

- 최종 인공지능 데이터: 간판 한글단어 바운딩 박스 45만 건이 포함된 1600*1200 해상도의 한글 이미지들과 한글이 입력되어 있는 해당 이미지별 어노테이션 JSON파일
  - JPG 이미지 약 450,000건: 한글 단어 45만 건에 해당하는 학습 데이터 구축
  - 어노테이션 JSON 파일 약 450,000개: 한글단어 45만 건 이상에 해당하는 학습 데이터 구축
  - 이미지와 JSON 파일 수량 비율은 1:1

2) 책표지 인식 학습 데이터

- 최종 인공지능 데이터: 책표지 한글단어 바운딩 박스 4만8천 건이 포함된 1600*1200 해상도의 한글 이미지들과 한글이 입력되어 있는 해당 이미지별 어노테이션 JSON파일
  - JPG 이미지 약 48,000건: 한글 단어 4만8천 건에 해당하는 학습 데이터 구축
  - 어노테이션 JSON 파일 약 48,000개: 한글단어 4만8천 건 이상에 해당하는 학습 데이터 구축
  - 이미지와 JSON 파일 수량 비율은 1:1

3) 메뉴판 인식 학습 데이터

- 최종 인공지능 데이터: 메뉴판 한글단어 바운딩 박스 1천 건이 포함된 1600*1200 해상도의 한글 이미지들과 한글이 입력되어 있는 해당 이미지별 어노테이션 JSON파일
  - JPG 이미지 약 1,000건: 한글 단어 1천 건에 해당하는 학습 데이터 구축
  - 어노테이션 JSON 파일 약 1,000개: 한글단어 1천 건 이상에 해당하는 학습 데이터 구축
  - 이미지와 JSON 파일 수량 비율은 1:1

4) 상품명 인식 학습 데이터

- 최종 인공지능 데이터: 상품명 한글단어 바운딩 박스 1천 건이 포함된 1600*1200 해상도의 한글 이미지들과 한글이 입력되어 있는 해당 이미지별 어노테이션 JSON파일
  - JPG 이미지 약 1,000건: 한글 단어 1천 건에 해당하는 학습 데이터 구축
  - 어노테이션 JSON 파일 약 1,000개: 한글단어 1천 건 이상에 해당하는 학습 데이터 구축
  - 이미지와 JSON 파일 수량 비율은 1:1

●○ 데이터셋의 분포

- 간판: 전국 17개 시도의 간판 이미지를 전국 인구분포현황과 유사비율로 촬영하며, 서로 다른 위치의 간판 중 동일한 고유명사 텍스트의 중복 4건 이하만 허용
- 책표지: 한국십진분류표(KDC)에 따라 분류간 유사비율로 촬영하며, 서로 다른 책 표지 이미지 중 동일한 고유명사 텍스트의 중복 4건이하만 허용
- 메뉴판: 한식, 중식, 카페 등 음식분류별 유사비율로 촬영하며, 동일 메뉴판 이미지가 아닌 한 고유명사 텍스트 중복 4건이하만 허용
- 상품: 생활 주변의 상품을 촬영하며, 동일 상품 이미지가 아닌 한 고유명사 텍스트 중복 4건 이하만 허용

●○ 데이터 구조

자연 이미지는 현재 간판, 메뉴판, 책 표지, 상품 등으로 분류되어 있고 그 중에서 간판 이미지는 간판이미지의 다양성을 위해 전국에서 간판을 수집하였고 지역을 기준으로 간판을 분류하였다. 메뉴판이나 책표지, 상품명 이미지는 지역이 크게 의미가 없으므로 분류는 하지 않았다.

분류	간판	메뉴판	책표지	상품명
수량	45만	4만8천	1천	1천

No	항목		길이	타입	필수여부	비고
	한글명	영문명				
1	데이터셋정보	info		Object		
1-1	데이터셋명	info.name	100	String	Y	
1-2	데이터셋설명	info.description	1000	String		
1-3	데이터셋URL	info.url	200	String		
1-4	데이터셋생성일자	info.date_created	100	String	Y	
1-5	데이터날씨	info.weather	100	String	Y	
1-6	밤낮	info.sun	100	String	Y	
2	이미지정보	images		List		
2-1	이미지식별자	images[].id	100	String	Y	
2-2	이미지너비	images[].width	4	Number	Y	
2-3	이미지높이	images[].height	4	Number	Y	
2-4	이미지파일명	images[].file_name	100	String	Y	
2-5	이미지라이선스	images[].license	100	String	Y	
2-6	이미지촬영일자	date_created	100	String	Y	

No	항목		길이	타입	필수여부	비고
	한글명	영문명				
3	어노테이션정보	annotations		List		
3-1	어노테이션 식별자	annotations[].id	100	String	Y	
3-2	인식문자이미지식별자	annotations[].image_id	100	String	Y	
3-3	어노테이션 텍스트	annotations[].text	1000	String	Y	
3-4	어노테이션 속성	annotations[].attributes	1	Object		
3-5	어노테이션 바운딩박스	annotations[].bbox	4	List		
4	라이선스	licenses		List		
4-1	라이선스명	licenses.name	100	String	Y	
4-2	라이선스URL	licenses.url	200	String	Y	

●○ 데이터 예시

이 데이터는 간판 데이터 기준이며, 메뉴판, 책표지, 상품명 이미지는 아래 예시에서 weather, sun이 없는 구조를 가진다.

