개요 : 한국형 사물 이미지 AI 데이터셋이란?

한국형 사물 이미지는 대한민국 내에 존재하고 유통되는 한국형 사물에 국한하여 학습용 데이터로 적합한 대상체로 정의하며, 국가지정 유적국가지정 유적건조물(탑, 성곽, 가옥, 궁궐, 사찰 등), 한국에서 제조되었거나 유통되는 상품(신발, 가방, 지갑, 잡화 등), 한국의 35개 도시 랜드마크를 범위로 한다.

한국형 사물 이미지 AI 데이터셋(Korean Type Object Image AI Training Dataset)은 사물 이미지 분야에서 유일한 대한민국의 유적건조물, 상품, 랜드마크(향후 범위 확장)에 특화된 분야 AI 학습 데이터셋으로서 (주)미디어그룹사람과숲에서 구축했으며, 2020년 11월 기준 총 360만장의 이미지로 구성되어 있다.

본 데이터셋의 구축목적은 대한민국의 건축물, 유형의 상품 등 실질적 형태를 지닌 사물에 대한 이미지를 인공지능이 학습할 수 있는 데이터 형태로 구축하여 사물 이미지 검색, 추천, 활용하는 서비스를 만드는데 도움을 주고, 산업발전 측면에서는 스마트관광, 스마트스토어 등 4차 산업에 수반되는 사물 임지의 데이터셋을 구축하기 위함이다.

한국형 사물 이미지 데이터셋의 대표적인 활용분야는 관광, 쇼핑, 건축, 제품개발, 디자인특허, 4차 산업 활용 등 다양한 분야에서 널리 쓰일 수 있다.

스마트 관광의 경우, 한국형 사물 이미지 데이터셋으로 학습된 AI 와 지도정보, 숙박 및 음식점 정보, 교육 정보 등 추가적인 정보를 활용하여 학습된 데이터에 한하여 외국인 관광객 (일반인 포함) 의 관광 안내, 여행지 주변 정보 제공, 청소년 현장 교육에 활용을 목적으로 활용 할 수 있다.

스마트 쇼핑의 경우, 한국형 사물 이미지 데이터셋의 상품 데이터셋 학습을 통한 무인 상품결제 시스템의 상품 검색 및 정보 출력에 활용 할 수 있으며, 학습된 상품에 한하여 상품 이미지를 업로드하여 유사한 이미지의 상품 검색 및 상품에 대한 정보와 구매 페이지로 이동할 수 있다.

대표적인 한국형 사물 이미지 데이터셋 활용사례는 유적건조물 및 랜드마크 데이터셋을 활용한 '문서작성 도우미' 서비스와 2021년 상반기 오픈 예정인 AI 상품 검색 시스템 'AI 캐셔' 가 있으며, 점진적인 서비스 확대를 통해 적용범위를 확장하고 있다.

데이터셋의 구성

데이터셋의 분류체계는 유적건조물, 상품, 랜드마크 총 3개 분야로 구분하고 상세설명, 대상체, 구축량, 제공 방식에 대한 설명은 다음과 같다.

분야	상세설명	대상체	구축량	제공 방식
유적건조물	1. 서울/경기/인천/경주에 위치한 국가지정문화재 2. 종교신앙, 정치국방을 포함한 15개 분야 3. 궁궐, 가옥, 탑, 무덤, 사찰, 교회, 성곽, 성당 등	약 840개	260만장	이미지, JSON 포맷 파일
상품	1. 외국관광객 쇼핑구매 우선순위 고려 2. 신발, 가방, 화장품, 악세서리 포함 8개 분야 28개 아이템 3. 금속, 화장품, 시계, 악세사리, 신발, 가방, 지갑, 모자, 아이웨어	약 251개	80만장	이미지, JSON 포맷 파일
랜드마크	1. 각 도시의 주요 상징물 2. 총 35개 도시 랜드마크 대상체 3. 동상, 타워, 전망대, 빌딩, 대교, 센터, 역사, 정자, 성문, 등대	약 66개	20만장	이미지, JSON 포맷 파일

데이터셋의 설계 기준과 분포

데이터셋을 설계할 때 가장 중요하게 고려했던 점은 분류체계이다. AI 학습을 위해 360만장의 한국형 사물 이미지에 대한 종별 수량, 다양성, 품질수준을 만족시킬 수 있는 분류체계 기준을 수립 및 적용한다.

