
2019

# Q&A로 풀어본 AI 학습용 데이터 상세 매뉴얼

과학기술정보통신부

NIA 한국정보화진흥원

2019

# Q&A로 풀어본 AI 학습용 데이터 상세 매뉴얼

Contents

1

# 한국어-영어 번역 말뭉치 AI 데이터

구축 기관 (주)솔트룩스파트너스, (주)에버트란, (주)플리토

## 필요성

- 세계는 이미 신경망 번역(NMT, Neural Machine Translation) 엔진을 사용하는 것이 일반화되어, 용도에 맞는 제품과 서비스를 만들어 기존의 번역을 혁신하는 단계가 되었으나, 국내 자동번역/인공지능 번역 연구개발은 대규모 고품질의 말뭉치가 부족하여 개별 기업이 자체적 기술 개발에 한계

- 최근 기업에 신경망 번역 엔진을 공식적으로 도입하여 사용하는 사례가 증가하고 향후 업무 활용 및 연구개발이 더욱 활발해질 것으로 기대함에 따라 공공 번역 말뭉치를 구축하여 공개함

## 구축 내용

- 신경망 번역 성능 향상을 위한 고품질 한국어–영어 번역 말뭉치 셋 160만 문장 구축; 뉴스 80만, 지자체 웹사이트 10만, 조례 10만, 한국문화 10만, 구어체 40만, 대화체 10만

## 구축 공정

## 수집

**Q 수집된 뉴스데이터 종류는 무엇인가요?**

A 과업 결과물에 포함된 디지털 뉴스의 이용 허락 기간은 『2019 한국어-영어 번역 말뭉치 AI데이터 구축』 사업의 결과물로 공개되는 기간까지로 정했습니다. 본 사업이 공개 데이터로 유지되는 기간은 저작권의 이슈가 발생하지 않습니다

**Q 수집된 뉴스데이터 종류는 무엇인가요?**

A 대상매체는 국민일보, 내일신문, 노컷뉴스, 미디어오늘, 서울경제, 스포츠서울, 전자신문, 파이낸셜뉴스, 한겨레, SBS의 10개 매체이며, 발행일이 2018년 6월 1일부터 2019년 5월 31일 사이의 뉴스데이터를 수집하였습니다.

**Q 수집된 뉴스데이터 세부종류는 무엇인가요?**

A 세부분류 항목은 문화, 경제, 정치, 스포츠, IT, 국제, 지역, 사회의 8개 항목입니다.

## 정제

**Q 원문 정제의 기준은 무엇인가요?**

A 160만 문장은 평균 20어절 이상입니다. 그리고 뉴스의 경우 비슷한 기사가 많아서 85% 이상 중복되는 문장은 수집하지 않았습니다.

## 번역

**Q 데이터 학습을 위한 번역 문장의 특성이 있을까요?**

A 문장과 문장이 쌍을 이루어 딥러닝 학습을 효율적으로 할 수 있도록 사업 초기에 구축 가이드라인을 설정하고 작업방식을 통일해서 데이터를 구축했습니다. 문장부호 및 고유명사, 단위나 수치 표시 지침을 정확하게 세우고 준수하도록 했습니다.

## 검수

**Q 검수 시 번역 품질기준은 무엇인가요?**

A 외부기관인 광운대학교 AI번역산업연구센터에서 본 결과물에 대한 품질검사를 진행했습니다. 목표량 검수 후 품질지수가 낮은 해당 작업자의 결과물을 분류하고 구축기관이 해당 작업물을 전수 검사하는 방식을 거쳤습니다. 품질을 평가하는 기준은 정확성과 유창성의 틀에서 세부적으로 평가하였습니다. 하기 표를 참고하세요.

| Error Code | | Issue Types | |
|---|---|---|---|
| 1 | ACm. | Accuracy 정확성 | Mistranslation |
| 2 | ACo. | | Omission |
| 3 | ACme. | | Miscellaneous Error(addition, source text error 포함) |
| 4 | FLg. | Fluency 유창성 | Grammar |
| 5 | FLa. | | Awkwardness |
| 6 | FLme. | | Miscellaneous Error |

## 배포

**Q 데이터의 구조나 메타정보를 알려주세요.**

A 구축 데이터는 엑셀 파일(*.xlsx)로 제공하고 번역 DB를 다운받아 수월하게 활용 가능합니다. 데이터의 모든 문장에는 문장번호를 부착하여 관리가 용이합니다.

| 구분 | 문장번호 | 출처 | 특성 |
|---|---|---|---|
| 뉴스 | ○ | 기사 URL | 기사 작성일 |
| 지자체 웹사이트 | ○ | | |
| 조례 | ○ | URL, 출판물 | |
| 구어체 | ○ | | |
| 대화체 | ○ | | 대화세트/화자구분 |

## Q 한국어-영어 말뭉치 데이터의 가치는 얼마일까요?

A 민간 기업이 본 데이터를 단독으로 구축하기 위해서는 단순 비용만 50억원(전문번역 비용의 1/3 정도라고 생각합니다.) 이상의 자원이 들어갈 것으로 예상됩니다. DB로 공개된 고품질의 번역학습을 위한 말뭉치가 많지 않기에 본 데이터가 가지는 데이터로서의 공공가치는 그 이상입니다.

## Q 데이터를 활용할 수 있는 방법은 무엇인가요?

A 본 말뭉치는 엑셀 파일로 제공하기에, 여러 조건의 메타데이터를 분류하여 사용하고자 하는 엔진에서 학습해 보세요. 기본 번역 학습엔진의 품질을 위해서는 뉴스데이터를 활용하실 수 있고, 특정 분야(조례, 한국문화, 구어체, 대화체 셋)를 증분 학습하는 용도로 활용할 수 있습니다.

### 활용 예시

**기업이나 기관이 보유한 말뭉치와 공개 말뭉치를 활용하여 최적의 맞춤형 인공지능 번역 엔진을 만들고 이를 자기 기관/기업 번역에 활용하거나 사업에 적용하는 등 사업 목적에 맞게 활용**

- 자동 법률번역 시스템 개발
- 한국문화뉴스 자동번역기 개발
- 판매촉진을 위한 CS 챗봇 번역기 개발

**자동번역 기반 크라우드 소싱 사업 활용**

- 번역 기업 활용 : 집단 지성 활용하여 인공지능 번역 성능 향상, 업무 생산성 제고
- 크라우드 번역 사업 : 기존 크라우드 소싱에 블록체인을 도입 수익 배분
- 다국어 스타 SNS : K-POP 스타 팬사이트 등 자발적 다국어 커뮤니케이션이 활발한 서비스에 인공지능 번역 적용, 수익성과 사용자 참여 학습 도모

2

# 한국형 사물 이미지 AI데이터

구축 기관 | (주)미디어그룹사람과숲, (주)인피닉, (주)솔트룩스

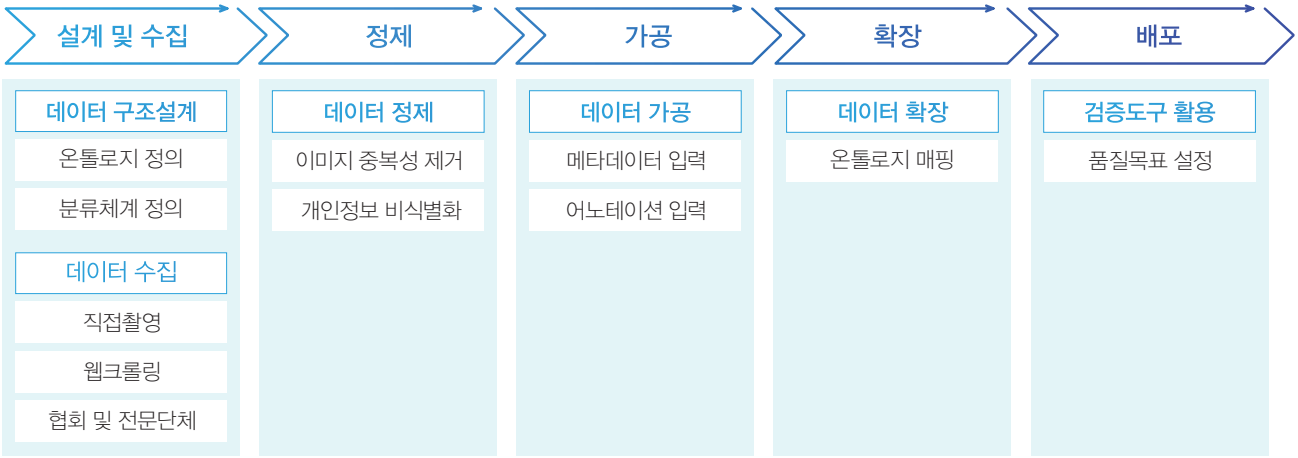
## 필요성

- 인공지능 사물 이미지 인식기술은 자율주행, 스마트시티, 스마트제조, 무인 스토어 등 다양한 산업분야에서 활용 가능한 기술임
- 구글, 이미지넷, COCO dataset 등 대용량 이미지 공개 사이트에는 국내 특성에 맞는 AI 학습데이터가 부족함
- 국내 장소, 객체에 대한 인공지능 기반의 시각지능 기술 개발 및 서비스 강화를 위해 한국형 사물 이미지 AI 데이터 공개 필요

## 구축 내용

- 한국형 사물 이미지 학습용 데이터 구축을 위한 객체 및 속성정보를 취득하여 정보이용자(산업계, 학계 및 연구소)가 연구개발에 쉽고 효율적으로 활용할 수 있는 고품질의 인공지능 학습데이터 구축
- 촬영(수집)데이터 : 한국형 이미지 360만장
- 국가지정 유적건조물(탑, 성곽, 가옥 등) 260만장, 상품(신발, 가방, 지갑, 잡화 등) 80만장, 35개 도시 랜드마크 20만장
- 데이터 촬영 및 이미지 수집을 통한 데이터 구축

## 구축 공정




### 수집

**Q 수집은 어떻게 이루어졌나요?**

A 수집 작업을 진행하다 보면 지식재산권의 문제가 발생하게 됩니다. 따라서 이번 사업의 수집 물량의 90% 이상은 직접 촬영을 통해 조달하고 나머지 10%의 수집 물량에 대해서도 협회와 계약 또는 공공저작물로 사용권이 허락된 데이터에 대해 수집을 진행했습니다.

**Q 직접 촬영은 어떻게 이루어졌나요?**

A 직접 촬영은 유적건조물, 상품, 랜드마크 전문 촬영 경험 인력 또는 사진 전공인력을 투입하여 사진 촬영의 전문성을 확보하였고 4대의 카메라(모니터 카메라 1대)로 다각도의 사진을 확보하였습니다.

### 정제

**Q 개인정보 비식별화는 어떻게 이루어졌나요?**

A 개인정보에 해당되는 얼굴 부분, 자동차 번호판에 흐림 효과(Blur)를 통해 비식별화를 진행합니다. 인식 가능한 부분 영역을 설정하여 얼굴, 번호판에 해당되는 영역을 알아볼 수 없는 수준으로 흐림 효과를 처리합니다.

**Q 유사 이미지 간 중복성 제거는 어떻게 이루어졌나요?**

A 수집을 통해 확보한 이미지를 대상체별로 분류하고 분류된 이미지를 솔루션을 통해 유사도에 따른 유사·중복성 체크된 이미지들을 제거합니다.

## 가공

**Q 가공은 어떻게 이루어지나요?**

A 한국형 사물 이미지 구축 대상에 대한 Bounding Box 처리, 객체별 속성정보를 입력하여 JSON 파일로 처리됩니다. 메타데이터 오류 발생 예방을 위하여 일부 속성값은 자동 입력되는 도구를 개발하였습니다.

**Q 데이터 구조는 어떤 속성으로 구성되어 있나요?**

A 데이터 구조는 속성별로 35개 구조로 구성되어 있으며 크게 3가지로 이미지 데이터 속성정보, 객체 속성정보, 환경 속성정보로 구성되어 있습니다.

## 확장

**Q 의미확장은 어떻게 이루어졌나요?**

A 온톨로지 사전에 해당되는 ADAM KB는 사람, 장소 등 7가지 도메인 영역에서 2,600만개의 인스턴스를 확보하고 있으며 추론 후 트리플 수는 4억 5,000만개로 아시아 최대 규모의 온톨로지 사전을 보유하고 있습니다. 본 사업의 대상체들은 ADAM KB 온톨로지 사전과 매핑되어 데이터의 의미확장 작업이 이루어집니다.

**Q 온톨로지 매핑이 뭔가요?**

A 1차 어노테이션된 레이블 정보를 기반으로 온톨로지를 조회할 수 있는 도구를 활용하거나 온톨로지 질의(SPARQL)를 생성하여 각 레이블에 해당하는 온톨로지 클래스, 프로퍼티, 인스턴스를 조회합니다. 조회된 결과에 대해 수동 검수를 수행하고 해당 레이블과 같은 의미이지만 다른 레이블로 생성된 클래스가 존재할 수 있으므로 온톨로지 상의 상/하위 의미 관계를 수동으로 점검하여 동일한 의미의 개념이 없는지 확인합니다.

