탑승자 인지 프롬프트와 이중 시각 경로를 활용한 경량화된 차량 실내 행동 인식 모델

간저릭, 유재웅, 조용현
한국전자기술연구원

gnzrg25@keti.re.kr, jaewoong.yoo@keti.re.kr, anorange7417@keti.re.kr

Lightweight In-Cabin Action Recognition via Passenger-Aware Prompting and Dual Visual Encoders

Ganzorig Gankhuyag, Jaewoong Yoo, Yonghyeon Cho
Korea Electronics Technology Institute (KETI)

요 약

차량 실내(In-Cabin) 행동 인식은 협소한 공간, 탑승자 간의 가림, 급격한 조명 변화, 그리고 카메라의 제한된 시야각 등으로 인해 기존의 범용 비디오-텍스트 정합 모델을 직접 적용하기에 한계가 있다. 본 연구에서는 이러한 제약을 극복하고 엣지 디바이스 환경에서도 효율적으로 동작할 수 있는 모델을 제안한다. 제안하는 프레임워크는 백본 선택이 가능한 Image Feature Extractor 와 포즈 정보를 처리하는 Keypoint Feature Extractor 의 이중 경로 구조를 갖는다. 특히 본 연구에서는 Image Feature Extractor 로 CLIP 을 채택하여 사전 학습된 풍부한 시각적 특징을 활용하였으며, 이를 탑승자 인지 프롬프팅과 결합하였다. 자체 구축한 인캐빈 RGB 데이터셋을 이용한 실험 결과, 제로샷 설정에서는 제한적인 성능을 보였으나, 파인 튜닝 시 95.2%의 높은 정확도를 달성하여 제안 모델의 실효성을 입증하였다.

Ⅰ. 서 론

최근 자율주행 기술의 고도화와 함께 차량 내 탑승자의 안전과 상태를 실시간으로 모니터링하는 인캐빈 모니터링 시스템(ICMS)의 중요성이 급증하고 있다 [1]. 탑승자의 행동을 인식하는 기술은 안전벨트 착용 여부 확인, 운전자 부주의 감지, 승객의 돌발 행동 예측 등 안전 시스템의 핵심 요소로 작용한다. 그러나 차량 실내 환경은 일반적인 실내외 환경과 달리 매우 협소하며, 시트나 다른 승객에 의한 신체 가림이 빈번하게 발생한다. 또한, 터널 통과나 야간 주행 시의 급격한 조명 변화와 노이즈는 인식 성능을 저하시키는 주요 원인이 된다.

기존의 행동 인식 연구들은 대규모 비디오 데이터셋으로 학습된 3D CNN 이나 트랜스포머(Transformer) 기반의 거대 모델을 주로 사용해왔으나, 이는 연산량이 과도하여 차량용 임베디드 시스템이나 엣지 디바이스에 탑재하기에는 부담이 크다. 이에 본 연구에서는 RGB 단일 모달리티만을 사용하여 계산 효율성을 확보하면서도, 텍스트-비전 멀티 모달 학습의 장점을 극대화할 수 있는 모델을 제안한다. 본 모델은 시각적 특징뿐만 아니라 신체 관절의 기하학적 정보를 활용하는 Keypoint Feature Extractor 를 병렬로 배치하여 가림이 심한 환경에서도 강인한 인식 성능을 목표로 한다.

Ⅱ. 본론

제안하는 시스템 구조 제안하는 시스템은 크게 탑승자 검출, 이중 특징 추출, 그리고 유사도 기반 행동 분류 단계로 구성된다.

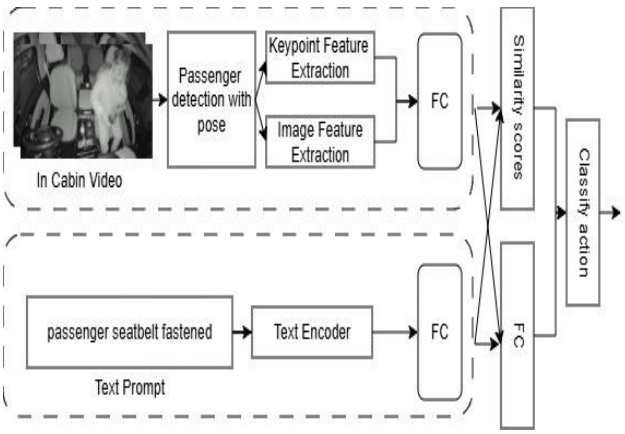

그림 1 차량 실내 행동 인식 모델

탑승자 및 포즈 검출 입력된 RGB 프레임에서 먼저 객체 검출 모델을 통해 좌석 별 탑승자를 개별적으로 인식하고, 각 탑승자에 대한 2D 키포인트 좌표(x, y)와 가림 여부를 추출한다. 이 과정은 복수의 탑승자가 존재하는 상황에서도 개별 행동을 정밀하게 분석하기

위한 전처리 단계이다. 본 연구에서는 신속하고 정확한 탑승자 검출 및 포즈 추정을 수행하기 위해, 객체 인식 분야에서 성능이 입증된 YOLO [2] 기반의 아키텍처를 채택하여 사용하였다.

