테크니컬 리포트

2020년 1차
인공지능
학습용
데이터 구축

기타 영역

한국인 재식별 이미지

개요: 한국인 재식별 데이터셋이란?

본 데이터셋은 보행자 재식별 기술 개발에 활용할 수 있는 한국인 재식별 데이터셋으로 한국과학기술연구원(KIST), ㈜휴먼ICT, SQI소프트㈜에서 구축하였으며, 보행자 1,000명 이상 총 400만장 이상의 보행자 이미지와 속성정보로 구성되어 있다.

재식별 기술은 다수의 CCTV에서 특정 보행자를 검색할 때 주로 사용되며, 특정 카메라에서 검출된 인물의 정보를 활용하여, 다른 시점/공간의 CCTV에서 해당 인물을 검색하는 기술이다. 해당 기술은 CCTV 관제시스템의 지능화를 위해 반드시 필요한 기술이며, 이에 대한 활발한 연구 및 상용화를 진행하고 있다. 재식별 기술의 사례에 대해서는 아래를 참고할 수 있다.

(a) 사람의 구조 정보 [1]

(b) 카메라 위치 정보 [2]

그림 1 재식별 연구 수행을 위한 정보의 예

[1] G. Zhao et al., "Improving person re-identification by body parts segmentation generated by gan". IEEE IJCNN (pp. 1-8), 2018.

[2] M. Gou et al., "DukeMTMC4ReID: A large-scale multi-camera person re-identification dataset".IEEE CVPRW (pp. 10-19), 2017.

데이터셋의 구성

본 데이터셋은 다양한 성별, 연령별, 의상, 소지품 등을 고려하여 한국에 거주하는 사람 1,000명을 대상으로 촬영한 실영상 CCTV 데이터라고 할 수 있다. 총 1,000명의 섭외된 연기자를 대상으로 촬영을 진행하여, 1명당 최소 10개의 CCTV에 10초 이상 노출 (30fps 기준 초당 10프레임 추출) 하는 것을 기본으로 하며, 다양한 환경(실내/외) 및 시간대(낮/일몰), 옷차림, 마스크, 소지품 등을 고려하여, 총 400만장 이상의 이미지로 구성되어 있다.

본 데이터셋은 개인정보 보호법 침해를 예방하기 위하여, 피촬영자 (섭외 연기자)를 대상으로 개인정보활용동의서를 수집하였으며, 해당 동의서에는 주관기관인 한국과학기술연구원에서 AI Hub 사이트(한국정보화진흥원; NIA)로의 데이터 배포 및 AI Hub사이트에서 AI Hub 사이트 회원으로의 재배포 동의를 받도록 하였다.