## 품질 점검

**Q 품질 점검은 어떻게 이루어졌나요?**

A 결과물은 .Net Framework로 자체 제작된 검수 프로그램을 통해 데이터의 적합성을 검증합니다.
학습데이터 유효성 검증은 RetinaNet 기반의 인공지능 엔진을 활용해서 진행되었으며 Training Data Set, Validation Data Set, Testing Data Set으로 구분하여 진행하며 Test Image Data에 대해서 Ground Truth와 결과를 비교하여 학습데이터의 유효성을 검증하였습니다.

## 활용

**Q** 한국형 사물 이미지 AI데이터에 특화된 데이터 활용법은 어떤 것들이 있을까요?

A 4차 산업혁명 시대의 각 산업별 활용 및 연구 분야에 활용 가능하고 스마트 교육(사업 기간 내 산출물), 스마트 관광, 스마트 공장/스토어 등 다방면에서 활용 가능한 데이터로 판단됩니다.

**활용 예시**

### 한국형 사물 이미지 활용 AI 문서작성 도우미(스마트 교육)

- 문서 작성 시 참고 정보 관련 이미지 및 단어 추천 서비스

### 한국형 사물 이미지 활용 AI 관광 도우미(스마트 관광)

- 문화재 및 도시별 랜드마크가 학습된 AI 관광 도우미를 통해 외국인 관광객(일반인 포함)의 관광 안내 및 청소년 현장 교육에 활용

### 한국형 사물 이미지 활용 AI 스마트 제조(스마트 공장/스토어)

- 상품에 대한 AI 자동인식 기반의 제품 분류/품질 검수 및 자동 솔루션 개발 가능

3

# 한국어 글자체 이미지 AI 데이터

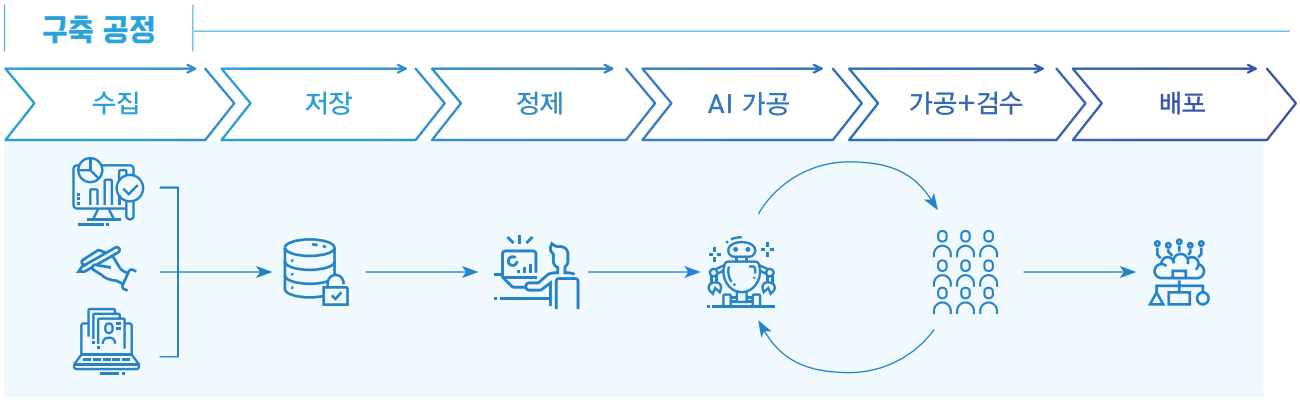
구축 기관 (주)키니앤파트너스, (주)슈퍼브에이아이

## 필요성

- OCR 기술은 자율주행, 증강현실(AR), IoT 등의 산업분야에서 사물의 문자를 인식해서 제공하는 서비스의 기반기술임
- 글로벌 기업(네이버, 구글 등)이 OCR 활용 인지 서비스를 제공하고 있으나, 공개된 한국어 글자체 데이터셋이 없어 기관, 기업의 연구개발에 어려움이 있음
- 전세계적으로 OCR은 AI 기반의 OCR로 변화하고 있으므로 한글 글자체에 대한 공개된 학습용 데이터셋이 필요
- 초성, 중성, 종성의 조합형 형태에 맞춘 데이터셋 구축을 통한 한글 인식 정확도 향상 기반 마련

## 구축 내용

- 현대 한글 글자체 데이터셋 : 현대 한글 문자 1만 1,172자에 대한 인쇄체 및 손글씨체 500만자, Text-in-the-wild 10만장
- 현대 한글, 국립국어원의 한국인이 가장 많이 쓰는 단어 6,000자, 뉴스 기반 문장 등으로 작성한 글자
- 간판, 표지판, 상표, 도서표지 등 Text-in-the-wild는 촬영한 이미지를 사용




## 수집

**Q 인공지능 학습용 데이터의 특성상 실제 일상생활 속에서 발생할 수 있는 다양한 장면의 데이터가 필요한데, 본 한글 데이터는 어떻게 수집되었나요?**

A 기본적으로 한글, 단어, 문장 등 여러 가지 글자 타입을 확보하기 위해 노력하였습니다. 손글씨의 경우, 다양한 성별/나이대의 인원을 약 200명 가량 모집하여 데이터를 구축하였고, 인쇄체의 경우 다양한 폰트를 기본으로 데이터를 구축하였으며, 실제 문서들의 노이즈를 변수요건으로 넣어 증강데이터를 추가 구축하였습니다. Text-in-the-wild의 경우, 일상생활에서 볼 수 있는 한글글자를 최대한 다양한 환경에서 수집하고자 하였습니다.

**Q 손글씨, 인쇄체의 경우, 한글-단어-문장 형태로 수집하였는데, 각각의 문자를 선정한 기준은 무엇이었나요?**

A 다양한 데이터를 확보하는 것을 기본으로 하였고, 그 중 사람들이 가장 많이 사용하는 단어, 문장을 확보하기 위해 노력하였습니다. 많을수록 좋지만, 선택과 집중을 위해서 아래 기준으로 글자를 선정하였습니다. 단어의 경우, 국립국어원에서 공개한 가장 사용 빈도가 높은 '한국어 학습용 어휘 6,000낱말'을 선정하였고, 문장의 경우 뉴스 데이터를 기반으로 만들어진 한국어 말뭉치 자료를 활용하였습니다.(AI 허브 내, AI데이터-기계독해 자료 활용)

## 정제

**Q 데이터 정제는 어떤 절차로, 어떤 기준으로 진행되었나요?**

A 손글씨의 경우, 스캔 후 jpg, jpeg, png 파일로 변환하여 데이터의 유효성을 판단하였으며, 인쇄체의 경우 별도의 소프트웨어를 만들어 인쇄체 글자를 생성하여 디지털 파일로 전환하였습니다. Text-in-the-wild는 촬영 후, 디지털 파일로 출력하였습니다. 수집된 데이터를 디지털화하는 것을 목표로 하였고, 아래 3가지 기준을 충족하도록 정제 기준을 선정하였습니다.

- 흔들림이 없어야 함
- 오탈자가 없어야 함
- 기타 오염이 없어야 함

**Q 데이터의 기본 스펙은 어떻게 되어있습니까?**

A 기본적으로 엔지니어분께서 직접 수정할 수 있도록 고화질 이미지 데이터를 확보하였습니다. 손글씨/인쇄체의 경우, 300dpi이상으로, Text-in-the-wild는 HD급 이상의 이미지를 기본으로 확보하였습니다.

## 가공

**Q 가공은 어떤 기준으로 진행되었습니까?**

A 수집된 이미지는 Crop된 상태로 저장되어 있습니다. 가공은 Crop된 이미지 내의 글자가 무엇인지를 '전사'하도록 하였습니다. 각각의 글자는 아래와 같이 Crop되어 있습니다.

가공 이미지

| 글자 단위의 가공 | 단어 단위의 가공 |
| --- | --- |
|  |  |

**Q 가공된 데이터의 Annotation 정보는 어떻게 기록되어 있습니까?**

A 구축 가이드라인에 Label Structure가 작성되어 있습니다. 수집된 이미지는 Crop된 상태로 저장되어 있습니다. 기본적으로 데이터셋에 대한 정보, 이미지 정보, Annotation 정보, 라이센스 정보가 기록되어 있으니 활용에 참고하세요.

## 검수

**Q 검수는 어떤 절차를 가지고 진행되었습니까?**

A 데이터 구축과정에서 가공 인력/검수 인력이 1차 검증을 하였고, 이후 품질검증 전문 그룹에서 Random Sampling을 하였습니다.

**Q 향후 데이터 문제가 발생할 경우 어떻게 되나요?**

A 공개 사이트에서 데이터를 수정할 수 있도록 Labeling System을 만들어 놓았습니다. 해당 Tool을 사용하여 수정하시면 되고, 혹시 추가 데이터가 구축될 경우, 추가 업로드하여 다른 연구/개발자들이 사용할 수 있도록 공개 부탁드리겠습니다.

## 배포

### Q 파일 용량이 큰데, 어떻게 확인할 수 있나요?

A 함께 공개된 Sample 데이터셋은 본 데이터셋과 동일한 폴더 구조, JSON Format을 갖고 있습니다. Sample 데이터셋과 Label Structure를 활용하여 데이터를 파악하세요.

### Q 활용에 관련된 기본 사항은 무엇인가요?

A 손글씨 및 인쇄체 파일은 Label Structure 이외에는 큰 어려운 부분이 없습니다. 단, Text-in-the-wild의 경우, 작업 기준을 정확히 아셔야 학습에 활용이 가능할 것으로 예상됩니다. Label Structure 하단부의 참고사항을 꼭 확인하세요.

- 허용 문자 : 한글, 영문 알파벳 대소문자, 숫자, 특수기호 7가지 . , ! ? ' " ( )
- 한 이미지에 글자가 너무 많을 때 이를 모두 작업하면 효율이 떨어지므로 한 이미지에서 가장 잘 보이는 20개 이내의 글자를 우선적으로 골라서 작업하고 남아있는 정답 부분은 혼동을 막기 위해 Ignored 박스로 Mask 함. 즉 Ignored는 비정답 영역이 아니라 정답(글자)이 포함될 수도 있는 영역이며, 모든 정답으로 간주되는 글자는 최소한 Character/Word로 어노테이션 되어있거나 Ignored로 가려져있어야 함

### 활용 예시

#### 자율주행 중 글자판독

- AI 산업 분야 중 현재 가장 큰 주요 관심사인 자율주행은 로봇공학과 소프트웨어 측면으로 나누어 볼 수 있으며 소프트웨어 개발 시 이정표 판독, 번호판 판독 등에 활용

#### 전자상거래 제품정보 서비스

- 전자상거래가 점점 활성화되어 경쟁이 치열해지고 있으며, 이에 제품의 정보 콘텐츠를 이미지, 영상 등으로까지 제작하기에 이르렀으나, 소상공인의 경우에는 인력, 비용 등의 문제로 제품의 정보를 제대로 전달하기 어려움
- AI OCR은 오프라인으로 제작된 제품설명서 및 제품의 라벨, 제품 패키지에 붙어있는 이미지 정보를 판독하여 전자상거래 정보로 손쉽게 변환

#### 의료/금융 정보의 개인정보보호

- 의료나 금융은 임상병리, 금융상품 개발을 위한 빅데이터 통계분석이 필요하나 개인정보를 많이 다루고 있어서 어려움이 있으므로 AI OCR을 통한 개인정보식별 및 개인정보보호에 활용

4

# 인도 보행 영상 AI 데이터

구축 기관 | (주)테스트웍스, (사)한국척수장애인협회, (주)셀렉트스타, 카이스트 RCV, (주)디투리소스

## 필요성

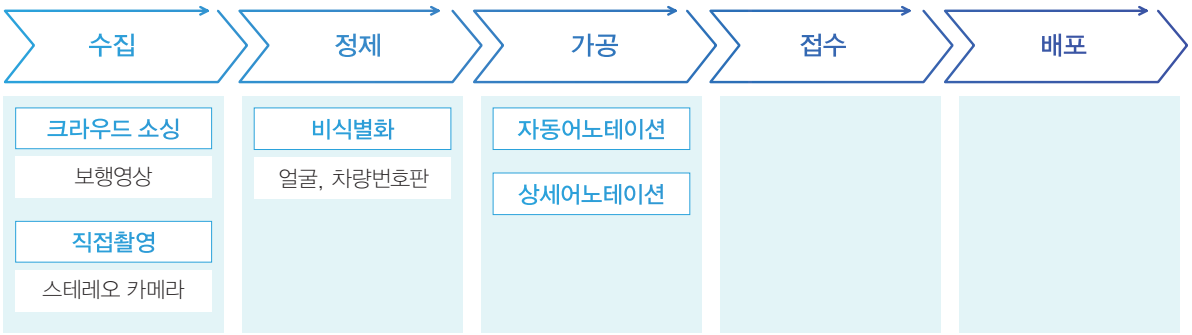
- 보행자용 인도는 차량용 도로와 다른 특성을 가지기 때문에 기존 데이터를 활용하여 보행자를 대상으로 한 인공지능 기술 및 응용서비스 개발에는 어려움이 존재함

