에지 로봇 내비게이션을 위한 실시간 이미지 임베딩 지도 작성

Building Real-time Image Embedded Map for Edge Robot Navigation

○윤 준 영[1], 김 필 은[2*]
[1)] 한국항공대학교 인공지능학과 (TEL: 02-300-0173; E-mail: a_o@kau.kr)
[2)] 한국항공대학교 AI 자율주행시스템공학과 (TEL: 02-300-0258; E-mail: pkim@kau.ac.kr)

Abstract As the scope of robotic applications expands, there is a growing demand for maps that possess general situational interpretation. We present CMAP, a mapping framework that incorporates the spatial meaning of specific locations. CMAP utilizes pre-trained VLM to generate maps and enables robots to search at general knowledge level through minimal computational power with language inputs. Also, creates viewpoints that allow users to visually recognize location semantics. Experiment demonstrates successful generation of appropriate goals at a rate of 58.7%, even with ambiguous keywords. This capability of mapping via general knowledge can significantly aid various robotic activities particularly with LLMs.

Keywords Robot Navigation, Real-time Mapping, CLIP, SLAM, VLM

1. 서론

로봇의 활용 범위가 넓어지면서 에지 로봇에서 수행하여야 하는 작업이 다양해졌다. 하지만 일반적인 센서(LiDAR, Camera)로 작성된 지도는 의미가 있는 공간 정보를 포함하지 않는다. 언어, 특히 LLM을 통해 로봇을 제어하기 위해 구체적이고 일반적인 공간 정보가 포함된 지도가 요구된다. 기존에도 Visual Language를 통해 지도를 작성[1][2]하는 시도가 있었으나, 에지 로봇에서 사용하기엔 큰 컴퓨팅 파워가 필요[1]하거나, 실제 로봇이 아닌 시뮬레이션 환경에서만 실험[2]되었다. 본 연구에서는 1. 적은 컴퓨팅 파워로 에지 로봇에서 사용할 수 있고, 2. 일반적인(general) 공간 정보를 담고 있는, 3. 공간 정보를 사용자가 직관적으로 판단할 수 있는 지도를 작성하는 프레임워크인 CMAP을 제시한다.

2. CMAP: 실시간 이미지 임베딩 지도 작성

2.1 시스템 구조

CMAP은 그림 1과 같이, 입력된 이미지와 2D LiDAR SLAM을 통해 얻은 자세를 토대로 키프레임(keyframe)인지 판별한 후, CLIP[3]을 이용하여 키프레임 이미지 임베딩과 자세를 저장한다. 이미지 임베딩은 PCA[4]를 이용하여 3차원으로 축소하여 시각화한다. 목표 문자가 입력되면 텍스트 임베딩과 저장된 이미지 임베딩의 유사도를 분석하여 가장 유사한 지점을 출력한다.

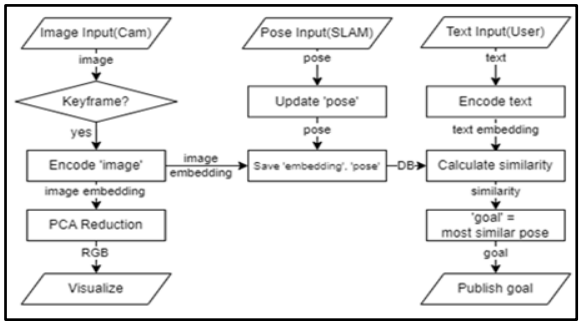

그림 1. Flowchart of CMAP.

2.2 CLIP[3] 이미지 임베딩

CLIP은 인터넷 규모로 사전 학습된 이미지-텍스트 인코딩 모델로, 이미지와 텍스트 임베딩을 동일한 차원 공간으로 매핑하도록 설계되었다. 이를 통해, 이미지의 공간 정보를 텍스트와 비교, 연산할 수 있으면서도, 이미지보다 더 적은 저장 공간을 사용할 수 있다. 따라서 로봇이 탐색한 공간을 효율적으로 분석하고 내비게이션하는 데 있어 큰 이점을 가진다

2.3 가중 맵 기반 키프레임(keyframe) 선정

상대적으로 컴퓨팅 파워가 낮은 에지 로봇을 이용할 때, 실시간 처리를 위해 관측 정보 기반 키프레임 선정 알고리즘이 필요하다. 본 연구에서는 거리 기반 가중치를 적용하여 새로운 프레임과 이전 키프레임의 관측값

※ 본 연구는 2024 년도 정부(교육부)의 재원으로 한국 연구재단의 지원을 받아 수행된 기초연구사업임(NRF-2022R1F1A1073799)

차이 k를 이용하여 키프레임을 선정하였다.

$$f_o(p) = w(d_{o,p}) = d^{0.5}e^{-0.25d+0.15} \quad (1)$$

$$k_{t=kf}: \ f_t(P_t) - f_{kf-1}(P_t \cap P_{kf-1}) > k_{th} \quad (2)$$

2.4 차원 축소를 통한 RGB 뷰포인트 생성

PCA[4]로 3차원으로 축소한 뒤 RGB에 대응하여 해당 공간이 어떤 의미를 가지는지 대략적으로 사용자가 인지할 수 있다. 다양한 환경으로 이루어진 데이터셋으로 피팅된 프로파일을 생성하여 변환하였다.