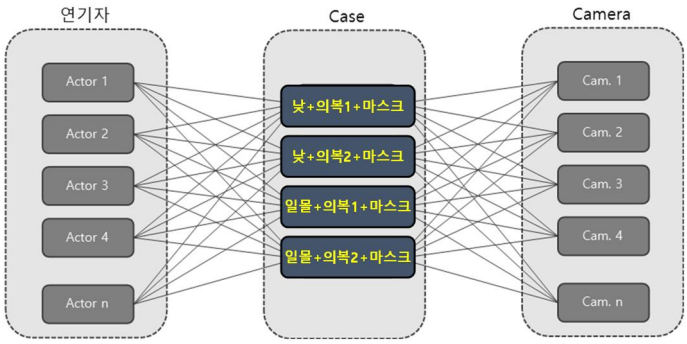

그림 2 실외 환경 데이터셋 구성

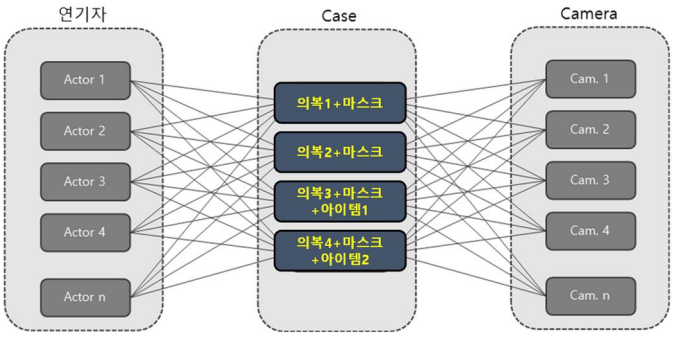

그림 3 실내 환경 데이터셋 구성

데이터 종류	포함 내용	제공 방식
보행자 Crop 이미지	약 400만장 1명당 4,000장의 이미지 생성 (CCTV 10개 x 10초 노출 x 10 frames x 4 Cases)	PNG
보행자 속성 정보	CCTV 및 보행자 정보	XML

데이터셋의 설계 기준과 분포

한국인 재식별 데이터셋은 기존 상용 데이터셋 대비 이미지 수량, 카메라 개수, ID당 평균 보행자 노출 카메라 개수, ID 당 평균 영상 수, 촬영환경, 마스크 착용, 의상 변화 등 다양한 측면이 고려되어 다양한 재식별 연구에 활용될 수 있도록 하였다.

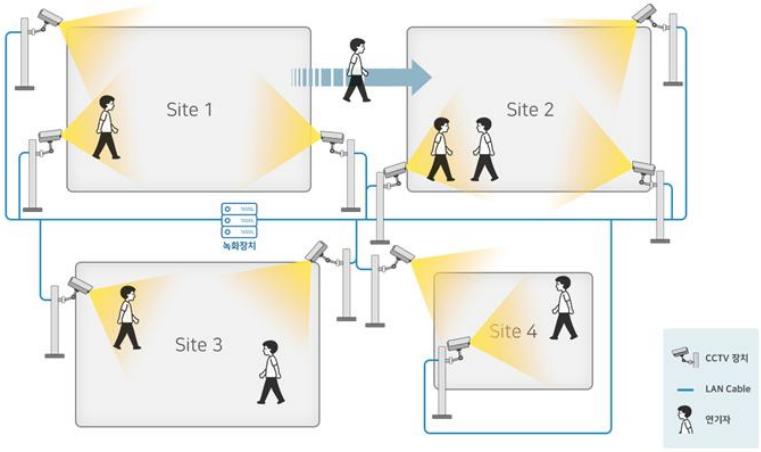

그림 4 재식별 데이터셋 구축 예시

- 다양한 장소에 촬영용 카메라 배치
- 다수의 모델이 모든 카메라에 노출되도록 동선 배치
- 실외/내 환경을 구분하여 촬영

또한, 실제 CCTV에서 촬영되는 상황을 고려하여, 실외와 실내비율 60:40, 성별 50:50, 연령비율 20~40대 (90%), 10, 50, 60대 10%로 구성하였다.

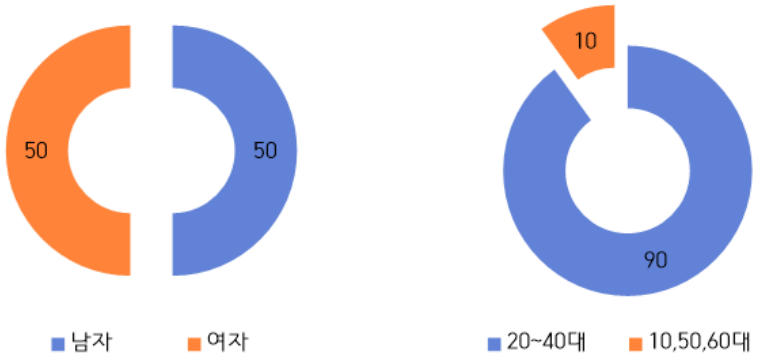

그림 5 한국인 재식별데이터셋 성별 및 연령대별 비율

데이터 구조

본 데이터셋에 포함되어 있는 항목과 정보는 아래와 같다.