- 인도 보행 영상 데이터셋 구축을 통한 사회 문제 해결의 토대 마련 및 AI 기술력 제고
  - 장애인 인도 보행의 어려움과 이동권 문제 해결을 위한 학습 데이터셋 구축
  - 인도 보행 인공지능 공개 데이터셋의 부족

## 구축 내용

| 데이터셋 종류 | 구축수량 | 작업 설명 |
|---|---|---|
| Bounding Box<br>(바운딩 박스) | 35만장 | 인도 보행에 장애가 되는 객체 29종에 대해 '박스' 형태로 어노테이션한 데이터 |
| Polygon Segmentation<br>(폴리곤 세그멘테이션) | 10만장 | 인도 보행에 장애가 되는 객체 29종에 대해 '폴리곤'형태로 어노테이션한 데이터 |
| Surface Masking<br>(서피스 마스킹) | 5만장 | 인도 노면 상태(재질, 특수성, 파손여부 등)정보를 '폴리곤' 형태로<br>어노테이션하고 마스킹한 데이터 |
| Depth Prediction<br>(뎁스 프리딕션) | 15만장 | 스테레오 카메라로 촬영한 인도 보행 이미지를 토대로 구축한 깊이 인식 데이터 |

## 구축 공정




### 수집

**Q 이미지 데이터는 어떻게 수집되었고 해당 이미지의 품질은 어떠한가요?**

A 이미지 데이터는 크라우드 소싱 방식과 직접 촬영 방식으로 수집되었습니다. 크라우드 소싱 방식은 모바일 앱 서비스를 기반으로 휴대폰 앱(1,920×1,080px)으로 촬영합니다. 직접 촬영 방식은 ZED Stereo 카메라(좌/우 이미지 각각 1,920×1,080px)를 사용하여 한국척수장애인협회 회원이 휠체어를 타고 5~10km/h로 이동하면서 1초에 한 번씩 사진을 찍는 방식으로 수집하였습니다. 이렇게 수집된 이미지 데이터 중에서 심하게 흔들리거나 심도효과가 있는 사진, 객체 형태가 구분되지 않는 사진 등은 데이터 품질을 위해 폐기하였으나 약간 흐린 이미지, 30도 미만으로 기울어진 이미지 등은 일부 존재합니다.

**Q 이미지 수집 시에 수집 조건은 없었나요?**

A 보행 가능한 구역에서 장소, 날씨, 시간 등 다양한 촬영 조건으로 이미지를 수집하여 AI 모델 학습을 위한 시간적, 공간적 다양성을 최대한 확보하도록 하였습니다. 수집 가공된 바운딩 박스 35만장과 폴리곤 세그멘테이션 10만장 이미지에 대해 낮/밤, 비/비오지 않음, 촬영 시각(일부 이미지 한함) 등의 환경정보를 .csv 파일로 제공합니다.

### 정제

**Q 촬영된 이미지는 원본 그대로 AI 데이터 구축에 사용되었나요?**

A 개인정보를 보호하기 위해 비식별화 과정을 거쳤습니다. 비식별화 과정은 1) 딥러닝 기반의 얼굴 및 차량 번호판 검출 알고리즘을 개발하여 수집된 이미지 데이터 내의 얼굴 및 차량 번호판을 자동으로 추출합니다. 2) 자동 추출된 얼굴과 차량 번호판에 대해 오검출, 미검출, 과검출 등을 수동으로 보완합니다. 3) 최종적으로 검출된 얼굴 및 번호판에 대해 OpenCV Library를 활용해 흐림처리하여 비식별화를 완료합니다.

## 가공

**Q 객체 클래스는 어떠한 기준으로 선정되었나요?**

A 보행 간 마주하게 되는 보행 장애요소들을 이동체와 고정체로 분류하여 정의했습니다.

| 대분류 | 소분류 | 레이블명 |
|---|---|---|
| 장애요소 | 이동체 | person(사람), car(승용차, 승합차), bus(버스), truck(트럭), bicycle(자전거), motorcycle(오토바이, 스쿠터), scooter(변속 기어가 없는 이륜차, 두 발로 탈 수 있는 기구), stroller(유모차), wheelchair(휠체어), dog(개), cat(고양이), movable_signage(이동식 홍보 전시물/안내판), carrier(리어카, 손수레) |
| | 고정체 | tree_trunk(가로수 기둥), potted_plant(화분), traffic_light(신호등), traffic_sign(교통 표지판), pole(대/기둥), bench(2인 이상이 앉는 기구), chair(1인이 앉을 수 있는 기구), table(탁자), stop(버스/택시 정류장), kiosk(한쪽이 열린 점포), fire_hydrant(소화전), parking_meter(주차요금정산기), bollard(볼라드), barricade(바리케이드), power_controller(전력제어함), traffic_light_controller(신호등제어기) |

노면의 안전성 부족으로 인한 위험유발가능성을 고려하여 노면 특성(재질, 특수성, 파손여부 등)에 따라 노면 객체 및 속성을 정의하였습니다.

| 대분류 | 레이블명 | 속성 |
|---|---|---|
| 노면 | alley(사람과 차가 함께 다닐 수 있는 길) | crosswalk(횡단보도)<br>damaged(파손)<br>normal(일반)<br>speed_dump(과속방지턱) |
| | bike_lane(자전거도로) | (속성값 없음) |
| | braille_guide_blocks(점자블록) | damaged(파손)<br>normal(일반) |
| | caution_zone(주의구역) | grating(그레이팅)<br>manhole(맨홀)<br>repair_zone(보수구역)<br>stairs(계단)<br>tree_zone(가로수영역) |
| | roadway(차만 다닐 수 있는 길) | crosswalk(횡단보도)<br>normal(일반) |
| | sidewalk(인도) | blocks(보도블럭)<br>cement(시멘트)<br>damaged(파손)<br>other(기타)<br>soil_stone(흙/돌)<br>urethane(우레탄) |

**Q 이미지에 나와 있는 모든 객체를 어노테이션하나요? 아니라면 어떤 기준으로 어노테이션했나요?**

A 데이터셋에 따라 각각의 가이드라인을 참고하여 어노테이션합니다

## 바운딩 박스 가이드라인

- 사진을 확대하지 않은 원본 상태에서, 레이블 목록에 포함되며 뚜렷하게 구분되는 가장 원거리 객체를 기준으로, 이 기준보다 먼 원거리의 (뒷부분의) 객체는 어노테이션하지 않음


- 단, 아래의 경우는 어노테이션하지 않음
  - 객체의 식별이 불가능한 경우(너무 흐리거나 차나 버스 같은 객체가 너무 빨리 이동해서 위치 파악이 어려운 경우 등)
  - 객체의 가로, 세로의 너비 중 긴 면이 64px 이하인 경우(단, Car, Bus, Truck, Person, Bollard, Traffic_light, Traffic_sign의 경우 32px 이하)
  - 거울이나 유리창에 반사된 객체
  - 광고물에 있어 인쇄된 객체
- 객체가 너무 밀집되어 있어 정확하게 구별하기 힘든 경우, 집단에서 가장 선명하고 외곽선이 뚜렷한 객체 2~3개를 어노테이션하며, 나머지는 하지 않음
- 객체끼리 겹치거나 화각으로 인해 객체가 잘려서 촬영된 경우에는 객체의 온전한 모양을 추정하여 사각형을 그려줌

## 폴리곤 세그멘테이션 가이드라인

- 기본적으로 바운딩 박스와 동일한 가이드라인을 적용함
- 단, 장애물에 의해 객체가 나눠져 가림 현상이 발생한 경우, 조각난 객체를 각각 그린 후 하나의 그룹으로 지정하여 하나의 객체임을 나타냄
- 객체 외곽선을 따라 어노테이션할 때 1 pixel 내외의 얇은 나뭇가지나 전선 등은 무시

## 서피스 마스킹 가이드라인

- 서피스 마스킹 레이블 목록에 정의되어 있는 노면은 모두 어노테이션함을 원칙으로 함
- 모든 객체는 겹치지 않도록 어노테이션해야 함
- 겹쳐진 상황으로 어노테이션해야 하는 아래의 경우에는, 큰 객체를 먼저 어노테이션한 후 그 위에 아래 객체를 어노테이션함
  - Caution_zone – manhole
  - 폭이 30 pixel 이하인 작은 Braille_guide_block
  - Sidewalk – damaged, Alley – damaged, Braille_guide_block – damaged

## 검수

**Q** 데이터셋의 어노테이션 정확도를 높이기 위해서 어떤 과정을 거쳤나요?

A 3단계의 검수 절차(Peer Review – Manager Review – Final Audit)를 거쳐 오검출, 미검출, 과검출, 어노테이션 범위 불량인 객체가 없는지 확인하여 데이터 품질을 관리하였습니다.

## 배포

**Q** 공개되는 각 데이터셋은 어떻게 구성되어 있나요?

A

| 데이터셋 종류 | 전체 이미지 수량 | xml 수량 | 작업(폴더) 수량 | 폴더 당 구성 |
|---|---|---|---|---|
| 바운딩 박스 | 352,901장 | 2,520개 | 2,520개 | Bbox_xxxx\image(.jpg or .png) 50~500장 + xml 1개 |
| 폴리곤 세그멘테이션 | 100,712장 | 1,994개 | 1,994개 | Polygon_xxxx\image(.jpg or .png) 50~300장 + xml 1개 |
| 서피스 마스킹 | 51,318장 (마스킹이미지 51,318장 별도) | 900개 | 900개 | Surface_xxx\image(.jpg) 31~120장 + xml 1개 + 'MASK'하위폴더 \mask image(.png) 31~120장 |
| 뎁스 프리딕션 | 174,080 set (1세트가 8장의 이미지로 구성됨) | 해당사항 없음 | 93개 | Depth_xxx\image(.jpg or .png) 8×n장 + calibration(.conf) 1개 |

**Q** 바운딩 박스/폴리곤 세그멘테이션/서피스마스킹 데이터셋에서 제공되는 어노테이션 정보를 담고 있는 XML Data Format은 어떻게 구성되어 있나요?

A 다음의 xml 예시를 통해 설명합니다.

xml 파일 예시

**공통정보 〈meta〉**

| Tag | | 설명 |
|---|---|---|
| task | id | task의 고유번호. 서버에서 관리하는 값으로 자동으로 부여 |
| | name | task의 이름 |
| | size | task의 사이즈 |
| | mode | task의 작업 모드(default : annotation)<br>annotation은 이미지, interpolation은 동영상 작업 |
| | overlap | task가 동영상인 경우 task간 겹쳐진 프레임의 개수 |
| | bugtracker | 이슈 관리 |
| | flipped | 뒤집힘 여부 |
| | created | task가 생성된 시간 |
| | updated | task가 수정된 시간 |
| | start_frame | task의 시작 프레임 정보(default : 0) |
| | stop_frame | task의 종료 프레임 정보(default : 0) |
| | frame_filter | 프레임 필터 정보(default : 0) |
| | labels | 라벨 리스트(라벨은 label name, attribute로 구성) |
| | segements | task를 등록했을 경우 segments 관련 정보 |
| | owner | annotation 작업자 관련 정보 |
| dumped | | xml 다운로드 받은 시간 |

**이미지별 어노테이션 정보 〈Image〉**

| Tag | Attribute | 설명 |
|---|---|---|
| image | id | image 고유번호 |
| | name | image 이름 |
| | width | image 너비 |
| | height | image 높이 |
| box | label | 레이블명 |
| | occluded | 가림/잘림 표시 |
| | xtl/ytl | box에서 좌상단 x/y 좌표(X Top Left/Y Top Left) |
| | xbr/ybr | box에서 우하단 x/y 좌표(X Bottom Right/Y Bottom Right) |
| | z_order | 객체가 그려진 순서 정보 |
| polygon | label | 레이블명 |
| | points | polygon에서 폴리곤의 포인트 좌표 |
| | occluded | 가림/잘림 표시 |
| | z_order | 객체가 그려진 순서 정보 |
| | group_id | polygon에서 그룹화 작업의 정보<br>group_id 값이 같으면 같은 그룹 |
| attribute | name | 속성값 |

## Q 깊이 추정 데이터셋의 구성은 어떻게 되어 있나요?

A Depth_xxx 폴더당 8×N 개의 파일과 calibration 파일 하나를 담고 있습니다. 하나의 뎁스 데이터셋에 해당하는 8개의 파일은 아래와 같습니다.

| 파일명 | 내용 |
|---|---|
| Disparity16 | GA-net을 통해 추출되는, disparity의 data(좌우 이미지에서 대응점의 x축 상 pixel 차이, 거리감을 나타내는 data)가 저장된 파일. grayscale로 저장 |
| Disparity | GA-net을 통해 추출되는, disparity를 시각화(visualization)한 파일. RGB로 저장 |
| Confidence_save | photo-consistency 방식을 통해 GA-net으로 추출된 disparity의 신뢰성을 측정하기 위해 생성된 이미지. 각 픽셀에 대해 좌우 이미지의 밝기 차이(0~255) 값이 들어있음 |
| Confidence | confidence_save의 값을 분석하여 2진화한 파일.<br>본 과제에서는 전체 pixel 개수 중 1의 개수의 합이 20% 이상인 것만 취함<br>* 1 : true confidence (백색)<br>* 0 : false confidence(흑색) |
| Crop_Left | GA-net의 입력 이미지로 쓰이는 Stereo Camara의 왼쪽 이미지. 1,920×592 |
| Crop_Right | GA-net의 입력 이미지로 쓰이는 Stereo Camera의 오른쪽 이미지. 1,920×592 |
| Raw_Left | 원본의 왼쪽 이미지. 1,920×1,080 |
| Raw_Right | 원본의 오른쪽 이미지. 1,920×1,080 |

*calibration 파일 : 좌우 카메라의 상대적 위치, 광학 특성 등 스테레오 영상에서 depth를 구하기 위해 좌우 영상 보정 정보를 담고 있는 파일입니다.