이중 시각 경로로 추출된 정보는 두 개의 병렬 인코더를 통해 처리된다.

1. Keypoint Feature Extractor: 본 연구에서는 단순히 좌표값만을 사용하는 것이 아니라, 관절 간의 공간적 연결성과 시간적 변화를 효과적으로 학습하기 위해 GNN 구조를 차용한 Keypoint Feature 를 설계하였다. 신체 관절을 그래프의 노드로, 뼈대를 엣지로 모델링함으로써, 시각 정보만으로는 구분이 어려운 미세한 자세 변화나 신체 일부가 가려진 상황에서도 구조적인 특징을 효과적으로 추출한다.

2. Image Feature Extractor: 탑승자의 시각적 외형 정보를 추출하기 위한 모듈이다. 이 모듈은 ResNet, ViT 등 다양한 백본 네트워크를 유연하게 적용할 수 있도록 설계되었다. 본 연구에서는 대규모 이미지-텍스트 데이터로 사전 학습되어 일반화 성능이 뛰어난 CLIP[3]의 Image Encoder 를 백본으로 채택하였다. 입력 이미지는 이 인코더를 거쳐 고차원의 시각 임베딩 벡터로 변환된다.

탑승자 인지 프롬프팅 단순한 클래스 이름(예: "drinking") 대신, "passenger seatbelt fastened", "driver talking on phone"과 같이 주체(승객/운전자)와 행위, 객체를 명시한 문장형 프롬프트를 구성한다. 이는 CLIP Text Encoder[3]를 통해 텍스트 임베딩으로 변환되며, 앞서 추출된 시각 및 키포인트 특징과 결합되어 최종 분류에 사용된다.

학습은 텍스트-이미지, 텍스트-키포인트 간의 일치도를 높이는 대칭 대비 손실(Lcon)과 최종 클래스 예측을 위한 분류 손실(Lcls)을 합산하여 최적화한다.

$$L = Lcon + Lcls \quad (1)$$

사전 학습된 지식을 보존하기 위해 Text Encoder 는 가중치를 고정하고, Image Feature Extractor 는 미세 조정만 수행하며, Keypoint Feature Extractor 와 분류 헤드는 처음부터 학습시키는 전략을 사용하여 학습 효율을 높였다.

표 1. 실험 결과

	TOP-1 Accuracy
Zero-shot	29.4%
Proposed (Fine-tune)	95.2%

실험을 위해 다양한 운전자와 주야간 조건, 그리고 승/하차 과정을 포함하는 자체 인캐빈 RGB 데이터셋을 구축하였다. 데이터셋은 '안전띠 착용/해제', '음료 섭취', '전화 통화' 등 9 가지 핵심 행동 클래스로 구성된다. 표 1 실험 결과, 제로샷 성능은 29.4%에 그쳤으나, 제안하는 방식으로 파인 튜닝(Fine-tune)을 진행했을 때 정확도가 95.2%로 대폭 향상되었다. 이는 탑승자 인지 프롬프팅이 복잡한 차내 환경에서 매우 효과적으로 작동함을 시사한다.

그림 2 결과 영상

Ⅲ. 결론

본 연구에서는 차량 실내 환경의 제약을 극복하기 위해 CLIP 기반의 Image Feature Extractor 와 Keypoint Feature Extractor 를 결합한 경량화된 행동 인식 모델을 제안하였다. 다양한 백본 적용이 가능한 유연한 구조 내에서, 본 연구는 CLIP 을 활용하여 제로샷 추론의 가능성을 탐색하고 파인 튜닝을 통해 95% 이상의 높은 정확도를 달성하였다. 특히 탑승자 인지 프롬프팅은 복잡한 차내 상황을 모델이 이해하는 데 핵심적인 역할을 수행하였다. 향후에는 야간 인식률 개선을 위해 IR(적외선) 센서 데이터를 결합한 멀티모달 연구와, 엣지 디바이스 탑재를 위한 모델 양자화(Quantization) 연구를 진행할 계획이다.

ACKNOWLEDGMENT

이 연구는 2025 년도 산업통상자원부 및 한국산업기술기획평가원(KEIT) 연구비 지원에 의한 연구임(과제번호: RS-2024-00506824, 과제명: 표준 인터페이스 기반 플릿 특화 개방형 제어기)

참 고 문 헌

[1] Martin, Manuel, et al. "Drive&act: A multi-modal dataset for fine-grained driver behavior recognition in autonomous vehicles." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019.

[2] Khanam, Rahima, and Muhammad Hussain. "Yolov11: An overview of the key architectural enhancements." arXiv preprint arXiv:2410.17725 (2024).

[2] Radford, Alec, et al. "Learning transferable visual models from natural language supervision." International conference on machine learning. PmLR, 2021.