장면 그래프 기반 공간 질의응답 시스템 설계 및 구현

여승연, 김유진, 이소현, 권용현, 김재호*

세종대학교

seungyeon.sejong@gmail.com, kimyj.sejong@gmail.com, sohyun.sejong@gmail.com,
yonghyun.sejong@gmail.com, *kimjh@sejong.ac.kr

Design and Implementation of Scene Graph-Based Spatial Question Answering System

Seungyeon Yeo, Yujin Kim, Sohyun Lee, Yonghyun Kwon, Jaeho Kim*

Sejong University

요 약

최근 지능형 에이전트의 활용이 확대됨에 따라, 공간적 맥락과 객체 간 관계를 이해하기 위한 수단으로 대규모 언어 모델의 활용이 주목받고 있다. 그러나 대규모 언어 모델에 2D 이미지나 단순 객체 검출 결과만을 입력으로 사용하는 경우 3차원 구조 및 공간 정보를 충분히 반영하는 데 한계가 존재한다. 이에 본 논문에서는 다중 시점에서 획득된 공간 정보를 공간 내 객체와 객체 간 관계로 명시적으로 표현하는 장면 그래프로 변환하고, 이를 기반으로 사용자의 자연어 질문에 대해 공간적 맥락이 반영된 답변을 생성하는 시스템을 제안한다.

Ⅰ. 서 론

최근 지능형 에이전트가 물리 환경에서 복잡한 작업을 수행함에 따라, 개별 객체의 인식을 넘어 공간 정보를 이해하는 공간 지능의 중요성이 대두되고 있다. 특히 사용자와의 자연스러운 상호작용을 위해서는 로봇이 공간의 구조를 이해하고 공간 내 객체의 위치 및 관계 변화를 파악하는 능력이 요구된다[1]. 이러한 흐름 속에서 대규모 언어 모델(Large Language Model, LLM)은 기존의 텍스트 중심의 처리 역할을 넘어, 시각 정보를 의미론적으로 해석하고 물리 환경 내 인과 관계를 추론하는 데 활용되고 있다[2]. 그러나 2D 이미지나 단순 객체 검출 결과만을 대규모 언어 모델의 입력으로 사용하는 경우 3차원 구조 및 공간 정보를 반영하는 데 한계가 있다[3].

이에 본 논문에서는 다중 카메라로부터 입력된 이미지 데이터를 전역 좌표계로 통합하여 공간 내 객체와 객체 간 관계로 표현하는 장면 그래프로 변환한다. 이를 기반으로 사용자의 자연어 질문에 대해 공간적 맥락이 반영된 답변을 생성하는 공간 질의응답 시스템을 제안한다. 제안하는 시스템은 oneM2M 기반 표준 IoT 플랫폼[4]을 기반으로 설계되었다. 장면 그래프 구축 과정에서는 객체 간 기하학적 연산과 의미론적 제약 조건을 적용하여 물리적으로 불가능한 관계를 제외하였다. 또한 이벤트 기반 장면 그래프 갱신 방식을 적용하여 공간 변화가 발생한 영역에 대해서만 그래프를 갱신함으로써 확장 가능한 공간 정보 관리 구조를 제공한다.

Ⅱ. 시스템 설계

1. 시스템 개요

그림 1은 본 시스템의 개요이다. 가상 환경의 카메라를 통해 발행되는 RGB 이미지와 깊이 정보로 인스턴스 분할 정보와 3차원 전역 좌표를 계산한다. 계산된 객체 정보는 표준 IoT 플랫폼으로 실시간 전송된다. 초기 수집된 객체 데이터를 기반으로 장면 그래프를 생성한 후, 객체의 위치 변화가 임계값을 초과하거나 객체의 생성 및 소멸이 감지되면 주기적으로

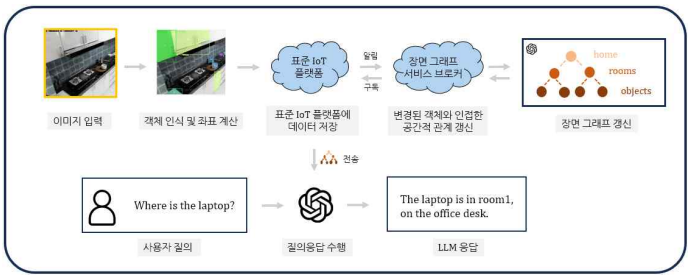

그림 1. 시스템 개요

그래프 갱신을 수행한다. 최종적으로 사용자의 자연어 질의가 입력되면, 시스템은 표준 IoT 플랫폼에 저장된 최신 장면 그래프를 참조하여 공간 정보를 분석하고 답변을 생성한다.