데이터베이스		**한국인재식별데이터셋**
공개여부		**O**
인원 수 (ID 수)		**1,000** **(2,000+)***
영상 수 (Patch 수)		**4,000,000+**
카메라 갯수		**20**
ID당 평균 노출 카메라 갯수		**10+**
ID당 평균 영상 수(반올림)		**4,000+**
crop size		**Vary**
촬영환경		**실외, 실내**
Multi-shot**		**O**
동일인의상변화		**O**
마스크 착용		**O**
색인정보	ID	**O**
	CAM ID	**O**
	시간대 (낮,일몰)	**O**
	성별	**O**
	키	**O**
	나이	**O**
	의상정보	**O**
	의상색상	**O**
	액세서리	**O**
	머리길이	**O**
	비고	액세서리: 모자, 안경, 선글라스, 마스크, 가방, 우산, 핸드폰, 캐리어, 서류가방, 기타 등

본 데이터셋의 저장 구조는 아래와 같다.

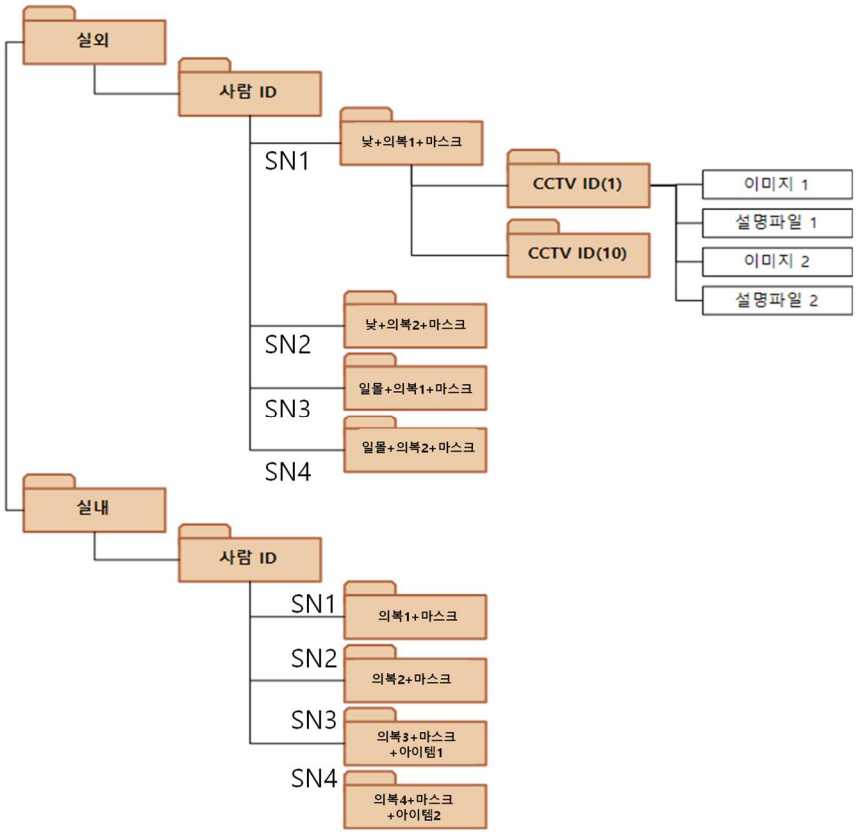

그림 6 데이터셋 저장 구조

데이터 예시

본 재식별 데이터셋의 이미지 별 속성정보는 아래와 같은 구조를 가진다.