## 활용 예시

### ≡ 안전한 도로(Safety Road) 서비스

- 보행자도로의 균열, 파손, 점자블럭의 파손 등 노면 인식 기술을 활용하여 사회적 배려 대상들이 안전하게 다닐 수 있는 안전한 길 안내 서비스를 구축함

### ≡ 한국형 딜리버리 봇 서비스

- 인도 상 객체/거리 인식 기술을 이용하여 보행 시 마주치게 되는 사람, 사물 등과 충돌하지 않고 안전하게 원하는 위치로 정상적으로 제품을 배달하는 서비스를 구축함

### ≡ 자율주행차 서비스

- 생활도로(이면도로) 위 객체/거리 인식 기술을 이용하여 자동차의 자율주행 단계 Level 3(조건적 자율 주행)로 발전시킴으로써 자율주행차 서비스를 구축함

## 부록: 도구 산출물

### ≡ 얼굴/번호판 비식별화 모델

- 원천 데이터 수집 및 가공 과정 중 개인의 초상권 등 개인정보 수집이 예상되는 바, 비식별화 기술을 활용한 권리 보호 보완 목적으로 활용하는 모델. 딥러닝 기반 학습데이터 Annotation 및 Validation 기술을 보유한 셀렉트스타(주)에서 개발

- 공개 및 사용법

- 공개 github 주소 : https://github.com/selectstarofficial/AI-Detection

5

# 멀티모달 영상 AI 데이터

구축 기관 (주)아크릴, (주)테스트웍스

## 필요성

- 기존 공개된 멀티모달 영상 데이터는 인공지능 연구 및 서비스를 개발하기에는 데이터 규모가 크지 않아 대규모 학습 데이터의 제작 필요성이 논의되어 왔으며, 저작권 및 초상권 문제, 연구개발 목적으로의 사용 제약으로 데이터 활용에 제한이 있음

- 한국의 문화에 맞는 인공지능 서비스 개발 및 적용을 위해서는 국내 상황에 맞는 인물, 대화 내용(맥락), 상황 정보 등이 포함된 대규모의 데이터가 필요함

## 구축 내용

- 감정, 성별, 연령대, 발화 스크립트, 개체관계, 발화 대화 의도, 대화 전략, 상황 설명 정보의 의미 정보가 부착된 멀티모달 영상 AI 데이터 110시간 분량 구축

- 멀티모달 영상 AI데이터 구축을 위한 저작도구 개발 및 배포

- 8종 AI 임무 유형의 데이터 검증용 인공지능 모델 개발 및 GUI 기반의 데이터 학습·테스트·배포가 가능한 시범서비스 개발
  *8종 AI 임무 유형 : 감정인식, 개체인식, 성별/나이 인식, 관계분석, 멀티모달 영상 질의응답, 단일 발화 의도 분석, 복수 발화 의도 분석, 음성인식

## 구축 공정

수집 → 정제 → 가공 → 검수 → 배포

## 수집

**Q** 수집된 데이터의 촬영 규모와 투입 인력, 촬영 데이터의 형태는 어떻게 되나요?

A 데이터의 총 규모는 중복되지 않는 6,000개의 시나리오(대화문 85,077개)를 기반으로 한 110시간 분량의 일상 연기 영상 데이터로 구성되어 있으며 총 300명의 연기자가 투입되었습니다.

**Q** 수집된 데이터의 시나리오에서 어떤 정보가 인공지능 분야에 활용될 수 있나요?

A 수집된 시나리오는 6,000개의 중복되지 않은 일상 상황에 대한 대본이며 해당 시나리오들은 각각 상황, 관심사에 대한 요약을 포함하고 있습니다. 이런 정보들은 전체적인 문맥으로 상황 정보 등을 유추하는 인공지능 분야에 적용 가능합니다.

**Q** 데이터 수집을 위한 촬영 시에 수집환경에 따른 편차가 존재하고 햇빛 또는 소음과 같은 노이즈가 발생할 수 있는데 해당 부분은 어떻게 해결했나요?

A 데이터 수집은 스튜디오 등 실내 작업 공간에서 수집하는 것을 원칙으로 하여 햇빛 또는 차량 소음과 같은 노이즈를 해결하였고, 여러 실내 환경에 대한 상황을 가정하여 촬영하였습니다.

## 정제

**Q** 수집된 데이터의 품질관리를 위한 방안은 어떤 것들이 있을까요?

A 수집된 촬영 데이터들은 촬영 현장에서 전문 촬영 프로듀서가 발화와 대본의 일치성 여부 검수, 연기 적절성 확인, 영상 상태 확인 과정을 거치게 되고, 전문 프로듀서의 결정에 따라 편집 또는 재촬영을 진행하여 데이터 품질관리를 수행하였습니다.

## 가공

**Q** 가공 데이터의 규모와 데이터는 어떤 형식으로 제공되나요?

A 가공 데이터는 수집된 영상 데이터를 기준으로 어노테이션을 수행합니다. 이를 통하여 JSON 파일과 발화를 기준으로 편집된 110시간 분량의 영상 데이터가 제공됩니다.

**Q** 수집된 영상 데이터의 가공을 통하여 얻을 수 있는 데이터의 항목은 어떻게 되나요?

A 영상 데이터의 가공을 통하여 얻을 수 있는 데이터의 항목은 인물별 감정, 인물 정보, 인물별 발화 스크립트, 개체 정보, 관계 정보, 상황 설명 정보, 발화별 대화 의도, 발화 대화 전략 분류 정보가 존재합니다.

**Q** 데이터의 가공은 어떤 방식으로 수행되었나요?

A 영상 데이터의 가공은 수행 목적에 맞게 커스터마이징한 오픈소스 기반 저작도구와 다양한 프로세스(발화 시점, 텍스트 감정, 발화 의도 등)로 업무 분할을 통하여 수행되었으며, 추후의 데이터 가공은 자체 개발한 저작도구를 통하여 단일 도구만으로도 데이터의 가공이 가능합니다.

## 검수

**Q** 고품질의 데이터를 만들기 위해 어떻게 진행하였나요?

A 어노테이션을 진행한 데이터는 3자 검수를 통하여 어노테이션 데이터의 이상 여부를 판단하는 것을 원칙으로 하였습니다. 데이터 검수 과정 중에 어노테이션 데이터의 이상 여부가 확인되면 해당 데이터는 재태깅 대상으로 지정되고 다시 한번 검수 과정을 거치게 됩니다.

**Q** 멀티모달 영상 AI 데이터의 경우 데이터의 복잡도가 높아 검수 과정이 어려워 보이는데 이런 부분을 해결하기 위해 어떤 검수 과정이 진행되었나요?

A 한 사람이 한 번에 모든 데이터가 적합한지 검수하기에는 복잡도가 높아 매우 까다롭습니다. 그래서 검수 이미지 제작, Bounding Box & 인물 ID 검수, 발화 시간 검수, 발화 내용 검수, 사물 라벨 태깅 및 관계 태깅 검수, 감정 태깅 검수로 총 6개 단계로 나누어 단계별로 검수를 진행하였습니다.

## 배포

**Q** 데이터의 복잡도가 높은데 사람들이 사용하기에 문제점이 없나요?

A 직관적인 데이터 확인이 가능하도록 JSON 형식을 택하였고 데이터 내에 key 이름 또한 최대한 직관적인 명칭을 사용하여 이해하기 쉽게 작성하였습니다. 제공하는 데이터는 영상 기반 데이터이므로 프레임마다 데이터의 형태가 변하게 됩니다.
이러한 부분을 고려하여 사용자가 필요로 하는 데이터에 쉽게 접근할 수 있도록 key 이름을 기반으로 하여 원하는 데이터에 접근할 수 있게 설계하였습니다.

**활용 예시**

### ≡ 외부인 침입 감지(개체인식, 인물인식)

- 사용자가 외출 시 시스템에 등록되지 않은 인물이 감지될 경우, 외부인 침입으로 간주하여 사용자에게 알리고 경찰에 신고

### ≡ 대화를 통한 감정분석(발화 의도 분석, 음성인식, 감정인식)

- 사용자와 일상 대화를 통해 얻어지는 영상과 음성을 분석하여 현재 사용자의 감정 상태를 파악

6

# 사람 동작 영상 AI데이터

구축 기관 | (주)스위트케이, 서울대학교, (주)케이티, (주)모션테크놀로지

## 필요성

- 사람의 자세 추정 연구는 컴퓨터 비전(CV, Computer Vision) 분야에서 오래전부터 연구되어온 분야이며 관절 데이터는 해당 연구에서 가장 필수적이며 기초적인 데이터
- 국내 기초 연구 및 여러 산업 분야에서 활용 가능한 내국인 및 국내 환경에 맞는 범용적인 관절 데이터가 부족한 상황

## 구축 내용

- 기초 동작 및 의미론적 동작에 대한 총 20만 클립 제작
  - 기초 동작 : 사람의 동작 가운데 가장 기초적인 동작인 쓰러짐, 걷기, 뛰기, 앉기, 놀기 등 20종
  - 의미론적 동작 : 일상생활, 홈피트니스, 운동, 악기 연주 등 30종의 동작

## 구축 공정

| 획득 | 가공 | 검증 |
|---|---|---|
| **데이터 적재** | **작업자 관리** | **데이터 검증** |
| 영상 촬영 | 사용자 등록 | 데이터 탐색 |
| 영상 업로드 | 검수자–작업자 배정 | 시각화 검증 |
| 영상 가공 | 작업자–영상 배정 | 전수 검수 |
| **데이터 전처리** | **데이터 전처리** | **데이터 분할** |
| 영상 메타 생성 | 어노테이션 | 학습 데이터셋 생성 |
| 이미지 추출 및 메타 생성 | 1차 검수 | 검증 데이터셋 생성 |
| Born detection | 2차 검수(샘플링) | 시험 데이터셋 생성 |

## 데이터 획득

**Q 다양한 동작 이미지들로 구성되어 있는데, 데이터는 어떻게 구성이 되어 있나요?**

A 데이터는 기초 및 의미론적 동작 50여종으로 되어 있습니다. 기초동작에는 동작 가운데 가장 기초적인 동작인 걷기, 뛰기, 앉기, 인사, 포옹, 쓰러짐, 놀기, 체조 등 20종으로 되어 있고 의미론적 동작은 탁구, 야구, 축구, 배드민턴 등 각종 운동 및 댄스, 악기 연주, 물건운반 등의 30종으로 되어 있습니다. 50종의 동작은 350여가지의 상세 시나리오로 구성되어 있습니다.

**Q 기초동작과 의미론적 동작의 액션은 어떻게 선별하였나요?**

A 기존에 오픈되어 있는 Human3.6과 MPII를 참고하여 걷기, 달리기, 앉기와 같은 기초동작을 주로 선별하였으며, 참여기관인 서울대와 KT의 향후 적용될 서비스 관점에서 의미론적 동작을 선별하였습니다. 무엇보다 2019 AI Expo를 통하여 담배피기, 푸시업, 윗몸일으키기, 포복, 태권도 등 많은 의견을 받아 전부는 아니지만 빈도가 높았던 항목에 대해 반영을 하였습니다

**Q 동작별로 클립 수가 다양한데 이렇게 다른 이유가 있나요?**

A 주로 동작이 단순하고 짧게 끝나는 경우 클립 수가 1,000~3,000 클립이 되며, 댄스나 요가처럼 길고 다양한 동작의 경우 1만클립 전후로 선정하였습니다. 상세하게는 각각의 세부 시나리오가 있어 세부 시나리오의 다양성에 따라 동작도 다양하여 수량이 달라지게 되었습니다. 참고로 1,000클립은 이미지로는 2만장 이상이 되는 적지 않은 수량입니다.