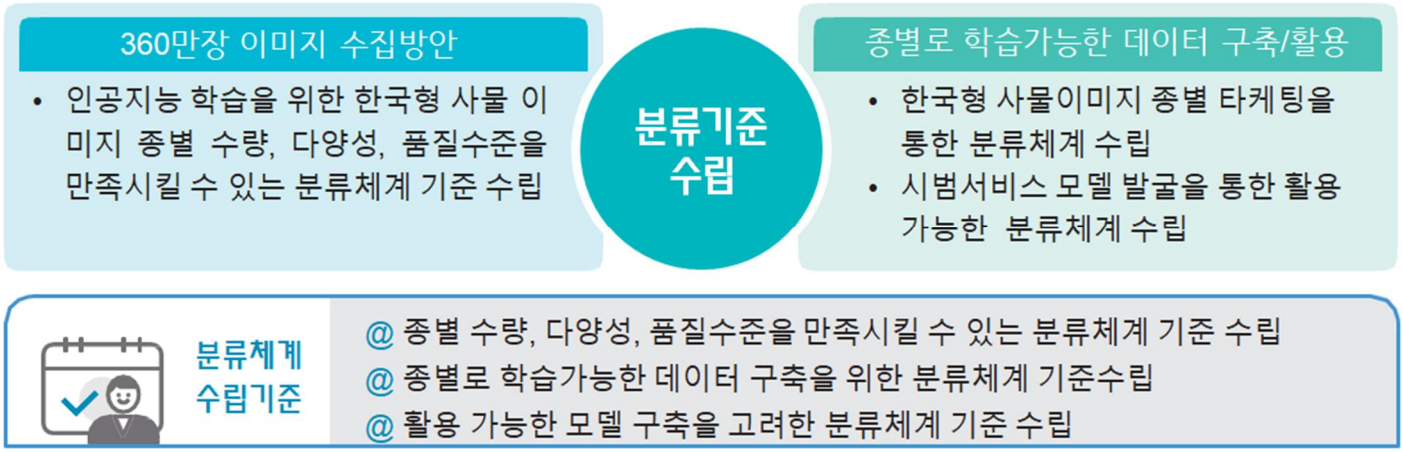


그림 1 분류체계 기준수립

데이터셋의 분류체계는 대분류, 중분류, 소분류로 구분하며 각 분류별 분포는 다음과 같다.

구분	대분류	중분류	소분류
명칭 및 수량	유적건조물	14개	30개
	상품	8개	28개
	랜드마크	35개	35개
합계	**3개**	**57개**	**93개**

데이터 구조

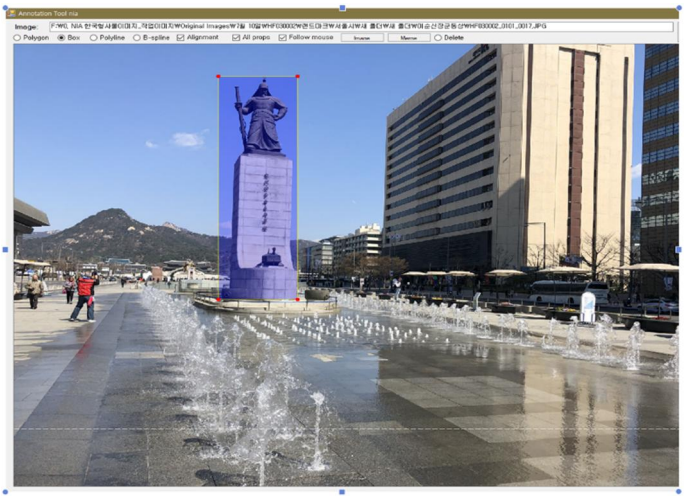
데이터셋에 따른 항목과 해당 값은 다음 테이블과 같다.

