최근 정보통신 기술
스마트홈에서 일상행동 인식을 위한 VLM-LLM 통합 프레임워크에 관한 연구

김종은, 윤동식
HDC 랩스

JongeunKim@hdc-labs.com, kevinds1106@hdc-labs.com

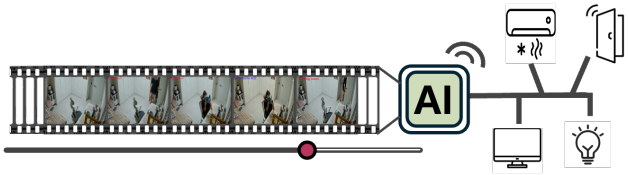
A Study on the Integrating VLM-LLM for ADL Recognition in Smart Home Environment

Kim Jong Eun, Yoon Dong Sik
HDC LABS

요 약

컨텍스트 인식형 스마트 홈 AI 시스템은 사용자의 필요를 사전에 지원하기 위해 신뢰성 있는 일상생활동작(ADL: Activities of Daily Living) 인식이 필수적이다. 이를 위해 본 연구에서는 40 개 활동 클래스를 포함한 가정 내 3 인칭 관찰 동영상 데이터셋을 구축하였다. 또한, Vision-Language Models(VLMs)와 Instruction-tuned LLMs 를 결합한 제로샷 프레임워크를 제안한다. 실험 결과, 이러한 결합은 ADL 인식 정확도를 유의하게 향상시키며, 스마트 홈 지능의 선제적 지원 가능성을 입증한다.

Ⅰ. 서 론



현대의 스마트홈은 조명, 에어컨, 도어락 등 다양한 IoT 기기를 통해 생활 편의성을 제공하지만, 여전히 사용자의 명시적인 조작에 의존한다는 한계가 있다. 진정한 스마트홈으로 발전하기 위해서는 사용자의 상태와 의도를 스스로 이해하고, 상황에 맞게 행동을 수행하는 자율형 AI 에이전트가 필요하다. 이를 위해서는 사용자의 일상행동(ADL)을 실시간으로 정확하게 인식하는 기술이 핵심이며, 특히 3 인칭 시점에서의 영상 분석은 전신 자세와 주변 맥락을 함께 파악할 수 있어 보다 정교한 행동 이해를 가능하게 한다.

본 연구에서는 이러한 목표를 달성하기 위해, 실제 가정 내 환경에서 다양한 일상행동을 촬영한 3 인칭 시점 기반 ADL 데이터셋을 구축하고, 이를 활용한 VLM- LLM 통합 프레임워크를 제안한다. 최근 VLM 은 복잡한 시각 장면을 이해할 수 있지만, 세밀한 지시(prompt)를 일관되게 따르지 못하는 한계가 있다[1]. 이에 따라 본 연구는 VLM 이 시각적 정보를 서술하고, LLM 이 이를 해석·추론하는 VLM(see)- LLM(reason) 구조를 적용하였다. 실험 결과, 명령 기반 학습을 거친 LLM 을 결합함으로써 제로샷 환경에서도 행동 인식의 정확도와 안정성이 크게 향상됨을 확인하였다.

Ⅱ. 본론

a. 데이터셋 구축

기존의 일상행동(Activities of Daily Living, ADL) 데이터셋은 주로 실내환경보다는 실외환경에서 촬영된 영상 또는 웹 크롤링을 통해 수집된 영상으로 구성되어 있다. 또한 일부 데이터셋은 공개 접근이 제한되어 있거나, 주석정보와 실제 영상 내용 간 불일치가 존재하는 문제를 포함하고 있다. 더불어, 다수의 기존 데이터셋은 3 인칭시점이 아닌 1 인칭시점의 영상만을 제공하고 있다[2]. 그러나 행동 인식에서 중요한 전신 자세와 주변 맥락 정보를 고려하기 위해서는 3 인칭 시점의 영상이 필수적이다. 이러한 한계를 극복하기 위하여, 본 연구에서는 실내 환경에서 다양한 일상행동을 포착한 3 인칭 시점 기반 ADL 데이터셋을 구축하였다. 데이터 다양성을 확보하기 위해 총 5 명의 피험자가 3~5 가지 서로 다른 복장으로 등장하였으며, 거실과

주방을 포함한 3 개의 방으로 구성된 실제 주거 공간에서 촬영을 진행하였다. 그 결과, 총 40 개 행동 클래스에 대해 430 개의 영상을 수집하였다.

b. 제로샷(zero-shot) ADL 인식 프레임워크

최근의 비전- 언어 모델(Vision- Language Model, VLM)은 복잡한 시각 장면에 대한 추론 능력을 보유하고 있으나, 복잡한 프롬프트에 일관성 있게 반응하지 못하는 한계가 존재한다. 이러한 문제를 보완하기 위하여, 본 연구에서는 행동인식 과제를 대상으로 VLM 과 LLM 을 결합한 통합 프레임워크를 제안한다. 이를 위해 제안하는 모델은 단순하지만 효과적인 구조로, 구축된 ADL 데이터셋 내 사용자 행동을 제로샷(Zero-shot) 환경에서 예측하도록 설계되었다.

