트랜스포머 기반 카메라-라이다 융합 서빙로봇 위치재추정

이지은, 오지용, *이학준
한국전자통신연구원 대경권연구센터, *Polaris3D

jieun.lee@etri.re.kr, jiyongoh@etri.re.kr, *hakjunlee@polaris3d.co

Transformer-Based Camera-LiDAR Fusion for Serving Robot Relocalization

Lee Jieun, Oh Jiyong, *Lee Hakjun
Electronics and Telecommunications Research Institute, *Polaris3D

요 약

본 논문은 트랜스포머 기반 카메라-라이다 융합을 활용한 서빙로봇의 종단간 위치재추정 연구를 제안한다. 모델 구조는 기존 트랜스포머와 동일하며, 두 센서 특징들은 각각 linear projection 을 통과한 다음, positional embedding 과 함께 트랜스포머 인코더에 입력되어 최종적으로 서빙로봇의 위치와 방향이 출력된다. 상용 서빙로봇을 활용해 수집한 데이터를 이용하여 비전 트랜스포머 위치재추정과 CNN 기반 카메라-라이다 융합 위치재추정과의 위치 오차 및 방향 오차를 비교한다. 실험을 수행한 결과, 트랜스포머 기반 카메라-라이다 융합 위치재추정이 더 작은 오차값들을 보이는 것을 확인하였다.

Ⅰ. 서 론

본 논문은 서빙로봇이 작성된 지도 상에서 자신의 위치를 잃어버리는 경우에 센서로 입력되는 정보들을 이용하여 위치를 알아내기 위한 위치재추정 연구를 수행한다. 종단간(end-to-end) 위치재추정은 전통적인 방법들에 비해 조명이나 움직임 변화에 강인한 장점을 가지는 PoseNet[1]을 기점으로 연구가 활발히 진행되고 있다. PoseNet 과 같이 카메라 영상으로부터 위치를 재추정하는 연구들[2], [3]과 3D 라이다 데이터를 사용하여 위치재추정의 정확도를 향상시킨 연구[4], IMU 센서만을 이용한 위치재추정 연구[5] 등 다양한 방법들이 제시되고 있다. 최근 컴퓨터 비전 분야에서 트랜스포머[6]를 활용한 많은 연구가 진행되고 있는데 그 중 대표적으로 카메라 영상을 이용하는 비전 트랜스포머(Vision transformer, ViT)[7]와 3D 포인트 클라우드를 이용하는 Point-MAE[8]가 있다.

본 논문에서는 트랜스포머를 기반으로 카메라-라이다 융합 위치재추정 연구를 제안하고자 한다. 트랜스포머에서 영상과 2D 라이다 포인트들이 위치를 추정하기 위한 상호보완적인 특징들을 선택할 수 있을 것이라는 가정하에 연구를 수행한다. 본 연구의 효용성을 알아보기 위해 상용 서빙로봇에서 수집한 데이터를 활용하여 비전 트랜스포머 모델 및 CNN 기반 카메라-라이다 융합 모델(FusionLoc)[9]과의 위치재추정 성능을 비교한다.

Ⅱ. 본론

본 연구는 그림 1 과 같이 기존 트랜스포머와 동일한 구조의 모델을 활용하며 트랜스포머 인코더에 카메라 영상 데이터 특징과 2D 라이다 데이터 특징, positional embedding 을 함께 입력한다. 각 데이터 특징은 linear projection 을 통과한 것이다. 트랜스포머 인코더를 거친 특징들은 MLP 레이어들을 사용하여 최종적으로 작성된 지도 상에서의 2 차원 위치와 2 차원 방향을 추정한다.