구분	필드	속성
image	identifier	파일명
	imsize	이미지 파일 크기
	copyright	이미지 저작권 소유자
	date	데이터 취득 일자
	resolution	해상도
	bit	비트값
	F-Stop	조리개 투과량
	exposure time	노출 시간
	ISO	ISO 감도
	focal length	초점 거리
	full aperture	조리개 최대 개방 수치
	view angle	화각
	white balance	화이트 밸런스

	depth	이미지 mode(RGB, Grayscale, bitmap)
regions	type	어노테이션 종류
	boxcorners	어노테이션 좌표 값
	ansize	이미지 영역 사이즈
	class	클래스명
	tags	분류항목
	truncated	대상체 잘림 여부
	종ID	종 아이디 찡보
	대분류	대분류 정보
	중분류	중분류 정보
	소분류	소분류 정보
	Instance	대상체
	Instance Upper	대상체 상위 레벨
	GPS 위도	GPS 위도 값
	GPS 경도	GPS 경도 값
	instance_uri	온톨로지 uri 정보
	sem_ext	의미확장 정보
	property : locatedIn	의미확장 지역정보
	property : relatedTerm	의미확장 연관정보

데이터 예시

이 샘플 데이터는 이순신 장군상에 대한 이미지 어노테이션을 수행한 JSON 형태의 데이터셋이다.



```
{
 "image": {
   "identifier":  "" , //파일명
   "imsize": [ 이미지 가로 사이즈, 이미지 세로 사이즈 ]
   "copyright":  "" , //저작권 정보
   "date":  "" , //촬영일
   "resolution":  "" , //해상도
   "bit":  "" , //비트값
   "F-Stop":  "" , //조리개 투과량
   "exposure time":  "" , //노출 시간
   "ISO":  "" , //ISO 감도
   "focal length":  "" , //초점 거리
   "full aperture":  "" , //조리개 최대 개방 수치
   "view angle":  "" , //화각
   "white balance":  "" , //화이트 밸런스
   "depth":  "" //RGB 여부
},
 "regions": [ {
   "type":  "" , //어노테이션 종류
   "boxcorners": [좌측 X,Y좌표/ 우측 X,Y좌표]
   "ansize": [ 이미지 가로 사이즈, 이미지 세로 사이즈 ]
   "class":  "" , //클래스명
   "tags":  //분류항목
  [  "truncated:0" , //대상체 짤림 여부
     "종ID:" ,
     "대분류:" ,
     "중분류:" ,
     "소분류:" ,
     "Instance:" , //대상체",
     "Instance Upper:", //대상체 상위",
     "GPS 위도:" ,
     "GPS 경도:" ]
     "instance_uri":  "" , //온톨로지 uri 정보
     "sem_ext": [ { // 온톨로지 의미확장 정보
     "property":  "locatedIn" ,  // 온톨로지 의미확장 지역정보
     "value":  ""
     }, {
     "property":  "relatedTerm" ,  // 온톨로지 의미확장 연관정보
```

```
    "value": ""
    }, {
    "property": "description" , //온톨로지 의미확장 상세정보
    "value": ""
    } ]
} ]
```

데이터 구축 과정

데이터 구축은 데이터 설계 및 수집, 정제, 가공, 확장, 검증의 프로세스를 거쳐 체계적으로 진행하였다. 데이터 수집은 직접촬영, 웹크롤링을 통한 자동수집, FTP를 통한 직접수집 방식을 혼용하였고, 데이터의 신뢰성과 작업효과성을 높이기 위한 위험관리, 의사소통관리, 품질관리, 진척관리 등 품질관리 활동을 추가 시행한다.

그림 2 데이터 구축 프로세스

효율적인 데이터 구축 작업을 위해 JAVA 기반의 이미지 어노테이션 툴을 자체 개발하여 적용하였다. 툴은 이미지를 불러오고 바운딩 박스 등 어노테이션 작업을 한 후 라벨링을 하는 방식으로 구성한다.