2. 시스템 세부 기능

2.1 객체 인식 및 공간 정보 통합

분산된 다중 카메라 환경에서 수집된 RGB 이미지를 기반으로 개방형 어휘 기반 객체 인식이 가능한 YOLOE[5] 모델을 활용하여 인스턴스 분할을 수행한다. 이후 헝가리안 알고리즘을 적용하여 다중 객체 추적을 통해 프레임 간 객체 ID의 일관성을 유지한다.

개별 카메라 좌표계에 인식된 객체 정보를 하나의 통합 공간에서 표현하기 위해 2차원 픽셀 좌표와 깊이 정보를 카메라 내부 파라미터와 결합하여 카메라 기준의 3차원 로컬 좌표를 계산한다. 이후 ROS2의 좌표 변환 프레임워크인 TF 시스템을 활용하여 계산된 로컬 좌표를 사전에 정의된 단일 전역 좌표계로 변환한다. 이를 통해 다중 시점에서 관측된 객체 위치를 일관된 기준으로 통합할 수 있다. 계산된 객체 위치 정보는 객체의 유클리드 거리 변화가 설정된 임계값을 초과한 경우에만 표준 IoT 플랫폼으로 전송하여 그래프 갱신에 필요한 정보만을 선별적으로 활용한다.

2.2 이벤트 기반 장면 그래프 구축

분산된 다중 카메라 환경에서 발생하는 객체 변화를 장면 그래프에 반영하기 위해 표준 IoT 플랫폼의 구독 및 알림 기능을 활용한다. 장면 그래프 서비스 브로커가 데이터 변화를 감지하면 노드로 반영하고 대규모 언어 모델을 활용하여 해당 노드를 중심으로 객체 간의 공간적 관계, 의미론적 관계를 추론한 뒤 이를 엣지로 구성한다. 이 과정에서는 변경된 객체와 연관된 관계만을 갱신하는 증분 업데이트 방식[6]을 적용한다.

2.3 공간 질의응답 시스템

구축된 장면 그래프를 기반으로 대규모 언어 모델을 활용하여 사용자의 자연어 질의에 대해 공간 정보가 반영된 질의응답을 수행한다. 대규모 언어 모델의 프롬프트는 그림 2와 같이 노드의 공간적 관계보다 의미론적 관계를 주요 판단 기준으로 활용하도록 설계되었다. 장면 그래프를 기반으로 객체의 위치 정보와 의미론적 관계가 입력으로 제공되며 이를 바탕으로 공간 정보를 고려한 추론이 수행된다. 이를 통해 모델은 갱신되는 물리 정보를 반영하여 공간에 대한 응답을 생성한다.

Ⅲ. 지능형 공간 질의응답 시스템 개발

실험 환경 구축을 위해 InteriorAgent 시뮬레이터의 kujiale_0004 데이터셋 환경[7]에 각 주요 구역별로 가상 카메라 1대를 배치하여 총 6대의 가상 카메라를 설치하였다. 이후 각 카메라에서 획득된 입력에 대해 YOLOE[5] 기반 인스턴스 분할을 수행하며, 다중 객체 환경에서도 안정적인 인식을 위해 신뢰도 임계값을 0.5로 설정하였다. 객체의 3차원 위치는 RGB 이미지와 깊이 정보를 결합하여 산출하며, 위치 변화가 임계값(δ > 20cm)을 초과하거나 객체의 생성 및 소멸이 발생한 경우에만 표준 IoT 플랫폼인 Mobius[8]에 저장 및 갱신된다. 장면 그래프 구축을 위해 객체를 노드로, 객체 간 관계를 엣지로 표현하며, 엣지는 공간적 관계(상, 하, 좌, 우), 의미론적 관계(대면, 포함), 거리 기반 인접 관계로 구성된다. 인접 관계는 객체 간 유클리드 거리가 1.5 m 미만인 경우로 정의했다. Mobius[8]는 방-객체 정보와 장면 그래프의 노드 및 엣지로 구성된 리소스 구조를 통해 공간 정보를 계층적으로 관리하며, 변화가 발생한 영역에 한해 장면 그래프를 증분 업데이트 방식으로 갱신한다. 사용자의 자연어 질의가 입력되면, 그림 3과 같이 시스템은 Mobius[8]에 저장된 최신 장면 그래프를 기반으로 GPT-4o-mini를 활용하여 의미론적 관계를 반영한 응답을 생성한다.