그림2 | 데이터 셋 이미지


```
{
    "images": [
        {
            "id": 1,
            "width": 1601,
            "height": 1200,
```

```
            "file_name": "20201017_111654.jpg",
            "date_captured": "2020-10-26 12:20:09"
        }
    ],
    "annotations": [
        {
            "id": 1,
            "image_id": 1,
            "text": "대창집",
            "bbox": [
                131,
                460,
                1290,
                479
            ]
        }
    ],
    "cropLabels": [],
    "info": {
        "name": "충남-2-8",
        "description": "",
        "weather":"맑음",
        "sun":"낮",
        "date_created": "2020-10-26 13:35:49"
    }
}
```

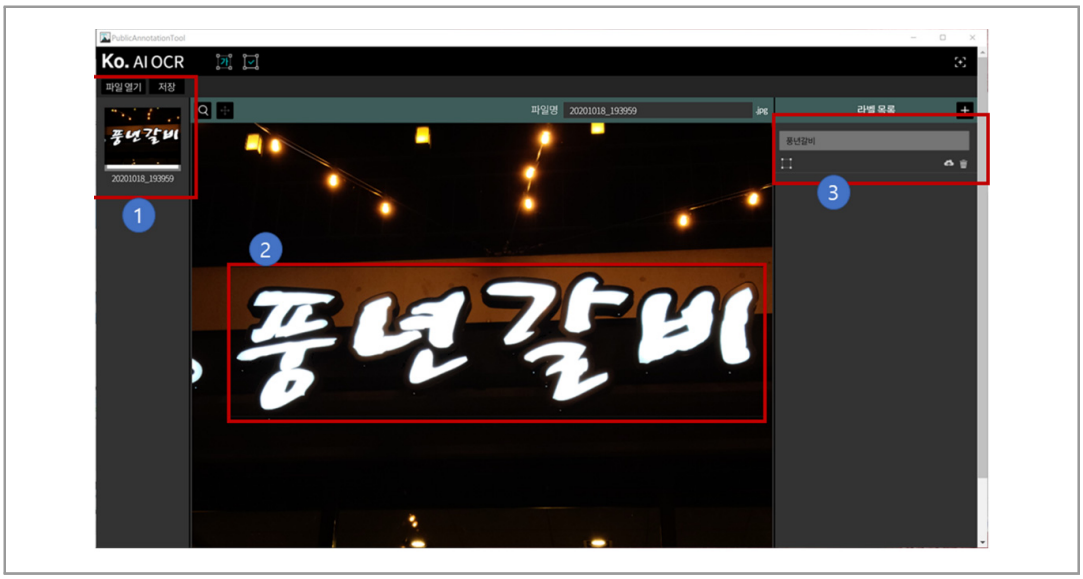

JSON 데이터 파일

※ 한 이미지에 텍스트는 여러 개일 수 있다. 생성규칙은 이미지의 [지역이름]_[일련번호]_N이다.

●○ 데이터 구축 과정

구축단계	세부절차	필수 여부	세부설명
수집	데이터 수집	필수	실내, 실외의 한글 이미지 데이터 촬영
정제	데이터 정제	필수	촬영된 이미지 데이터의 유효성 자체 검증
가공	1차 가공	필수	이미지 데이터 내 한글 글자 영역 Bounding Box 표시
	2차 가공	필수	표시된 영역내 Label(한글 텍스트) 입력
검수	전수 검사/교차 검사	필수	가공 검수 체크리스트에 따른 이미지의 유효성 검수
	최종 검수	필수	수행기관 및 검수업체를 통한 품질 인증

가공 단계에서는 어노테이션 저작툴(Public AnnotationTool)을 사용한다. 어노테이션 툴을 이용하여 사용자가 직접 박스 및 한글 텍스트를 생성하고 한글 텍스트를 입력한다.



작업 순서는 (1)파일읽기, (2)bounding Box 작업, (3)텍스트 입력 순으로 진행한다.

파일을 읽어 2번과 같이 바운딩 박스에 해당하는 한글 텍스트를 3번에 입력한다. 그후 1번 위의 저장버튼을 누르면 어노테이션 과정이 완료된다. 텍스트입력을 통해 UTF-8 JSON 데이터가 생산된다.

이미지의 한글을 입력하여 JSON을 생성할 때 이미지에 여러 한글이 있는 경우 가장 잘 보이는 한글 위주로 최대 3개까지 라벨링 작업을 실시한다 하나의 라벨링은 한글 10자 이하로 포함시켜야 하고 인식대상을 제외한 한글의 경우 바운딩 박스를 그린 후 don't care처리(xxx 표기)를 해준다.

●○ 검수와 품질 확보

검수는 구축 데이터 50만장에 대하여 공정을 모두 마친 후 실시한다. 전수 검수, 전수 교차 검수, 최종 검수 등 총 3차에 걸쳐 수행한다. 일반 검수자가 데이터를 복수로 검증하여 중복 데이터 및 반려대상 데이터를 검수하고 최종 검수에서 샘플링 데이터에 대해 수행기관과 검수기관이 검수를 진행한다. 사진의 초점이 흔들리지 않았는지, 바운딩 박스가 한글을 정상적으로 포함하고 있는지, 바운딩 박스안의 한글과 텍스트안의 한글이 동일한지 육안으로 식별하여 검사한다. 검수를 진행할 때는 어노테이션 저작툴(Public AnnotationTool)을 사용한다.