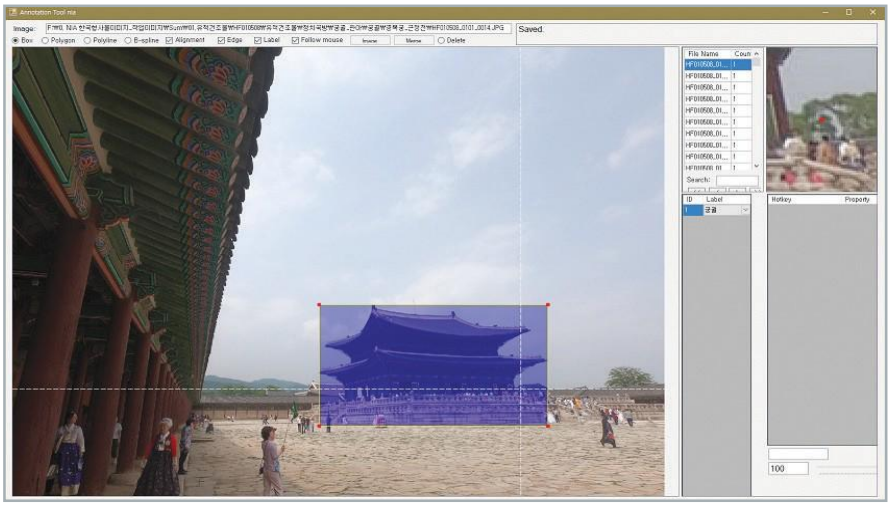


그림 3 효율적인 데이터 구축을 위한JAVA 기반의 이미지 어노테이션 툴

이미지 유사도 분류 툴을 활용하여 이미지 유사도 분류 및 유사 이미지 중복성을 제거하였다. 이러한 이미지 전처리 자동화를 통해 작업의 효율성과 정확도를 높인다.

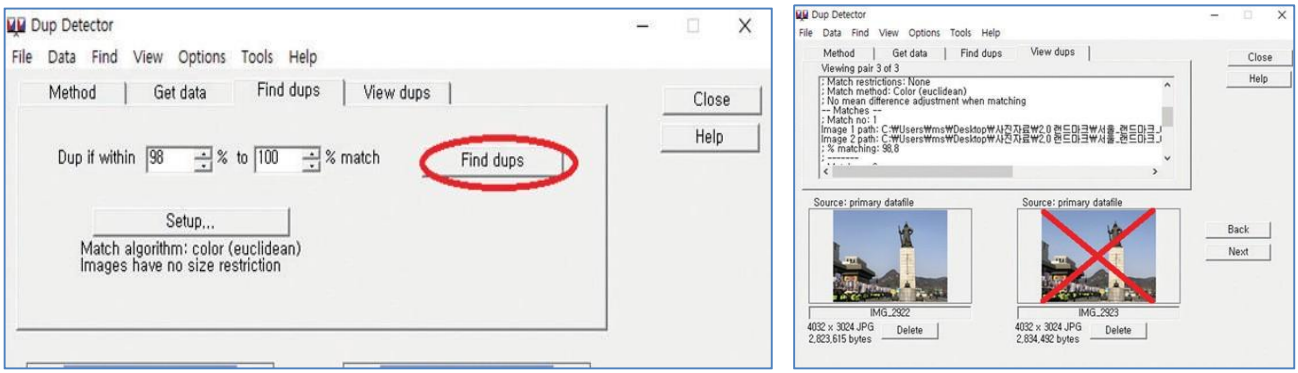


그림 4 이미지 유사도 분류 툴

검수와 품질 확보

품질관리를 위해 품질관리 체계 및 프로세스를 정립하여 적용하였고, 공공데이터 품질관리 매뉴얼 절차를 준수하여 진행하였다. 데이터 프로파일링 기법을 적용하고 단계별 검증 및 피드백을 거쳤다. 데이터 프로파일링을 위한 데이터 전문가 및 품질진단을 위한 품질관리 전문가 등 분야별 업무전문가를 활용하여 품질수준을 높인다.