① Camera List: CCTV 카메라 목록 및 정보 (Camera_List.xml)

속성	설명	값 예시
카메라 ID	카메라 구분	207041
GPS 좌표	GPS 좌표 (위도, 경도)	37.401594, 126.972768
방향	카메라가 바라보는 방향	NE(North East)
해상도	동영상 해상도	SD, HD, FHD

② Human List: 보행자 목록 및 정보 (Human_List.xml)

속성	설명	값 예시
ID	배우의 Unit ID	H0000~H9999
성별	배우의 성별	male
나이	배우의 나이	23

신장	배우의 신장	170
머리 형태	헤어스타일	normal
머리 색상	검정, 노랑, 갈색, 흰색	black
상의 타입	긴팔, 반팔, 나시, 원피스	long_sleeve
정의된 상의 색상	정의된 상의 색상 유무	True (1)/False (0)
상의 색상	배우의 상의 색상	red
하의 타입	긴바지, 반바지, 치마	long_pants
정의된 하의 색상	정의된 하의 색상 유무	True (1)/False (0)
하의 색상	배우의 하의 색상	navy

③ Item List: 소지품 목록 및 정보 (Item_List.xml)

속성	설명	값 예시
ID	배우의 Unit ID	0000~9999
종류	모자, 안경, 썬글라스, 마스크, 가방, 우산, 핸드폰	cellphone
색상여부	정의된 상의 색상 유무	true
색상	악세서리의 색상	black

데이터 구축 과정

본 데이터 구축은 2020년 7월부터 11월까지 다양한 실내외 장소에서 실제와 동일한 CCTV 장비를 구축하여, 섭외된 연기자를 촬영하고, 이를 정제 및 가공, 검수하여 구축하였다.

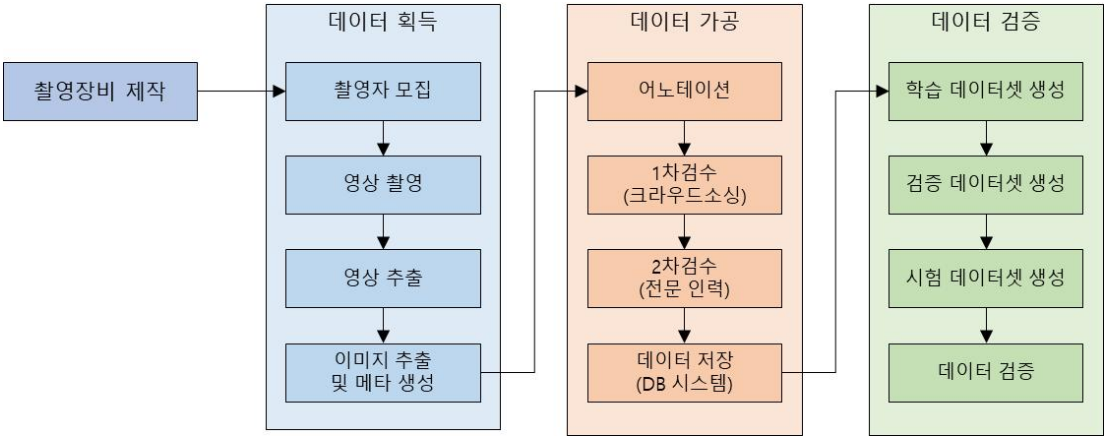

그림 7 데이터 구축 개요

한국인 재식별 데이터셋 구축을 위한 개요는 위의 그림과 같이 촬영 장비 제작, 데이터 획득, 데이터 가공, 데이터 검증의 순서를 따른다. 먼저 실내외 한국인 재식별 데이터셋 구축을 위하여, CCTV와 동일한 환경으로 촬영이 가능하도록 촬영장비를 제작하였다. 개인정보가 보호된 데이터 획득을 위하여, 촬영자를 모집하고 개인정보활용 동의서를 취득한 이후에 촬영을 진행하며 이에 대한 영상 및 이미지 추출, 메타데이터를 생성하였다.