장면 그래프 기반 공간 질의응답

```
1: 당신은 실내 환경에서 동작하는 지능형 어시스턴트이다.
2: 장면 그래프로 표현된 주거 환경 정보를 기반으로 사용자의 질문에 답하라.
3: 추론 과정에서는 객체 간 관계 정보를 단계적으로 고려한다.

4: 실내 환경에서 관측된 객체 목록은 다음과 같다:
5: OBJECT_LIST = [OBJECT_LIST]

7: [추론 규칙]
8: 객체 간의 의미론적 관계 정보가 존재하는 경우 이를 우선적으로 활용한다.
9: 관계가 역방향으로 주어질 경우 이를 보완적으로 추론한다.
10: 공간 좌표 정보는 보조적인 근거로만 사용한다.
11: 모든 답변에는 객체가 속한 방 정보를 포함한다.

12: [출력]
13: 공간적 맥락이 반영된 자연어 답변만을 출력한다.

---

OBJECT_LIST: 환경 내 객체 목록
(라벨, 속성, X좌표, Y좌표, z좌표, 방 정보, 엣지 관계)
```

그림 2. 장면 그래프 기반 공간 질의응답 프롬프트

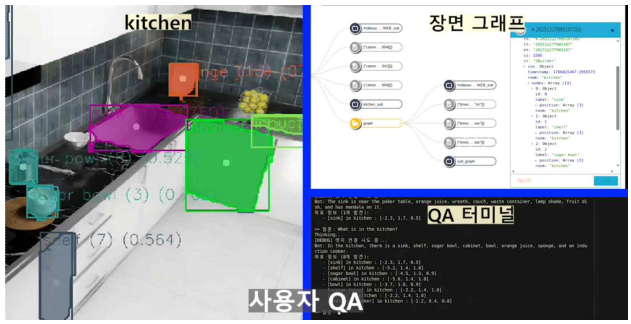

그림 3. 장면 그래프 기반 공간 질의응답 시스템 실행 화면 (https://youtu.be/FC7_Y7GURmk)

Ⅳ. 결론

본 논문에서는 분산된 다중 카메라 환경에서 획득되는 시각 정보를 표준 IoT 플랫폼에 저장 및 관리하고, 이를 장면 그래프로 구조화하여 대규모 언어 모델과 결합함으로써 실내 공간의 맥락을 이해하고 추론할 수 있는 공간 질의응답 시스템을 제안하였다. 제안한 시스템은 다중 시점의 공간 정보를 전역 좌표계로 통합하고, 이벤트 기반 장면 그래프 갱신을 통해 확장 가능한 공간 정보 관리 구조를 제공한다. 장면 그래프에 기반한 의미론적 관계 추론을 질의응답 과정에 활용함으로써 물리 환경의 맥락이 반영된 응답을 생성할 수 있음을 보인다. 본 시스템은 로봇 및 스마트홈 환경에서 공간 인지와 효율적인 데이터 관리를 동시에 고려한 지능형 시스템 설계의 가능성을 제시하였다.

ACKNOWLEDGMENT

본 연구는 과학기술정보통신부의 재원으로 한국연구재단, 무인이동체원천기술개발사업단의 지원을 받아 무인이동체원천기술개발사업(RS-2020-NR117734)을 통해 수행되었으며, 산업통상자원부 및 산업기술평가관리원(KEIT) 연구비 지원에 의한 연구(RS-2022-00154678)이며, 과학기술정보통신부 및 정보통신기획평가원의 정보통신방송혁신인재양성(메타버스융합대학원)사업 연구 결과로 수행되었음 (IITP-2026-RS-2023-00254529)

참 고 문 헌

[1] I. Armeni et al., "3D scene graph: A structure for unified semantics, 3D space, and camera," in ICCV, 2019.

[2] Driess, D., et al. "Palm-e: An embodied multimodal language model." ICML, 2023.

[3] Johnson, J., et al. "Image Retrieval using Scene Graphs." CVPR. 2015.

[4] J. Kim, et al., “Standard-based IoT platforms interworking: Implementation, experiences, and lessons learned,” in IEEE Commun. Mag., vol. 54, no. 7, pp. 48-54, Jul. 2016.

[5] Ao Wang., et al. “YOLOE: Real-Time Seeing Anything” ICCV. 2025.

[6] S. Wu et al., "SceneGraphFusion: Incremental 3D scene graph generation from RGB-D frames," CVPR, 2021.

[7] H. Fu et al., "3D-FRONT: 3D Furnished Rooms with LayOuts and New Terminology," IEEE TPAMI, 2021.

[8] Mobius, Retrieved Dec. 29, 2025, from https://github.com/IoTKETI/Mobius.