대규모 언어 모델과 프롬프트 엔지니어링의 결합을 통한 로봇의 물체 인식 능력 향상

오혁진, 오현수, 김재호*
세종대학교

[hyukjin.sejong, hyeonsu.sejong]@gmail.com,*kimjh@sejong.ac.kr

Enhancing Robot Object Recognition Capabilities through the Integration of Multimodal Large Language Models and Prompt Engineering

Hyuk Jin Oh, Hyeon Su Oh, Jae Ho Kim*
Sejong University.

요 약

로봇 공학의 발전에 따라 사람과 로봇이 상호작용하는 휴먼-로봇 인터랙션에 관한 연구도 활발히 진행되고 있다. 하지만 사람의 명령에 따라 진행되는 상호작용의 특성상 자연어 형식으로 진행되는 명령의 복잡성과 다양한 지시 유형에 대한 처리 문제에 의해 로봇의 인지 능력 향상에 어려움을 겪고 있다. 본 논문에서는 대규모 언어 모델과 프롬프트 엔지니어링을 결합하여 로봇의 물체 인식 능력을 향상하는 시스템을 제안한다. 로봇 시뮬레이션 툴인 Isaac Sim 을 활용하여 제안한 시스템을 구현한다.

Ⅰ. 서 론

최근 로봇 공학이 발전함에 따라 휴먼-로봇 인터랙션에 관한 연구도 활발히 진행되고 있다. 휴먼-로봇 인터랙션은 사람과 로봇이 어떻게 상호작용하고 의사소통 하는지에 대한 연구 분야로, 교육, 재활 및 치료, 산업 자동화 등 많은 분야에서 응용된다[1]. 휴먼-로봇 인터랙션은 로봇이 사람의 의도를 인지하고 이해하는 인지 모델링에 대해 주로 연구가 진행되고 있다[2]. 이러한 연구의 주된 방향은 사람과 로봇 간의 상호작용이며, 이 상호작용은 사람의 명령에 의해 행해진다. 사람의 명령에는 자연어의 복잡성과 다양한 지시 유형에 대한 처리 문제가 존재하고, 사용자와 로봇이 위치해 있는 공간에 대한 이해가 반영되어야 하므로 사람의 의도를 정확히 파악하도록 로봇에게 명령을 내리는 것은 어렵다[3]. 이러한 문제를 해결하기 위해 본 논문에서는 사용자의 명령인 언어적 정보와 로봇의 시각적 데이터를 결합하여 맥락 정보를 모델링 하고, 이를 통해 사람과 로봇 간의 효과적인 상호작용을 위한 객체 인식 향상 시스템을 제안한다. 대규모 언어모델(M-LLM, Multimodal Large Language Model)을 이용하여 사용자가 자연어로 요청한 객체를 로봇이 선택할 수 있는 시스템을 개발한다. 사용자의 요청을 받아들이고, 로봇이 취득한 데이터를 분석하여 물체를 식별하는 기능을 구현한다.

Ⅱ. 본론

1. 시스템 구성

본 연구에서 전체적인 시스템의 구성은 그림 1과 같이 4개의 모듈로 구성된다. 가상 공간을 구현하기 위해 시뮬레이션과 로봇 제어 등의 기능을 제공하는 로봇 시뮬레이션 툴인 Isaac Sim을 사용한다. 또한 Isaac Sim의 논코딩 프로그래밍 기능인 OmniGraph와 그 노드들의 연결을 통해 이미지 처리 과정을 구현하고 ROS2(Robot Operating System 2) 플랫폼과 통신한다. ROS2는 로봇 프로그래밍 플랫폼으로, 노드 단위로 기능이 구현되며 토픽을 통해 통신이 이루어진다. Isaac Sim과 ROS2 모듈의 연결을 통해 이미지에 대한 전처리를 진행한다. 언어적