그림 8 촬영용 CCTV 장치 예시 및 기제작 시제품 모습

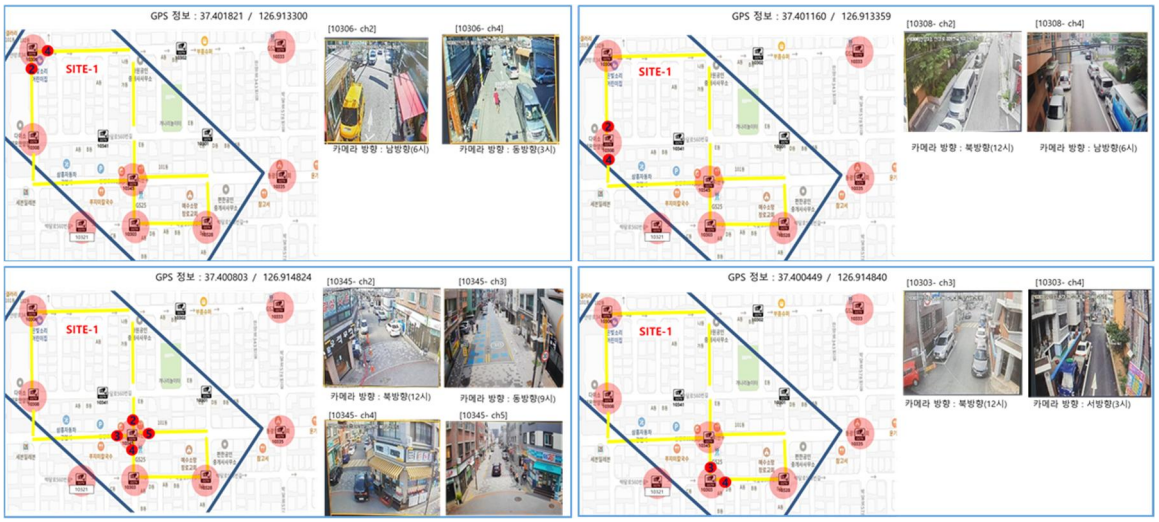

그림 9 실내외 동선 설계 및 촬영

데이터 가공을 위한 어노테이션을 고용된 전문인력을 통해 수행하며, 이에 대한 전문 인력을 통한 1차 검수 및 크라우드 소싱 방법을 활용한 2차검수를 진행한 이후 검수가 완료된 데이터를 DB 시스템에 저장하였다. 구축된 데이터의 검증 및 인공지능 학습을 위하여, 저장된 데이터를 학습, 검증, 시험 데이터셋으로 분할하고 이에 대한 인공지능 모델 개발 및 학습을 통해 데이터셋을 검증하였다.

검수와 품질 확보

본 데이터셋의 품질 확보를 위해, 두 가지 검수 시스템을 아래와 같이 운영하였다.