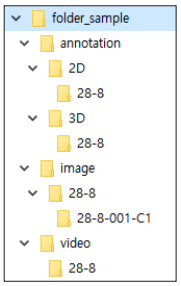
**Q 촬영 환경과 관련해서 촬영장소, 액터, 카메라 등은 어떻게 구성하였는지요?**

A 촬영장소는 주로 2곳의 스튜디오에서 촬영을 하였으며, 조금 더 다양한 장면을 연출하기 위해 체육관, 옥상, 공터, 큰 창고 등에서 촬영을 하였습니다. 액터는 남성 8분, 여성 11분 총 19분이 참여를 했으며, 주로 20대 분들로 구성을 했습니다. 기존에 데이터셋이 단면에 의한 이미지였다면 본 데이터셋은 기존에 단면 촬영에 의해 가려졌던 부분이 없도록 12대의 카메라를 상/중/하, 앞/뒤/좌/우에 배치하여 촬영하였습니다. 특히 12대의 카메라를 시간 동기화하여 동일한 동작에 대해 다양한 각도에서의 영상을 확보하였습니다.

## 데이터 가공

### Q 데이터 구조는 어떻게 되나요?

A 오른쪽 예시와 같이 Annotation 아래에 JSON기반의 2D와 3D 데이터 파일이 있으며, 2D, 3D, image, video 아래에는 모두 동일하게 액션시나리오 번호(ex, 28-8) 폴더가 있습니다. 3D의 경우 전체 수량의 10% 정도만을 시범으로 구축하여 모든 폴더가 있지는 않습니다. image의 경우에만 촬영 회차 및 카메라 위치별 정보를 담고 있는 폴더를 두어 그 아래에 동일 영상에 대한 추출 이미지를 두었습니다.(ex, 28-8-001-C1) 여기서 001은 28-8번 시나리오를 반복 촬영한 번호가 됩니다.



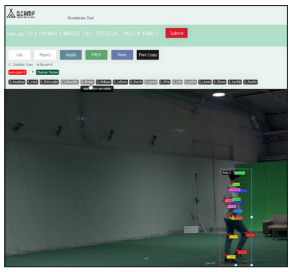

### Q 데이터를 가공하는 데 별도의 툴이 있나요?

A 


주관기관 ㈜스위트케이에서 보유하고 있는 인공지능학습데이터 가공 시스템인 D.CAMP가 있으며, 해당 시스템 안에는 이미지, 동영상, 3D, Text, MRC 등 다양한 유형의 인공지능 학습 데이터를 만드는 기능을 포함하고 있습니다.
이중 사람 동작 영상에 대한 가공기능 일부에 대하여 별도로 시스템을 구성하여 AI 허브(www.aihub.or.kr)에 오픈하였습니다. 가공툴은 Java 기반으로 구동되며, 로컬에 다운받아 경로 수정 후 실행을 하면 바로 웹상에서 사용 가능하도록 구성하였습니다.

### Q 데이터 가공은 어떤 식으로 하는지요?

A 촬영된 영상에서 1초에 3장씩 이미지를 추출하며, 추출된 이미지에 대하여 body detection 및 keypoint detection이 이루어집니다. 추출된 이미지와 2D 관절정보는 아래와 같이 웹상에 표시되며, 제대로 추출이 되지 않은 관절에 대하여 사람이 마우스로 수정을 합니다. 상단의 회색과 검은색은 해당 관절이 보이고(검은색) 안보이고(회색)를 의미하며 이 정보는 나중에 학습 시 보이는 관절만을 대상으로 할 수도 있습니다.



## 데이터 검증

### Q 데이터의 정확성은 어떻게 검증하였나요?

A 데이터가 일관적이지 않으면 학습이 제대로 이루어지지 않는 것에 착안하여 구축된 데이터를 이용하여 모델을 학습시키고 손실함수 값이 epoch에 따라 잘 떨어지는지 확인하였습니다. 무엇보다 20만 클립의 많은 데이터를 만들기 위해서는 1차적으로 keypoint detection에 의해 자동으로 추출되는 관절정보가 잘 나와야 고품질의 데이터를 더 빠르게 확보할 수 있어 만들어지는 데이터로 학습시켜 추출작업의 성능을 높이는 데 활용하였습니다.

**Q 검증을 위한 데이터는 어떤 기준으로 추출했나요?**

A 20만 클립의 많은 데이터를 모두 사용하지는 않았으며, 400만장 이상의 많은 이미지 중 랜덤 샘플링을 통해 20만장을 추출하여 검증하였습니다.

**Q 관절 좌표값이 사물이나 다른 관절에 의해 가려진 값은 어떻게 구분하였나요?**

A 실제 환경에서는 다른 사람이나 사물 등에 가려 관절이 보이지 않는 경우가 많으며 이러한 상황을 반영하기 위해서 Visualize flag를 두어 0은 관절이 사진상에 존재하지 않을 때, 1은 관절이 사진상에 존재하나 가려졌을 때 그리고 2는 사진상에 존재하고 가려지지 않고 잘 나올 때를 기준으로 구분하였으며 이를 통해 산업에서 필요한 서비스에 따라 flag 값을 통해 데이터를 구분하여 활용할 수 있도록 하였습니다.

**Q 영상 데이터의 다양성을 어떻게 확보하였나요?**

A 첫째로 실내촬영과 실외촬영을 통해 여러 촬영환경을 조성하였으며 특히 실외촬영은 체육관, 옥상 등 다양한 촬영환경에서 데이터를 수집하였습니다.
둘째로 액터의 의상과 조도를 다양하게 하여 다양한 영상 데이터를 수집하였으며 실제 서비스적 관점에서 영상의 시나리오를 다양하게 하여 여러 상황에서 데이터가 활용될 수 있도록 구축하였습니다.

**활용 예시**

**홈트레이닝**

- 카메라를 통하여 운동하는 자세를 촬영하고 실제 트레이너와 얼마나 유사하게 하고 있는지 점수화하여 매일 달라지고 있는 운동자세를 리포트해주는 서비스

**댄스동작 따라하기**

- 노래를 선곡하여 댄스동작을 따라하고 따라하는 나의 모습을 촬영하여 댄스동작 안무와 어느 정도 정확도를 보이고 있는지 수치화하여 자세 정확도를 리포트해주는 서비스

**무인 재활운동 지원**

- 증강현실 기반으로 나의 모습을 보여주고 팔, 다리 등 운동을 할 수 있는 미션을 증강현실 공간상에 투입하여 해당 물체를 터트리면서 재활운동의 재미를 부여해 주는 서비스

7

# 한국인 안면 이미지 AI 데이터

구축 기관 | 한국과학기술연구원, (주)휴먼ICT, (주)SQI소프트

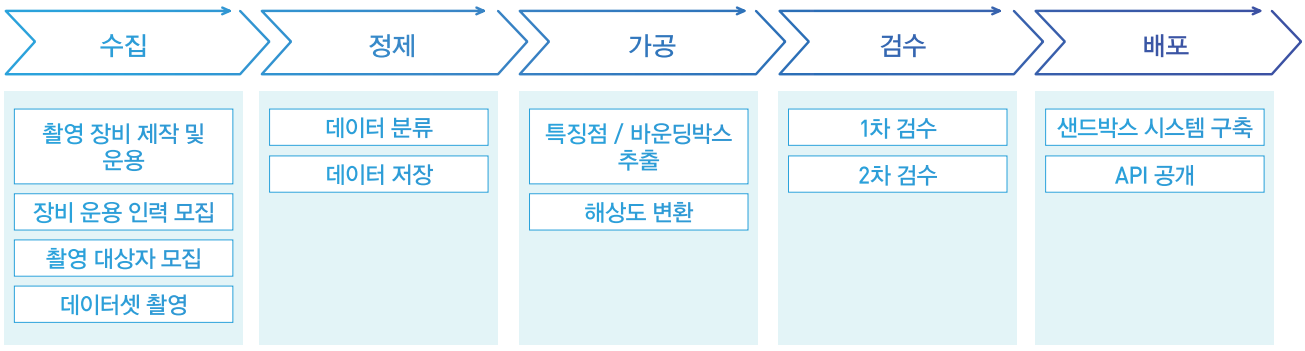
## 필요성

- 현재까지 대규모 얼굴 영상 데이터베이스는 국외 연구기관에서 주도적으로 구축한 사례들이 존재
- 대부분 서양인 위주의 얼굴 영상을 포함하고 있어서 한국인 특성에 맞는 신원 확인 기술 및 얼굴 관련 응용 기술에 대한 성능 최적화 등에 어려움 존재

## 구축 내용

- 얼굴 데이터의 실효성 제고를 위해 각도 20종, 조도 30종, 가림 6종, 표정 3종, 해상도 3종을 반영한 600명(인당 약 3만장)의 안면 이미지 데이터 제작

## 구축 공정



## 수집

**Q 각도와 조도가 다양한 이미지들을 구축하였는데, 그 이유가 무엇인가요?**

A CCTV와 같은 비제약환경에서의 신원확인 기술 개발을 위해서는 다양한 환경에 대한 데이터베이스가 필요합니다. 특히, CCTV의 경우에는 다양한 날씨, 시간, 각도에 따라 인물이 촬영되므로, 기존의 기술로는 신원 확인에 한계가 있다고 할 수 있습니다.
본 데이터베이스에는 다양한 조도, 각도가 포함되어 있으므로, 이를 학습에 활용하면, 비제약환경에서의 신원 확인 기술 성능을 높일 수 있을 것으로 예상할 수 있습니다.

**Q 데이터베이스 수집을 위해서는 많은 사람이 동원되어야 할 것 같은데, 장비 구축과 촬영은 어떤 식으로 진행되었습니까?**

A 데이터베이스 수집을 위해서 본 사업에서는 총 4대의 장비를 구축하였고, 동시에 운용하였습니다. 장비당 촬영 인원은 총 8명으로 각각 촬영관리자, 촬영 요원, 1차 검수 및 특징점 보정 요원, 2차 검수 및 특징점 보정 요원의 역할을 수행하였습니다. 많은 데이터를 촬영하므로 촬영과 동시에 검수가 진행될 수 있도록 하여 효율적인 데이터셋 구축을 가능하게 하였습니다.

**Q 다양한 연령대의 촬영자를 모집하기가 어려웠을 것 같은데, 어떤 식으로 촬영자를 모집하였습니까?**

A 촬영 대상자는 성별 및 연령대의 균형을 맞춰서 섭외하였으며, 인력채용 사이트, 엑스트라 협회, 지자체 협조 등을 통한 다양한 경로를 통해서 모집하였습니다.

## 정제

**Q 대규모의 데이터베이스이므로, 정제 작업에 많은 인력이 동원되었을 것 같은데, 데이터 정제는 어떤 식으로 진행되었습니까?**

A 본 사업에서는 1인당 3만 2,400장의 영상을 촬영하도록 하며, 집단 지성을 활용하여 촬영 대상자 및 영상 정보 분류를 진행하였습니다. 수작업 정제 및 파일 오류 자동 검출 도구를 사용하여 사용불가 사진 제거 및 미촬영 사진에 대한 즉각 추가 촬영을 진행함으로써 효율성을 높였습니다.
또한, 촬영된 영상을 자동으로 K-Face 구조에 맞추어 구성된 디렉토리에 저장함으로써, 작업 시간 및 영상 정리 작업에 대한 효율성을 높였습니다.

## 가공

**Q 구축된 안면 이미지의 특징점 및 바운딩 박스는 어떤 식으로 활용하는 것이 좋을까요?**

A 안면 이미지의 특징점 및 바운딩 박스는 신원 확인, 얼굴 변환, 나이 인식 등 다양한 안면 응용 기술에 활용될 수 있습니다. 본 데이터셋의 특징점 및 바운딩 박스를 활용하여, 얼굴의 부위별 검출, 인식 등에 대한 기술 개발에 활용할 수 있을 것으로 예상됩니다.

**Q 구축된 가공 데이터의 자세한 정보를 알고 싶습니다.**

A 가공데이터는 이미지 파일과 함께 텍스트파일로 저장되어 있으며, 특징점의 경우 총 7종류로 텍스트파일에 적혀있는 순서대로 "0 : 코 끝, 1 : 오른쪽 눈 중심, 2 : 왼쪽 눈 중심, 3 : 오른쪽 입꼬리, 4 : 왼쪽 입꼬리, 5 : 오른쪽 귀 중심, 6 : 왼쪽 귀 중심"을 의미합니다. 바운딩 박스의 경우 순서대로 "0 : 얼굴 전체, 1 : 왼쪽 눈, 2 : 오른쪽 눈, 3 : 코, 4 : 입, 5 : 왼쪽 귀, 6 : 오른쪽 귀"를 의미합니다.

## 검수

**Q 상당히 많은 수의 이미지가 구축되어, 검수에 많은 시간과 인력이 동원되었을 것 같은데, 어떤 식으로 검수가 진행되었습니까?**

A 촬영된 안면 이미지는 촬영 당시 1차 확인을 거치고, 검수 팀에서 2차 확인을 실시하도록 하였으며, 촬영 조건이 일치하지 않는 경우에는 재촬영을 실시하도록 하였습니다. 또한 검수 효율화를 위해 촬영된 안면 이미지의 누락, 번짐 및 조건 오류를 실시간으로 점검하기 위한 도구를 개발하여 활용하였습니다. 이를 통해, 실시간으로 모든 촬영된 이미지에 대한 전수 검수를 실시하며, 중복 없이 빠른 검수가 가능하게 하였습니다.