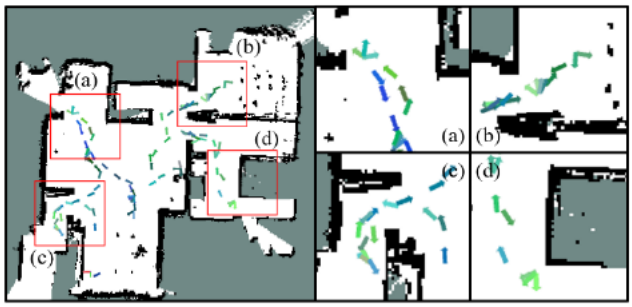

그림 2. Viewpoint generation of CMAP

3. 실험

3.1 실험 환경

사용자가 가정 환경에서 로봇을 조종하여 지도를 생성한 뒤, 입력 문자로 해당하는 곳이 보이는 목표 지점을 생성하면 성공이다. 실험 환경에 존재하는 '화장실', '거실' 등 장소(location)와 'TV', '냉장고' 등 물체(object), '씻을 수 있는 곳' 등 설명하는(description) 구문으로 이루어진 100개의 입력 문자(keyword)로 실험하였다. Nvidia GeForce MX450이 탑재된 Turtlebot4로 진행하였다.

3.2 실험 결과

표 1. Experiment of CMAP in real world.

	Simple selection			Weighted selection		
	1/60	1/30	1/10	k=1k	k=0.1k	k=0
keyframes ↓	127.0	243.0	760.7	175.3	364.0	436.0
Accuracy ↑	51.0	54.0	58.0	56.0	56.3	58.7

실험 결과 단순한 키프레임 선정으로는 저장되는 키프레임에 따라 51%에서 58%의 정확도를 보여주었으며, 가중 맵 기반 키프레임 선정을 이용하였을 때에는 k_{th} 값에 따라 비슷한 키프레임을 저장한 단순 선정 알고리즘보다 좋은 정확도를 가졌다. 특히 k_{th}=0으로 설정하여도 움직이지 않을 때의 프레임을 걸러주어 효율적이었다. 그림 2를 보면 (a)주방, (b)식당, (c,d)침실에 대해 FOV에 따라 다른 색의 뷰포인트가 표시된 것을 확인할 수 있다. 또한, 그림 3을 보면 같은 화장실이어도 'bathtub' 명령으로 욕조가 있는 식당 근처의 화장실로 목표 지점을 설정하여 정확한 내비게이션이 가능했다.

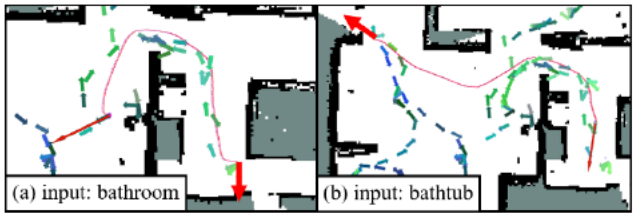

그림 3. Goal point generation of CMAP

4. 결론

LiDAR나 카메라로만 제작한 지도는 위치를 특정하는 데에는 유용하나 의미 있는 공간 정보를 얻기가 힘들다. 로봇의 활용 범위가 넓어지면서 일반적이고 범용적인 공간 해석이 가능한 지도 작성이 요구된다. 본 논문에서는 해당 위치가 가지고 있는 공간적 의미를 포함한 지도 작성 프레임워크인 CMAP을 제시한다. CMAP은 VLM을 이용하여 제작되는 지도로, 작은 컴퓨팅 파워로 키워드 입력을 통해 일반적인 지식 수준에 해당하는 위치를 탐색할 수 있으며, 뷰포인트 생성으로 사용자가 시각적으로 인지가 가능하다. 실제 환경에서도 58.7%로 적합한 목표 지점 생성에 성공하였다. 추후 연구에서는 여러 대의 카메라로 FOV를 확장하거나 탐색 알고리즘을 추가하여 더 정확도를 높힐 수 있을 것으로 기대된다. CMAP으로 제작된 지도는 언어, 특히 LLM을 통해 내비게이션 할 수 있으며, 다양한 로봇 활동에 도움이 될 것이다.

참고문헌

[1] C. Huang, O. Mees, A. Zeng, and W. Burgard, "Visual language maps for robot navigation," *2023 IEEE International Conference on Robotics and Automation*, pp. 10608-10615, May, 2023.

[2] S. Y. Gadre, M. Wortsman, G. Ilharco, L. Schmidt, and S. Song, "Cows on pasture: Baselines and benchmarks for language-driven zero-shot object navigation," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23171-23181, June, 2023.

[3] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, ..., and I. Sutskever, "Learning Transferable Visual Models From Natural Language Supervision," *arXiv preprint arXiv:2103.00020,* 2021.

[4] H. Abdi, and L. J. Williams, "Principal component analysis," *Wiley interdisciplinary reviews: computational statistics*, vol. 2(4), pp. 433-459, 2010