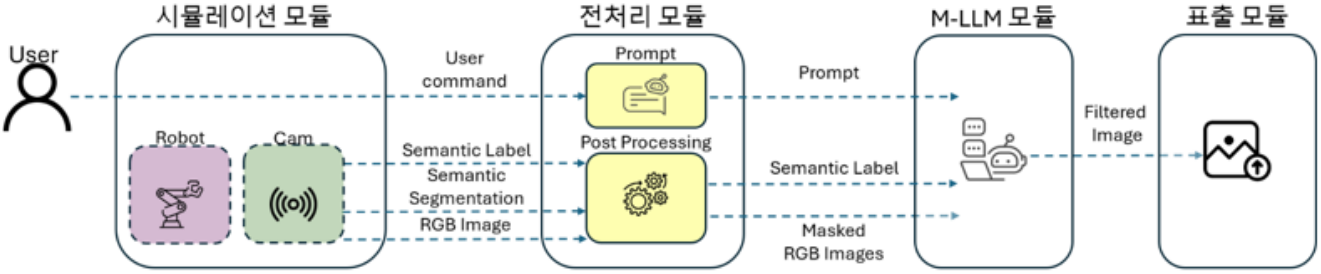

그림 1. 시스템 구성도

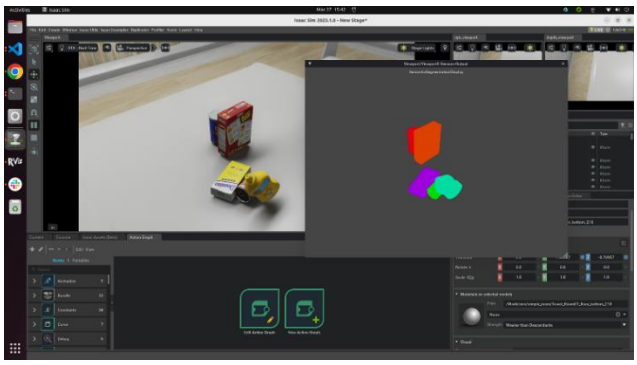

그림 2. 객체 식별 결과 화면

정보를 수집하기 위해 M-LLM 모듈과 프롬프트 엔지니어링을 결합하여 사용하고, 시각적 데이터를 수집하기 위해 시맨틱 세그멘테이션을 사용한다. 생성된 객체 별 마스킹 이미지와 라벨 데이터는 M-LLM 모듈로 전송하여 사용자 명령에 따라 선택된 이미지에 대한 시맨틱 라벨을 출력한다. 이를 위해 이미지 배열 인덱스를 토픽으로 반환한다. 그리고 출력된 시맨틱 라벨에 따라 이미지를 필터링하여 보여준다.

2. 시스템 시뮬레이션

Isaac Sim 으로 구현한 가상 공간에서 그림 2와 같이 3D 에셋을 불러온다. 로봇에 장착된 카메라에 접근하여 이미지 데이터를 생성하기 위해 Isaac Sim 의 OmniGraph 를 사용한다. OmniGraph 의 모듈을 연결하여 시뮬레이션 실행 시 카메라의 viewport 를 통해 입력 받은 이미지에 대해 처리하는 과정들을 구현한다. 이때, 사용자의 명령에 따라 로봇이 확인하고자 하는 물체에 대한 시맨틱 정보를 부여하여 시맨틱 세그멘테이션을 수행할 수 있도록 한다. 전처리를 위해 Isaac Sim 으로부터 발행된 RGB 이미지 토픽, 시맨틱 세그멘테이션 토픽, 시멘틱 라벨 토픽을 전처리 모듈에서 수신한다. 시맨틱 라벨과 시맨틱 세그멘테이션을 이용하여 바이너리 마스크를 만들고 이를 RGB 이미지에 적용하여 각 객체의 마스킹 이미지를 생성한다.