1) 1차 전문인력 검수

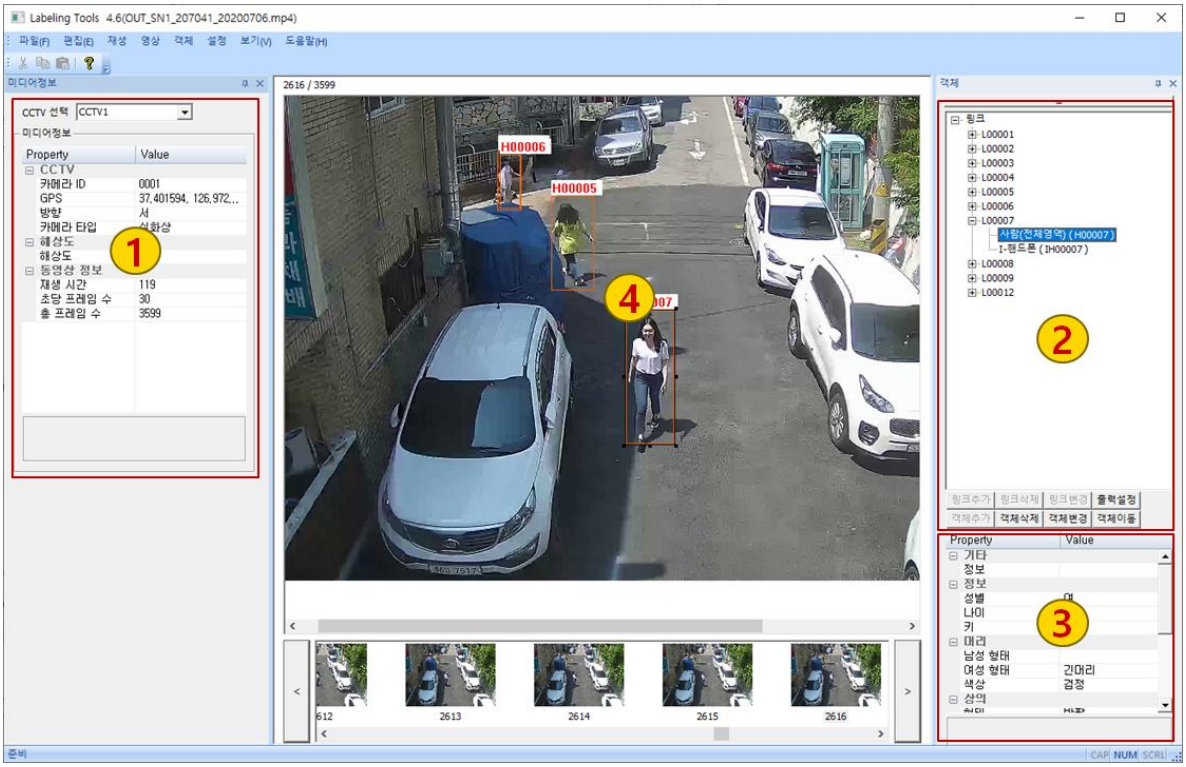

그림 10 1차 전문인력 검수 순서 예시

① 미디어 정보가 맞는지 확인

② 링크 및 객체(보행자)가 모두 있는지 확인

- 사전에 동영상 별로 출연하는 배우들의 정보가 준비되어야 함

③ 추가한 대상 객체의 정보가 맞는지 확인

④ 영상을 플레이하여 Bounding-Box가 대상 객체를 규칙에 맞게 잘 추적하는지 확인

2) 2차 크라우드소싱 검수

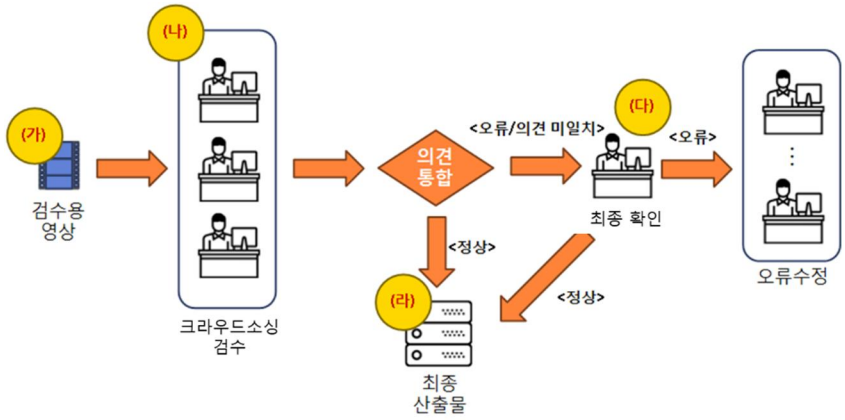

그림 11 2차 크라우드소싱 검수 순서 예시

(가) 검수용 영상 생성

- 어노테이션/라벨링 단계에서 생성된 Crop Image의 개인정보 비식별 처리

(나) 크라우드 소싱을 활용한 2차 검수

(다) 최종확인

- 2차 크라우드소싱 검수에서 지적된 항목을 확인

(라) 검수가 완료된 영상과 라벨링 정보를 최종 산출물로 저장

- 최종 통과한 결과물은 지정된 디렉터리 구조에 맞게 수집
- 수집되는 용량과 접근 편의성을 위해 NAS에 수집
- 손실의 가능성을 배제하기 위해 전용 백업 NAS 사용

데이터 구축 담당자

수행기관(주관) : 한국과학기술연구원, 이메일: jhcho@kist.re.kr