## 배포

**Q 개인정보보호를 위한 안면 이미지 배포용 샌드박스 시스템이란 것은 어떤 것인지요?**

A 안면 이미지는 개인정보에 민감한 데이터로, 이에 대한 배포를 위해서는 반드시 적법한 절차를 거쳐야 합니다. 이를 위해서 데이터 취득 시, 개인정보활용 동의서를 촬영대상자에게 서명받았으며, 데이터의 유지/관리를 위해서 '가'급 보안 국가 연구소인 한국과학기술연구원(KIST)에 샌드박스 시스템을 구축하였습니다. KIST는 자체 보안서버를 통해 데이터를 안전하게 저장하고 배포하고 있으며, AI 허브사이트를 통해 안면 이미지 데이터셋을 요청하고 검증된 활용자만 데이터에 접근할 수 있도록 하고 있습니다.

**Q 기업에서의 데이터 활용도 가능한 것인지요?**

A 상업적 목적으로 활용하는 기업의 경우에는 학습하고자 하는 인공지능 네트워크를 제공 받아, 이를 샌드박스 시스템 내에서 학습을 진행한 이후 학습된 모델을 기업에 제공하도록 합니다. 상업적 목적으로 기업이 데이터 샌드박스 시스템을 이용할 경우 정보 보안 문제없이 대용량의 데이터를 효율적으로 처리할 수 있도록 데이터 접근 및 학습 API를 개발하여 활용하도록 하였습니다.

**Q 안면 이미지 데이터를 활용하여 인공지능 모델을 만들고, 학습 및 테스트를 진행하려면 어떻게 해야할까요?**

A 본 데이터의 활용도를 높이기 위해, 본 사업에서는 두 가지 안면 응용 기술(안면인식, 가상얼굴생성)에 대한 학습 및 테스트 샘플 코드를 공개하고 있습니다. 해당 샘플 코드는 안면 이미지 데이터셋을 코드상으로 읽어들일 수 있으며, 이에 대한 학습 및 테스트를 손쉽게 진행하실 수 있습니다. AI 허브(www.aihub.or.kr)의 안면 이미지 부분에서 다운로드하여 활용하실 수 있으며, 샘플 코드 내에 보다 자세하게 설명되어 있습니다.

**활용 예시**

**본인인증(1vs1 비교) 서비스**

- 본인인증(1vs1 비교) 서비스는 사진이나, 사진 촬영을 통한 안면 이미지에 대해 본인여부를 확인하는 서비스
- 출입통제 시 신분증에 있는 사진과 라이브로 촬영된 사진이 동일인물 여부를 인증하는 데 사용

**유사인물검색(1vsN 검색) 서비스**

- 유사인물검색(1 vs N 검색) 서비스는 사진이나, 사진 촬영을 통한 안면 이미지에 대해 기등록된 인물리스트에서 유사인물 리스트를 검색하는 서비스
- 실종아동 찾기 등을 위한 검색서비스, 범죄수사를 위한 검색서비스 등에 활용

**기타 서비스 활용 모델**

- 안면인식 기술을 활용한 응용서비스는 안면인식 기술의 속도와 성능의 비약적인 발전으로, 기존의 전통적인 사업모델인 출입통제, 범죄수사 뿐만 아니라, 금융결제, CCTV 영상분석의 영역까지 다양하게 확대

8

# 위해물품 엑스레이 이미지 AI 데이터

구축 기관 (주)엠폴시스템, (주)인씨스, (사)국가사이버안전연합회

## 필요성

- 현재까지 엑스레이 검색 업무는 보안요원이 육안으로 검색하는 방법에 의존하여 개인의 능력(경력, 컨디션, 판단력)에 따라 판독의 결과가 다르기 때문에 업무의 프로세스가 정규화되어 있지 않고 위해물품이 소형화되어감에 따라 Human-Error가 발생

- AI 기반 위해물품 탐지 기술은 데이터를 통해 물품을 탐지하기 때문에 해당 기술은 새로운 시장을 발굴 가능

## 구축 내용

- 엑스레이 데이터의 실효성 제고를 위해 국토부 고시(항공/항만 기내 반입 금지 물품) 대상 물품 40만건과 물리보안 대상 물품(저장매체) 7만건의 데이터셋을 구축
  - 다양한 모델 네트워크를 사용하기 위해 Segmentation Annotation과 Pascal VOC Annotation을 동시에 제공

## 구축 공정

수집 → 정제 → 가공 → 검수 → 배포

## 수집

**Q** 엑스레이 물품 스캐너에서 영상을 생성하는 경우 별도의 장비가 필요한가요?
아니면 엑스레이에 백업된 데이터를 이용하여 수집하나요?

A 엑스레이 스캐너 이미지는 고화질의 데이터를 생성해야 하기 때문에 백업된 이미지로는 좋은 학습 효과를 낼 수 없습니다.
특히 손실 압축을 사용하는 JPG 이미지는 실시간 판독을 하는 비교 대상 이미지와 DPI 및 픽셀 구성에 줄 수 있기 때문에 PNG 형태의 무손실 압축 영상을 사용하는 것이 좋습니다. 이미지 생성을 위해 장비가 필요합니다. HD급 영상을 가공할 수 있는 HDMI 인코더 또는 HDMI Grab 보드는 필수 입니다.

**Q** 이미지 생성을 위한 물품의 선정에 있어 반영되어야 할 사항은 어떤 것들이 있을까요?

A 당 과제의 데이터셋은 실제 공항, 항만, 연구소의 보안 검색 업무를 담당하는 CBT 교육 이수자를 대상으로 실제 데이터화하는 작업을 진행하였습니다.
국토부에서 고시한 기내 반입 금지 물품과 저장매체를 대상으로 패턴이 다른 샘플을 대상으로 데이터를 생성하였으며 실제 사용되는 캐리어, 가방, 바스켓 등에 적재하여 데이터를 생성하였습니다. 위와 같이 실제 데이터의 생성도 중요하지만 기본 데이터(물체 원형)의 생성도 중요합니다.

**Q** 엑스레이 제조사별 이미지 생성 비율을 1:1:1로 정의하여 데이터를 구축한 사유는 무엇인가요?

A 국내 판매되는 제조사 중 시장 점유율이 높은 3개의 제조사 제품으로 선정한 것은 데이터 활용에 대한 효율성을 반영한 것입니다. 최소 학습에 필요한 물품별 이미지는 AI 학습을 위해 필요한 데이터이기 때문에 엑스레이 시장 점유율이 낮다고 해서 데이터 수량을 감소시키면 AI 학습을 하는 데 있어 어려워집니다. 이에 물품 학습을 위해 필요한 데이터를 1:1:1의 비율로 생성하였습니다. 제조사 별로 Pseudo-6 Color가 조금씩 다르기에 동일 데이터가 필요합니다.

## 정제

**Q** 엑스레이 이미지의 정제는 어떤 사항을 정제 대상으로 판단하나요?

A 물리적인 장애로 인해 물품 이미지가 왜곡되거나 원래의 형태가 변형이 되었을 경우 정제 대상 이미지입니다. 스캔 도중 내부에서 물품이 밀린 경우 이미지가 왜곡되는 경우가 발생합니다. 또 제너레이터 이상으로 노이즈가 발생한 이미지 역시 정제 대상 이미지입니다. 엑스레이의 설정 실수로 발생하는 이미지 왜곡도 정제 대상 이미지입니다. 금속 등과 같은 밀도가 높은 물체의 경우(아래쪽 물품이 보이지 않는 경우) 데이터셋으로 활용이 가능하지만 어노테이션은 제외합니다.

**Q 데이터 중첩에 촬영의 경우 데이터셋에서 필수 요소인가요?**

A 엑스레이 검색 분야의 경우 Detection, Classification, Localization의 세 가지 알고리즘을 구현해야 하는 고난이도 기술이기 때문에 정상적인 이미지와 중첩되는 이미지의 다양화가 중요한 요소입니다. 카메라 영상과 달리 중첩의 경우 투과율에 따라 무시되는 상황이 있기 때문에 어떤 물체와 중첩되었는지 얼마나 중첩되었는지가 중요합니다. 그래서 데이터셋 구축 시 단일품목(해당 물품 30%) 중 어떠한 물체도 넣지 않은 Pure한 상태와 일반 품목을 중첩한 비율을 5:5로 계획하여 구축하였습니다.

## 가공

**Q 각 물품별 생성 규칙에 대한 비율을 선정한 이유는 무엇인가요?**

A 해당 데이터셋의 경우 기본 학습과 실제 데이터화를 접목하여 구성하였습니다. 실제 공항, 항만의 엑스레이 검색의 경우 단일/기본(15%)은 물체의 원래 형태를 다양한 각도와 가방, 캐리어 같은 Baggage를 고려하여 기본 학습을 위한 데이터셋을 구성하였고 단일/비품목(15%)의 경우 일반적인 옷 등과 같이 위해물품이 아닌 경우에 대한 상황을 고려하여 배정하였고, 복합/품목(40%)의 상황은 위해물품을 다수 소지하였을 경우에 알고리즘(Detection, Classification, Localization)상에 복잡하게 표현이 되고 일반적으로 보안요원들의 육안감시가 가장 어려운 이미지이기 때문에 많은 비중으로 데이터셋을 다양하게 구축하였습니다. 복합/비품목(30%)도 위 내용과 동일한 개념으로 구축하였습니다.

**Q 중첩된 물품 이미지가 생성이 되어 구성이 되는데, 이런 중첩된 사항에 대한 어노테이션은 어떻게 처리하였나요?**

A Z-Number에 따른 엑스레이 투과율이 다르게 표현이 되는 엑스레이 장비 특성상 투과대상 물체가 육안으로 특징이 보이지 않는 경우는 레이블링 영역에서 배제가 됩니다. 다만 투과하여도 물체의 특징이 육안으로 보여지는 경우 레이블링에 포함합니다. 그리고 중첩의 정도가 과반 이상일 경우에는 중첩 부분을 제외하여 레이블링 작업을 진행하였습니다.

## 검수

**Q Segmentation Annotation 정보를 읽어 자동으로 Pascal Voc Annotation 파일로 변환하는데 이를 효율적으로 검증할 수 있는 방법이 있을까요?**

A Poly-Line Annotation을 자동으로 Pascal Voc로 변환하는 알고리즘을 개발하여 공개하였고 이를 이용하여 좌표를 Crop하는 기능을 추가로 개발하여 검수 시 썸네일 형태로 Annotation을 확인할 수 있습니다. 데이터셋은 장비-물품-샘플의 구조로 폴더링되어 있어 검증 시 편리하게 이용할 수 있습니다.

**Q 검수 단계에서 오류가 발생한 데이터에 대한 재가공 작업 방법 및 데이터셋 포함 절차는 무엇인가요?**

A 검수 단계에서 잘못 가공된 데이터를 확인하고 이에 대한 이미지 파일명을 COCO Annotator에서 다시 열기를 통해 잘못된 어노테이션을 삭제 또는 수정을 하여 업로드하면 JSON 파일의 형태로 저장이 되고 이를 다시 Pascal Voc 자동 변환 프로그램으로 Load하여 확인 가능합니다. COCO Annotator를 통해 수정된 Annotation은 DBMS에 자동으로 업데이트됩니다.

## 배포

**Q 배포된 데이터셋은 두가지 어노테이션이 제공되는데 각각의 차이점이 무엇인가요?**

A 두 개의 어노테이션의 경우 활용적인 측면에서 공통점은 Object Detection과 Classification을 하기 위한 어노테이션 정보라는 점입니다. 그러나 서로 다른 모델 네트워크를 사용하기 때문에 하나의 이미지를 중심으로 두 개의 포맷으로 나눠지게 됩니다

Segmentation 방식의 Poly Line의 경우 Mask RCNN 모델 및 네트워크를 이용하여 학습이 되고 Pascal Voc의 경우 Yolo, SSD, RCNN 모델 네트워크로 학습하는 어노테이션 데이터입니다.