$$\textbf{VLM+LLM}: \quad \hat{y} = \mathrm{LLM}(\mathrm{VLM}(v, p'), p) \qquad (1)$$

다음 식 (1)은 제안한 통합 프레임워크의 전체 과정을 수식으로 나타낸 것이다. 여기서 v 는 입력 영상을, p 는 행동 예측을 위한 프롬프트를, p'는 행동 설명을 위한 프롬프트를, $\hat{y}$ 는 최종적으로 예측된 주요 행동을 각각 의미한다. 제안한 프레임워크에서는 먼저 VLM 이 설명형 프롬프트(p') 를 기반으로 영상 내 행동을 서술적으로 기술하고, 이후 LLM 이 해당 설명과 예측 프롬프트(p) 를 입력받아 최종 행동을 판별한다. 이러한 방식은 시각적 정보에 대한 정밀한 비주얼 그라운딩(visual grounding) 과 LLM 의 추론 및 명령 이해 능력을 결합하여, 제로샷 환경에서의 ADL 인식 성능을 향상시킨다.

c. 실험

수집된 데이터셋과 YouHome [4] 벤치마크를 대상으로 다양한 VLM- LLM 조합을 평가하였다. 실험 결과, 대규모 LLM 일수록 성능이 향상되었으며, 특히 Gemma-3-27B-it 모델이 가장 높은 정확도를 보였다. 이는 모델 용량과 지시문 이해 능력이 ADL 인식 성능 향상에 중요한 역할을 함을 시사한다.

표1 제안된 ADL 데이터셋에서 모델 및 프롬프트 조합별 행동 인식 정확도

LLM 모델	prompt1	prompt2	prompt3
Qwen2.5-VL-7B-Instruct	0.5625	0.597	0.5625
+ Gemma-3-4B-it	0.5177	0.5524	0.5941
+ Gemma-3-12B-it	0.6072	0.608	0.594
+ Gemma-3-27B-it	0.6072	0.6142	0.5918
+ Llama-3.1-8B-Instruct	0.5979	0.608	0.456
VideoLLaMA3-7B	0.5902	0.557	0.510
+ Gemma-3-4B-it	0.5787	0.5501	0.439
+ Gemma-3-12B-it	0.6288	0.6157	0.5671
+ Gemma-3-27B-it	0.6419	0.6350	0.6161
+ Llama-3.1-8B-Instruct	0.6057	0.5709	0.5162

표 1 결과, 모델의 규모가 커질수록 성능이 향상되는 경향을 보였으며, 특히 Gemma-3-27B-it 모델이 모든 프롬프트와 VLM 조합에서 가장 높은 정확도를 기록하였다. 이는 대규모 LLM 의 용량이 복합 추론 능력 향상에 중요한 역할을 한다는 점을 시사한다. prompt3 의 결과는 prompt1 및 prompt2 보다 낮게 나타났는데, 이는 보다 복잡한 instruction 구조로 인해 모델의 일관된 응답이 어려웠기 때문으로 판단된다. Llama-3.1-8B-Instruct 모델에서 이러한 경향이 뚜렷하게 나타났으며, 반면 Gemma-3 계열 모델은 높은 지시 순응도(instruction adherence)를 보였다.

표2 Youhome 데이터셋에 LLM 통합 전후 행동인식 정확도 비교

LLM 모델	prompt1	prompt2	prompt3
Qwen2.5-VL-7B-Instruct	0.5522	0.5391	0.5574
+ Gemma-3-27B-it	0.6003	0.5941	0.5848
VideoLLaMA3-7B	0.5532	0.4809	0.4986
+ Gemma-3-27B-it	0.6172	0.6090	0.6111

추가적으로, 제안한 프레임워크의 일반화 성능을 검증하기 위해 YouHome ADL 데이터셋에서 추가 평가를 수행한 결과를 표 2 에 나타냈다. 표 2 는 LLM 통합 유무에 따른 비교 결과를 보여준다. 표 1 의 결과와 일관되게, Gemma-3-27B-it 모델이 모든 프롬프트에서 가장 높은 정확도를 기록하였다. 이 결과는 대규모 LLM 이 제로샷 환경에서도 행동 인식의 추론 능력을 강화한다는 점을 다시 한 번 확인시켜준다.

Ⅲ. 결론

본 연구는 스마트 홈 지능 강화를 위한 초기 단계로서, 3 인칭 관찰 ADL 데이터셋을 구축하고 VLM- LLM 결합 프레임워크를 제안하였다. 실험을 통해 본 접근 방식의 효과성을 검증하였으며, 이는 향후 스마트 홈에서의 자율적이고 선제적인 사용자 지원을 위한 핵심 기반이 될 수 있다.

참 고 문 헌

[1] Yang, Te, “Enhancing instruction-following capability of visual-language models by reducing image redundancy”, 2024.

[2] Cartas, Alejandro, “Recognizing activities of daily living from egocentric images”, Springer International Publishing, 2017.

[3] Ye, Hanrong, “MM-Ego: Towards Building Egocentric Multimodal LLMs for Video QA”, 2024.

[4] Pan, Junhao, “Youhome system and dataset: Making your home know you better”, IEEE, 2022.