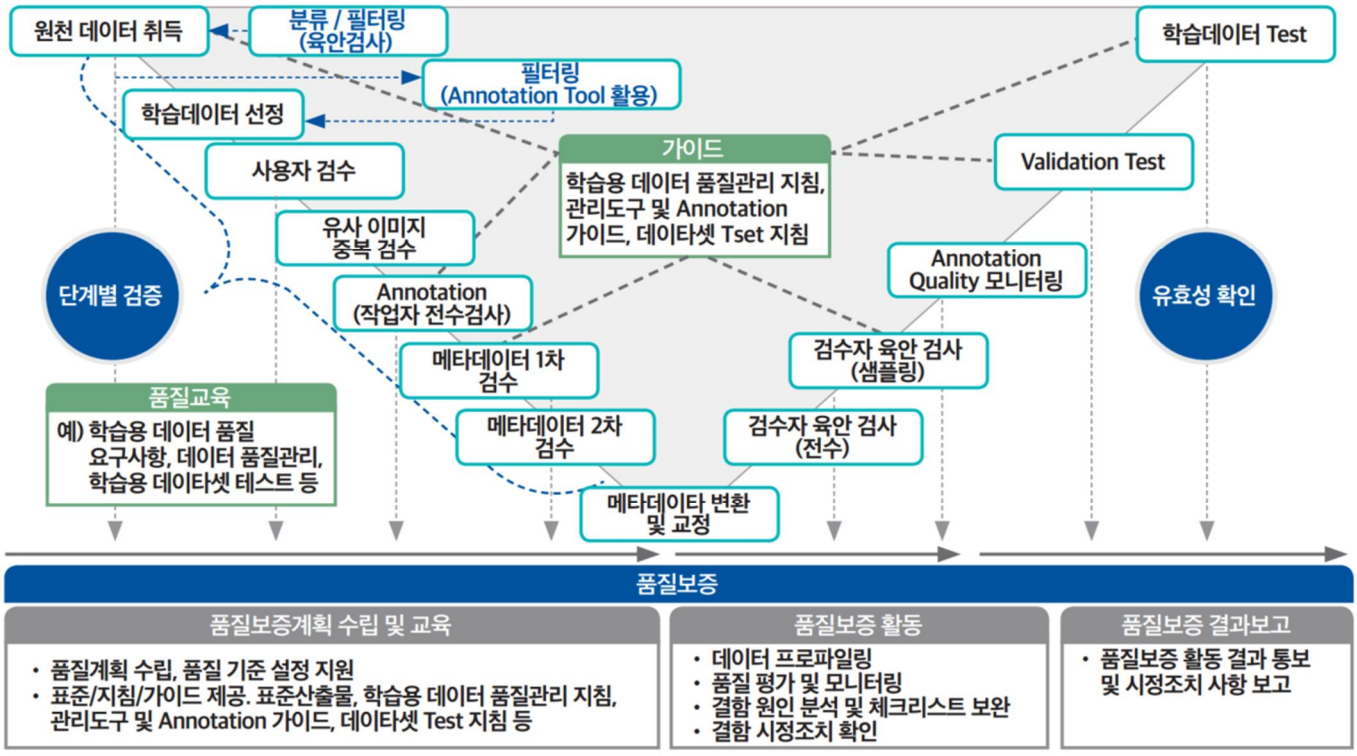


그림 5 데이터 품질관리 체계 및 프로세스

구축 공정별 검수절차, 검수방법, 목표기준율은 다음과 같으며, 원천데이터 취득부터 검증 및 테스트까지 전과정을 체계적으로 검수하여 데이터의 신뢰성을 확보한다.

구축 공정	검수절차	검수방법
원천데이터 취득	- 원천데이터의 분류를 목적으로 육안으로 검수	- 검수자 육안 검사
학습데이터 선정	- 프레임 연속성을 고려하여 카테고리 기준 수집목표 종에 대하여 선정 - 유효 학습데이터 선정	- 자동추출 도구 활용
사용자 검수	- 협회 및 전문단체를 통한 수집(FTP)에서 수집된 이미지에 대한 적합성을 검수	- 육안 전수 검사
유사 이미지 중복검수	- 유사도 분류 툴을 활용하여 중복 이미지 검수	- 자동화 전수 검사
메타데이터 1차 검수	- 입력된 메타데이터에 대한 20% 샘플링 검수	- 샘플링에 의한 검사
메타데이터 2차 검수	- 데이터베이스 카테고리에 대하여 카테고리 적합성 확인	- 전수검사

학습	- 4개월 단위로 테스트 데이터를 활용하여 학습데이터 정확도 평가	- 자동화 CNN 모델 활용
검증 및 테스트	- 4개월 단위로 학습 데이터셋으로 학습하고 검증 데이터를 이용하여 학습 데이터 정확도 평가 - 최종단계 유효성 검증 실패를 대비하여 구축 초기 단계부터 점진적 검증 진행	- 자동화 CNN 모델 적용 - 초기 학습용 데이터 30%로 선 학습 진행 후 10% 단위 검증 진행

데이터 구축 담당자

수행기관(주관) : ㈜미디어그룹사람과숲 (전화: 02-6959-6632), 이메일: lsc@humanf.co.kr