**활용 예시**

**공항 항만 위해물품 탐지 서비스**

- 공항 및 항만 등의 여행자 위해물품을 자동으로 탐지하는 서비스

**연구소 및 국가시설 저장매체 탐지 서비스**

- 기업 및 국가의 비밀 문서 또는 정보를 대외로 반출하는 상황에 대한 저장 매체의 외부 반출 및 내부 반입을 금지하기 위한 자동 탐지 서비스

**물류시스템에서 이상 물체 탐지 서비스**

- 세관 및 물류센터의 배송 물품 이외의 금지 물품 탐지 서비스

9

# 질병진단 이미지 AI 데이터

| 구축 기관 | 국립암센터, (주)루닛, (주)필라테크, (주)헬스허브, 건양대학교 병원, (주)에임즈, (주)인피니그루 |
|---|---|

## 필요성

- 해외는 다양한 분야에서 인공지능 기술 개발에 활용할 인공지능 학습 데이터 및 학습 환경 등 기반 인프라가 활성화되는 반면, 국내의 경우 개인정보 침해, 저작권, 초상권 등 다양한 이슈로 양질의 데이터 개방 및 활용이 어려운 상황
- 의료 산업 발전을 위한 인공지능 관련 신생기업들의 활성화를 위한 효율적이고 체계적인 의료 지식베이스 구축이 필요

## 구축 내용

- 유방암 데이터(총 3만장) : 정상 4,500명(1만 8,000장), 양성종양 1,500명(6,000장), 악성종양 1,500명(6,000장)
- 안저질환 데이터(총 4,500장): 일반안저영상(2,500장)[정상소견 1,500장, 녹내장소견 1,000장], 광각안저영상(2,000장)[정상소견 600장, 당뇨성 망막변증 700장, 황반변성 500장, 정맥폐쇄 200장]

## 구축 공정

수집 → 정제 → 가공 → 검수 → 배포

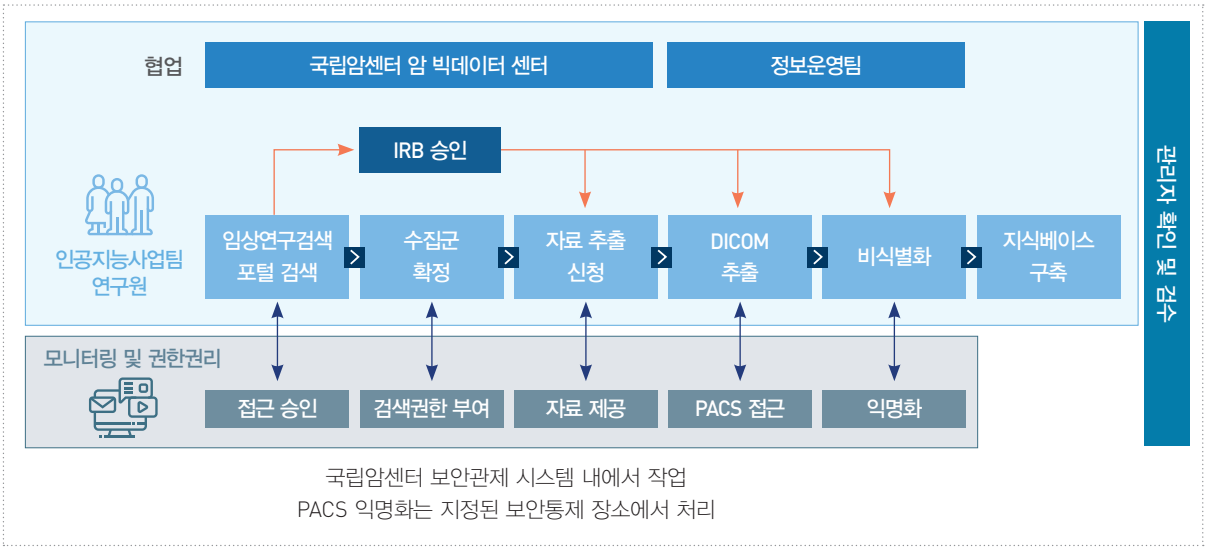
## 수집

**Q** 의료 데이터 수집 방법에 대해서 설명해 주십시오

A

**국립암센터**

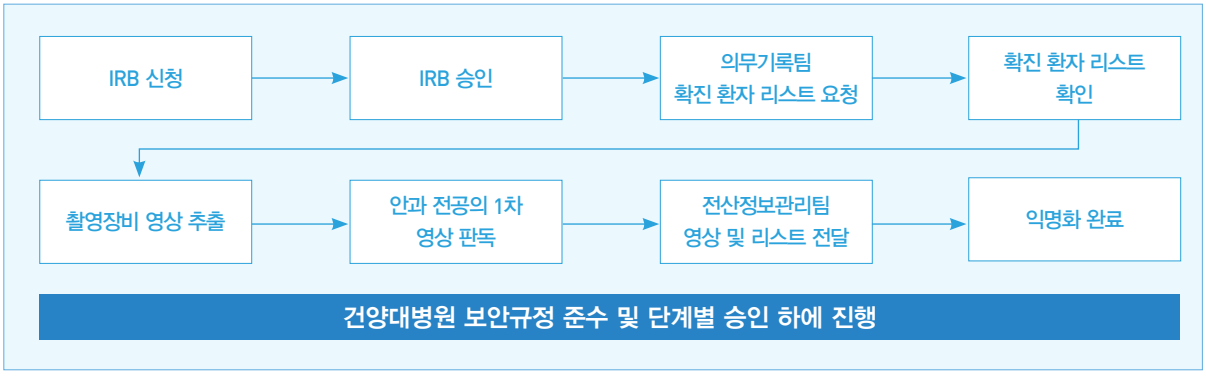
국립암센터에서 운영하고 있는 의료데이터베이스(EMR, PACS)에서 데이터베이스 질의를 통해, 유방암 진단을 위해 촬영한 유방촬영영상 및 해당 영상의 임상정보(Radiology Report, Pathology Report) 등을 수집. 개인정보보호법 등의 이유로 개인 또는 타 기관에서 각 기관의 의료데이터베이스에 직접적인 접근은 불가능합니다. 또한, 데이터베이스 접근을 위해 반드시 IRB 승인이 선행되어야 하며, 승인된 연구자만이 의료데이터베이스에 접근하여 수집할 수 있습니다.




국립암센터 보안관제 시스템 내에서 작업
PACS 익명화는 지정된 보안통제 장소에서 처리

**건양대병원**

건양대병원에서 운영하고 있는 의료데이터베이스(EMR)에서 데이터베이스 질의를 통해, 안저질환별 확진 환자 리스트 확인 후 안저촬영장비에서 영상을 추출합니다. 개인정보보호법 등의 이유로 개인 또는 타 기관에서 각 기관의 의료데이터베이스에 직접적인 접근은 불가능합니다. 또한, 데이터베이스 접근을 위해 반드시 IRB 승인이 선행되어야 하며, 승인된 연구자에 한하여 의료데이터베이스에 접근하여 수집할 수 있습니다.



**Q 데이터 종류 및 구성이 어떻게 되어있나요?**

A **국립암센터** 만 19세 이상 성인 여성환자들을 대상으로 유방암 1,500명(6,000장), 양성종양 1,500명(6,000장), 정상유방 4,500명(18,000장) 유방촬영영상 데이터를 수집하였습니다. 4-view (RCC, RMLO, LCC, LMLO) paired 유방촬영영상으로 한 환자당 4개의 이미지 데이터를 수집하였습니다. 유방촬영영상 이미지 데이터는 국제 표준인 의료용 디지털 영상 및 통신(Digital Imaging and Communications in Medicine, DICOM) 형식으로 구성되어 있습니다.

**건양대병원** 일반정상영상 1,500장, 일반녹내장영상 1,000장, 광각정상영상 600장, 광각황반변성영상 500장, 광각당뇨성망막병증영상 700장, 광각망막정맥폐쇄영상 200장의 안저촬영영상 데이터를 수집. jpg 형식의 이미지 파일로 수집하였으며, 영상에 해당하는 메타데이터도 함께 구성하였습니다.

## 정제

**Q 각 질병에 대한 수집 기준은 무엇입니까?**

A **국립암센터**

1) 유방암 세부 기준 : 조직검사 결과 악성으로 확진된 환자의 유방촬영영상 이미지이며, 악성의 경우 조직검사를 기준으로 선정하는 것이 일반적(ScreenPoint Medical BV, 2018)이며, 내원해서 진료받은 경우 수술일 기준으로 210일 안에 병리검사를 하고, 300일 안에 유방촬영영상을 확보한 경우와 외부에서 진료 후 내원한 경우 내원하기 전 외부에서 병리검사 후 항암치료로 진단결과의 bias가 수술일 기준으로 90일 이내에 병리검사와 유방촬영영상을 확보한 이미지를 수집하였습니다.
2) 양성 세부 기준 : 영상 판독상 양성(BI-RADS category 2,3)이며, 1년 이상 추적검사 결과 여전히 양성인 환자의 유방촬영영상을 수집하였습니다.
3) 정상 세부 기준 : 영상 판독상 정상(BI-RADS category 1)이며, 1년 이상 추적검사 결과 여전히 정상인 환자의 유방촬영영상을 수집하였습니다.

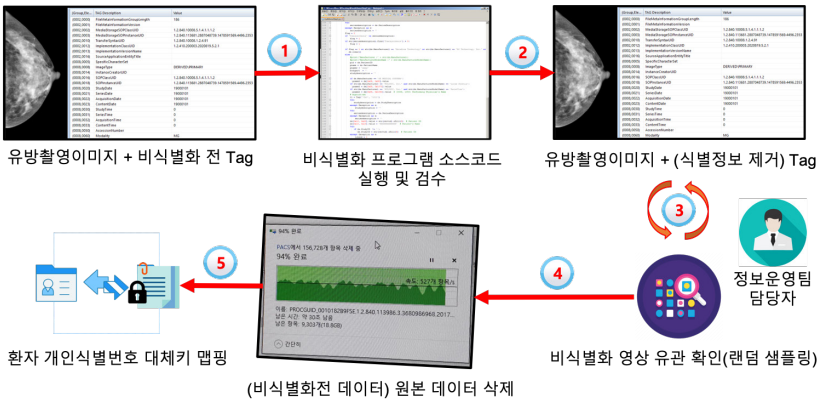
**건양대병원** 안저질환 데이터는 일반안저영상(녹내장)과 광각안저영상(황반변성, 당뇨성망막병증, 망막정맥폐쇄)을 나누어 수집하고, 각 질환별 판독 기준은 아래와 같습니다.

- 녹내장 : 1) C/D ratio 0.5 이상, 2) optic disc hem orrhage, 3) RNFL defect
- 당뇨성망막병증 : 1) 망막출혈, 2) 망막의 exudate, 3) 미세동맥류
- 황반변성 : 1) 황반출혈, 2) 황반의 drusen
- 망막정맥폐쇄 : 1) 망막의 화염상 출혈, 2) 망막 혈관의 thrombosis

## Q 의료 데이터의 비식별화 과정은 어떠한 절차로 이루어지나요?

A **국립암센터**

국립암센터의 EMR, PACS 등은 직접 접근이 아닌 이미 익명화된 임상연구 검색 포털 시스템을 통해 1차적인 개인정보보호가 이루어지므로, 2차적으로 일부 조건을 추가하여 수집군을 최종 확정하고 이후에 개인 식별화가 가능한 정보를 삭제하여 비식별화를 실행하였습니다. 각 단계별 승인과 추출에 대한 모니터링 및 작업 검수가 이루어져 개인정보보호법 준수와 국립암센터 보안절차를 준수하였습니다.




**건양대병원**

개인정보에 대한 문제가 없도록 각 기관에서는 기관생명윤리위원회(Institutional Review Board, IRB)의 허가가 통과된 이후에 데이터셋을 구축. 건양대병원은 안저질환별 확진 환자리스트 확인을 위해서만 EMR 접근이 필요하며, 환자 리스트를 기반으로 1차 판독이 완료된 데이터에 한해 익명화를 진행하였습니다. 각 단계별 승인과 추출에 대한 모니터링이 이루어져 개인정보보호법 준수와 건양대병원 보안절차를 준수하고, 이후 익명화가 완료된 데이터에 대해서만 2차 판독 및 검수를 진행하였습니다.

## 가공

## Q 메타정보는 어떻게 구성되어 있나요?

A **국립암센터**

영상 자료 이외의 해당 영상의 매칭되는 환자에 대해 검사장비, 모델 BIRADS Category가 있으며, 악성 환자에 한해서는 병변에 대한 추가적인 정보를 위해 수술 후 병리보고서의 일부가 포함되어 있습니다.

**건양대병원**

안저영상은 jpg 형식의 이미지 파일로 의료영상 표준인 DICOM의 메타데이터와는 다릅니다. 안저영상에 해당하는 내용을 새롭게 구성하여 메타데이터를 정리하고 필드는 일련번호, 이미지파일명, 일반/광각, 정상/질환 코드, 레이저치료 유무, 좌/우, 촬영 장비와 같으며, 서버 DB에 저장된 라벨결과 및 영상을 기반으로 작성하였습니다.

**Q 이미지정보는 어떻게 구성되어 있나요?**

A

**국립암센터**

각 케이스는 4장의 DICOM 영상, 즉 RCC, LCC, RMLO, LMLO 영상으로 구성됩니다. DICOM 데이터의 비식별화를 위해 공개 시 개인 식별이 가능한 모든 태그를 삭제하였습니다. 환자 ID는 환자의 병변 상태를 기준으로 새로 아이디를 부여합니다 (ex: 음성:00001, 양성:10001, 정상:20001). 환자 생년월일과 수행된 프로세스 시작 날짜와 같은 형식의 태그는 '19000101'로 고정하였습니다.