사용자 입력과 생성된 객체 별 마스킹 이미지, 시맨틱 라벨은 M-LLM 모듈로 이동하여 사용자가 요청한 이미지를 출력하도록 이미지 배열 인덱스를 토픽으로 반환한다. 마지막으로 표출 모듈에서 시맨틱 라벨을 기반으로 최종 출력될 이미지를 결정한다.

M-LLM 프롬프트는 세가지의 부분으로 구성한다. 첫째로, M-LLM 의 역할과 기대하는 기능을 정의하는 문장으로 첫 부분을 구성한다. 그리고 M-LLM 이 사용자로부터 입력을 받을 형식에 대한 예시와 M-LLM이 출력해야 할 예시를 정의한다. 다음으로 시맨틱 라벨과 그림3에 보이는 이미지와 같이 각 라벨에 해당하는 객체의 이미지의 배열 요소를 포함하도록 M-LLM을 구성한다. 마지막으로, 사용자의 명령을 자연어 입력으로 받아 사용자가 기대하는 이미지가 출력이 되도록 구성한다.

그림 3. 사용자 명령에 맞는 객체 이미지 출력

M-LLM 프롬프트를 구성한 뒤 사용자가 설탕이 들어간 물건을 집어 달라는 요청을 했을 때 그림2와 같이 해당하는 시맨틱 라벨을 출력한다. 그림3을 보면 출력된 시맨틱 라벨에 맞춰서 물건을 정확하게 인식하는 것을 확인할 수 있다.

Ⅲ. 결론

본 논문에서는 Isaac Sim Omniverse를 활용하여 가상 공간을 구축하고, M-LLM과 프롬프트 엔지니어링의 결합을 통해 로봇의 물체 인식 능력을 개선하기 위한 시스템을 제안한다. 가상 공간 내의 카메라 데이터와 시맨틱을 부여하여 생성된 합성 데이터를 통해 객체를 설정하고, 이를 인지하도록 프롬프트를 구성하여 프롬프트 엔지니어링을 진행하였다. 시스템 시뮬레이션을 통해 로봇의 물체 인식 능력이 향상됨을 확인하였다. 이 연구를 통해 기존 시스템의 한계를 극복하고, 보다 정교한 데이터 처리와 분석을 통한 기술발전을 실현하였다. 또한 향상된 물체 인식 능력은 후에 진행될 로봇 공학 분야의 연구에 많은 도움이 될 것이라고 기대한다.

향후 연구에서는 사용자 명령에 부합하지 않는 결과의 정확도를 개선하고, 객체의 이미지의 불완전한 출력 문제를 해결하고자 한다. 또한 향상된 로봇의 인지 능력을 바탕으로 물체를 잡아서 정확한 위치에 내려 놓는 작업(Pick and Place) 등의 복잡한 작업을 구현하고자 한다.

ACKNOWLEDGMENT

이 연구는 2024 년도 산업통상자원부 및 산업기술평가관리원 (KEIT) 연구비 지원에 의한 연구임 (RS-2022-00154678)

참 고 문 헌

[1] Murphy, R. R., Nomura, T., Billard, A., & Burke, J. L. (2010). An Exclusive Course for Computer Scientists and Engineers. IEEE Robotics & Automation Magazine, 23(2), 85-89.

[2] Kim, Y.C., Yoon, W.C., Kwon, H.T., Yoon, Y.S., Kim, H.J. (2007). A Cognitive Approach to Enhancing Human-Robot Interaction for Service Robots. In: Smith, M.J., Salvendy, G. (eds) Human Interface and the Management of Information. Methods, Techniques and Tools in Information Design. Human Interface 2007. Lecture Notes in Computer Science, vol 4557. Springer, Berlin, Heidelberg.

[3] Li, Z., Mu, Y., Sun, Z., Song, S., Su, J., & Zhang, J. (2021). Intention understanding in human–robot interaction based on visual-NLP semantics. Frontiers in Neurorobotics, 14, 610139.