**건양대병원**

각 안저질환별 폴더로 구분하여 이미지정보를 구성하였습니다. 이미지파일명은 비식별화를 위해 새로운 명명규칙을 적용하고, 일반은 General, 광각은 Wide를 부여하여 일반안저영상과 광각안저영상을 구분합니다. 질환은 정상 Normal, 녹내장 Glaucoma, 황반변성 AMD, 당뇨성망막병증 DMR, 망막정맥폐쇄 RVO를 부여하여 구분합니다.
(ex : 일반정상영상 – General_Normal_0001_L.jpg

**Q 어노테이션 정보는 어떻게 구성되어 있나요?**

A

**국립암센터**

악성 환자의 유방촬영영상 이미지에 대해서만 악성 병변의 어노테이션을 수행하였으며, 해당 악성 병변에 대한 어노테이션은 JSON 파일로 저장하였습니다.

**건양대병원**

해당없음

## 검수

**Q 검수는 어떠한 방식으로 진행되었나요?**

A

**국립암센터**

데이터 가공 절차를 통해 Type 1(1차 판독 전문의 판단 하에 병변이 실제로 존재하고 그에 대한 병변을 충분히 polygon 형태로 그릴 수 있는 상태)으로 분류된 영상에 한해서 2차 데이터 품질 검수를 진행하였습니다.
2명의 판독의가 교차검증을 시행하여, 오류가 있을 경우 저작도구를 활용하여 수정하고, 해당 내용에 대해 기록을 남겨 2명의 전문의의 합의하에 최종 판단합니다. 2차 교차판독을 통해 1차 & 2차 Double Reading의 불일치 혹은 Case 오류 등의 문제가 있는 환자는 제외합니다.
악성 유방촬영영상 선정기준에 부합하는 악성 영상자료에 대해, 추적 조직검사가 수행된 악성 병변의 위치를 Radiology Report 및 Pathology Report를 참고하여 영상의학과 전문의가 어노테이션을 수행하여 검수합니다.

**건양대병원** 안과 전공의들이 1차 판독 후 익명화된 데이터에 대해 저작도구를 활용하여 파트별(녹내장, 망막) 전문의 2인이 2차 판독 및 교차검증을 진행합니다. 판독자별 프로젝트 부여 후 라벨링을 진행하고, 서버 DB로 저장된 라벨 결과를 바탕으로 2인의 전문의 간 합의도가 일치하는 데이터만을 최종 데이터셋으로 활용합니다. 라벨링 진행 시 판독자 본인이 저작도구의 '리뷰모드'기능을 통해 자체 점검을 할 수 있으며, 안저질환별 확진된 환자의 영상으로 정의하기 때문에 100% Ground-Truth로 신뢰도가 높은 데이터를 보장할 수 있습니다.

**Q 데이터에 대한 유효성은 어떻게 평가하였나요?**

A **국립암센터** 유방촬영영상 이미지의 압축과 비압축본에 대하여 정상과 악성의 판독 시범 모델의 검증을 위하여 5-Fold Cross Validation을 수행하였습니다. 악성, 정상(양성포함)의 데이터를 5개로 Random Split를 하고 4개의 Sset으로 학습하고 1개의 Set으로 성능검증하는 방법으로, 전체 5개의 set에 대하여 성능 검증을 5번 반복하여 검증 성능의 평균으로 목표치 달성여부를 확인하였습니다. 평가 척도로는 민감도(Sensitivity), 특이도(Specificity), ROC(Receiver Operating Characteristics) Curve의 AUC(Area Under the Curve)를 측정하여 성능을 평가하였습니다.

최근 FDA 허가 완료(Rodríguez-Ruiz, et al., 2018)된 딥러닝 기반 유방촬영영상 판독보조 소프트웨어에서는 약 1만 8,000케이스의 유방촬영영상 이미지에 대하여 판독모델과 비교하여, 본 과제 제안 시 목표로 설정했던 성능 목표치를 초과 달성하였습니다.

**건양대병원** 일반안저영상과 광각안저영상의 판독 시범 모델의 검증을 위하여 10-Fold Cross Validation을 수행하였습니다. 일반안저영상의 경우 정상과 녹내장, 광각안저영상의 경우 정상과 황반변성, 정상과 당뇨성망막병증, 정상과 망막정맥폐쇄 데이터셋을 10개로 Random Split하였습니다. 9개의 set으로 학습하고 1개의 set으로 성능 검증하는 방법으로, 전체 10개의 set에 대하여 성능 검증을 10번 반복하여 검증 성능의 평균으로 목표치 달성 여부를 확인하였습니다. 평가 척도로는 민감도(Sensitivity), 특이도(Specificity), ROC(Receiver Operating Characteristics) Curve의 AUC(Area Under the Curve)를 측정하여 성능을 평가하였습니다. 본 과제 제안 시 목표로 설정했던 성능 목표치를 초과 달성하였습니다.

## 배포

**Q 배포방식과 사용방법에 대해서 설명해주세요.**

A **국립암센터, 건양대병원** IRB를 사전에 승인받은 연구자들에 한해, IRB 통지서 제출 후 데이터에 대한 접근이 가능합니다. 자세한 절차는 각 기관 담당자에게 문의 바랍니다.

**활용 예시**

### ≡ 교육 콘텐츠 제작, 스타트업 취·창업 촉진

- 국내외 적으로 의료 인공지능 관련 업무 종사자가 부족한 상황이며 기초 인력부터 전문 인력양성이 시급한 시점으로, 본 사업에서 구축된 데이터는 완전히 익명화된 데이터이기 때문에 개인 정보 침해의 염려 없이 의료 인공지능 전문 인력을 조기에 양성하기 위한 기초 실습자료로 다양하게 활용 가능

- 초·중·고 학생에서부터 일반 전문직에 이르기까지 교육자료 콘텐츠로 널리 활용될 수 있으며 산업계에서도 의료 인공지능 분야의 스타트업 취창업을 촉진하기 위한 밑바탕으로 다양하게 활용

### ≡ 영상판독 보조 시스템 개발

- 한국의 의료 인공지능 헬스케어 업체 뷰노(VUNO)에서 개발한 'VUNO MED – FUNDUS AI'는 안저영상을 분석하여 대표적 당뇨병 합병증인 당뇨성망막병증이나 녹내장 등 총 12가지 병변에 대해 판독하는 소프트웨어

- 해당 소프웨어는 안저의 정상, 비정상 여부를 판독하고 빈도가 높은 소견에 대한 병변 구역을 표시해 주는 인공지능 소프트웨어 의료기기로써 소프트웨어의 판독 결과를 보다 직관적으로 이해할 수 있어 안저촬영을 실시하는 모든 의료기관과 건강검진센터 등에서 편리하게 사용할 수 있을 것으로 기대

- 특히, 안과의사가 없는 1, 2차 의료기관을 위한 스크리닝 도구로서 큰 역할 수행 가능

10

# 이상행동 CCTV 영상 AI 데이터

| 구축 기관 | (주)마인즈랩, (주)마인즈앤컴퍼니, 수원시청 |
|---|---|

## 필요성

- 이상행동 CCTV 영상 데이터는 안면 데이터, 의료 데이터와 더불어 규제가 특히 심한 데이터로서 실제 위급 상황에 대한 영상 데이터를 확보하기가 불가능
- 해외의 영상 데이터는 국내 현실에 맞지 않는 데이터가 대다수로 현실의 영상 데이터를 확보하기가 어려움

## 구축 내용

- 폭행, 싸움, 주취행동 등 12가지로 정의한 이상행동 분류를 기반으로 핵심 시나리오를 작성하여 8400컷(약 5분) 이상의 총 700시간(실내 촬영 300시간, 실외 촬영 400시간)분량의 이상행동 영상 데이터 촬영 제작

## 구축 공정

수집 → 정제 → 가공 → 검수 → 배포

## 수집

**Q 12가지로 분류된 이상행동 영상의 비율이 다른 이유가 있나요?**

A 수원시의 협조를 받아 12가지 이상행동의 분류체계를 정리하였으며 최근 3년 간 실제로 발생한 사건들을 중심으로 데이터를 제작하여 현실의 비중에 맞게 데이터를 제작하였습니다. 다만 전체 비율에서 9%를 기준으로 4% 내외의 오차로 비중을 맞추었습니다. 최소한의 학습을 위하여 각 이상행동별로 최소 20시간 이상의 데이터는 확보하였습니다.

**Q 현재 대부분의 CCTV가 FHD인데 4K로 촬영하신 이유는 무엇인가요?**

A 4K로 찍은 영상은 FHD로 품질을 변환하기도 쉬우며 미래지향적으로 봤을 때 CCTV의 품질도 4K로 향상될 것이기 때문에 앞으로도 지속적으로 활용할 수 있는 데이터로 만들기 위해서 4K로 촬영하였습니다.

**Q 동일한 시나리오가 있는 것 같습니다.**

A 하나의 시나리오를 기준으로 다양한 각도를 확보하기 위해 실내 촬영은 다른 각도의 3대의 카메라로 촬영하였고 실외 촬영은 다른 각도의 2대의 카메라로 촬영하여 다양한 각도의 영상을 확보하여 학습데이터로서 품질을 높였습니다.

## 정제

**Q 실내 크로마키 촬영분에서 배경이 없는 이유는 무엇인가요?**

A 영상 데이터를 활용하여 학습 진행 시 배경이 오히려 학습 데이터의 품질을 떨어뜨리는 경우가 있어 배경을 따로 입히지 않고 데이터를 제작하였습니다. 단 실내 촬영 배경을 입히기 쉽도록 크로마키 키를 추출한 영상과 실내 촬영 원본 데이터를 동시에 제공을 해드리고 있습니다. 또한 배경이 필요한 경우 실외 촬영분을 같이 제공해드리고 있습니다.

**Q 개인정보가 들어가 있는 데이터가 있나요?**

A 기본적으로 등장하는 배우들 이외의 일반인들은 촬영 시 협조를 구하고 신체가 등장해도 문제가 없다는 확답을 받은 분만 영상데이터에 들어가 있으며 배우들은 모두 영상 공개에 대한 동의서를 받아두었습니다. 차량 번호판의 경우 촬영 차량 이외에는 학습 시 문제가 없도록 옅은 블러처리를 통해서 사람의 시각으로만 보이지가 않도록 정제를 하였습니다.

## 가공

**Q** XML에서 제공되는 태그를 단 기준은 무엇인가요?

A 해당 기준은 이상행동의 시작지점과 끝나는 지점, 액션이 시작하는 지점과 끝나는 지점을 태그를 달아 이상행동과 정상행동을 분류하였습니다. 필요하시다면 태깅작업의 가이드를 제공해드릴 수 있습니다.

**Q** 전체 분량은 얼마나 되나요?

A 전체 이상행동 영상 촬영 분은 실내 데이터(실내 크로마키 촬영) 300시간, 실외 데이터(실외 외부 촬영) 400시간을 기본으로 제공하며 실내 원본 데이터(실내 크로마키 촬영) 300시간을 추가로 제공합니다.

## 검수

**Q** 데이터에 대한 검수 방식은 어떻게 진행을 하셨나요?

A 데이터는 촬영 및 정제 후에 1차 검수를 Checker(1차 검수자)가 진행을 하고 2차 검수를 Header(2차 검수자)가 진행 후 Chief(3차 검수자) 샘플링을 통해서 검수를 완료합니다.

## 배포

**Q** 이상행동 데이터셋의 배포 형식에 대해서 알려주세요.

A 데이터셋은 기본적으로 영상 데이터(.mp4)와 태그 데이터(.xml)로 짝을 이뤄 구성되어 있으며 태그 데이터에는 영상에 대한 정보 및 태그 정보가 입력되어 있습니다.

**Q** 레이블링 도구에서 추가로 태그 작업을 할 수가 있나요?

A 레이블링 도구는 영상 데이터를 업로드하고 추가로 태그 작업을 할 수 있도록 구성이 되어 있습니다.

### 활용 예시

**이상행동 감지 시스템**

- 각 지방자치 단체에서 사용하고 있는 CCTV에 연동하여 이상행동 발생 시 관제사들에게 알람을 제공

**요보호자 검색 시스템**

- 주취행동과 배회 영상을 활용하여 행동 검출 시스템을 이용, CCTV에서 치매나 길을 잃어버린 아이를 찾고 해당 행동이 등장하는 지역을 검색하여 위험한 상황을 예방 가능

**발 행 일** 2020년 3월 20일

**발 행 처** 과학기술정보통신부 · 한국정보화진흥원

**작 성** 한국정보화진흥원 오성탁 본부장
한국정보화진흥원 신다울 팀장
한국정보화진흥원 김재욱 수석, 유호진 수석, 홍효진 수석
한국정보화진흥원 김민준 책임, 송민경 책임, 이용진 책임
한국정보화진흥원 이진희 선임, 전진우 선임, 최형인 선임

**디자인, 인쇄** 전우용사촌(주) 02-426-4415

---

과학기술정보통신부

NIA 한국정보화진흥원