서빙로봇의 위치재추정 실험을 위해 Polaris3D 사의 상용 서빙로봇을 이용하여 Realsense D435 의 영상 데이터와 RPLIDAR S1 의 2D 라이다 데이터를 수집하였다. 서빙로봇의 위치와 방향은 SLAM 기술을 적용하여 함께 수집하였다. 수집된 데이터는 10 개의 시퀀스로, 총 3,964 개의 영상과 라이다 데이터들로 구성되어 있다. 이 중 7 개의 시퀀스 (2,766 개)를 학습 데이터로 사용하였으며, 3 개의 시퀀스 (1,198 개)를 테스트 데이터로 사용하였다. 입력 영상의 크기는 비전 트랜스포머에서는 미리 학습된 모델을 활용하기 위해 224x224x3 으로, CNN 및 트랜스포머 기반 융합 모델에서는 256x256x3 으로 조정하였다. 2D 라이다 포인트들은 매번 들어오는 포인트들의 개수가 다르기 때문에 850 개의 포인트들만 랜덤으로 선택하였다. 트랜스포머 모델의 embed 차원은 768, 영상 패치 수는 16 으로 맞춰주었고, 2D 라이다 포인트들의 그룹화 설정값은 그룹 크기 32 개, 그룹 개수 64 개로 정하였다. Multi-head self-attention 은 미리 학습된 비전 트랜스포머 모델의 경우 12 개의 head 와 12 개의 layer 를 가지며, CNN 및 트랜스포머 기반 융합 모델의 경우 6 개의 head 와 4 개의 layer 로 설정하였다.

표 1 에는 각 모델에 대한 실험 결과를 나타내었다. 표 1 을 통해 확인할 수 있듯이, 영상 데이터만을 이용하는 모델보다는 영상 데이터와 라이다 데이터를 함께 활용하는 모델들이 서빙로봇의 위치를 더 정확하게 추정하였다. 또한 트랜스포머 기반 융합 모델이 다른 모델들과 비교하여 가장 작은 오차값을 보였으며, 평균 위치 오차 및 평균 방향 오차가 비전 트랜스포머 모델보다 0.57 m, 33.9°, CNN 기반 융합 모델보다 0.08 m, 3.71° 더 작은 것을 확인하였다.

Ⅲ. 결론

본 논문에서는 서빙로봇의 종단간 위치재추정 연구를 위해 트랜스포머 기반 카메라-라이다 융합 모델을 활용하였다. 상용 서빙로봇에서 수집한 데이터를 사용하여 실험을 수행하였으며, 트랜스포머 기반 카메라-라이다 데이터 융합 모델의 위치재추정 성능을 카메라 영상 데이터만을 입력 받는 비전 트랜스포머 모델과 CNN 을 적용한 카메라-라이다 데이터 융합 모델과 비교하여 제안한 모델의 위치재추정 오차가 가장 작은 것을 확인하였다. 추후에는 다양한 장소와 상황을 포함하는 데이터에 대해 트랜스포머 기반 융합 모델에 대한 실험적인 검증이 필요한 것으로 판단된다.

ACKNOWLEDGMENT

본 논문은 한국전자통신연구원 연구운영지원사업의 일환으로 수행되었음[24ZD1110, 대경권 지역산업 기반 ICT 융합기술 고도화 지원사업].

참 고 문 헌

[1] A. Kendall, M. Grimes, and R. Cipolla, “PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization,” 2015 IEEE International Conference on Computer Vision, pp. 2938-2946.

Davies R. W." The Data Encryption standard in perspective,"Computer Security and the Data Encryption Standard, pp. 129-132.

[2] S. Brahmbhatt et al., “Geometry Aware Learning of Maps for Camera Localization,” 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition, pp. 2616-2625.

[3] B. Wang et al., “AtLoc: Attention Guided Camera, Localization,” in Proc. AAAI Conf., vol. 34, no. 06, pp. 10393-10401, Apr. 2020.

[4] W. Wang et al., “PointLoc: Deep Pose Regressor for LiDAR Point Cloud Localization,” IEEE Sensors Journal, vol. 22, issue. 1, pp. 959-968, Jan. 2022.

[5] S. Sun, D. Melamed, and K. Kitani, ‘‘IDOL: Inertial deep orientationestimation and localization,’’ in Proc. AAAI Conf., vol. 35, no. 7, pp. 6128- 6137, May 2021.

[6] A. Vaswani et al., ‘‘Attention is all you need,’’ in Advances in Neural Information Processing Systems, vol. 30, 2017.

[7] A. Dosovitskiy et al., “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” International Conference on Learning Representations. 2021.

[8] P. Yatian et al., “Masked Autoencoders for Point Cloud Self-supervised Learning,” in Proc. European Conf. Computer Vision, Oct 2022.

[9] J. Lee, H. Lee, and J, Oh, “FusionLoc: Camera-2D LiDAR Fusion Using Multi-Head Self-Attention for End-to-End Serving Robot Relocalization,” IEEE Access, vo. 11, pp. 75121-75133, 2023.

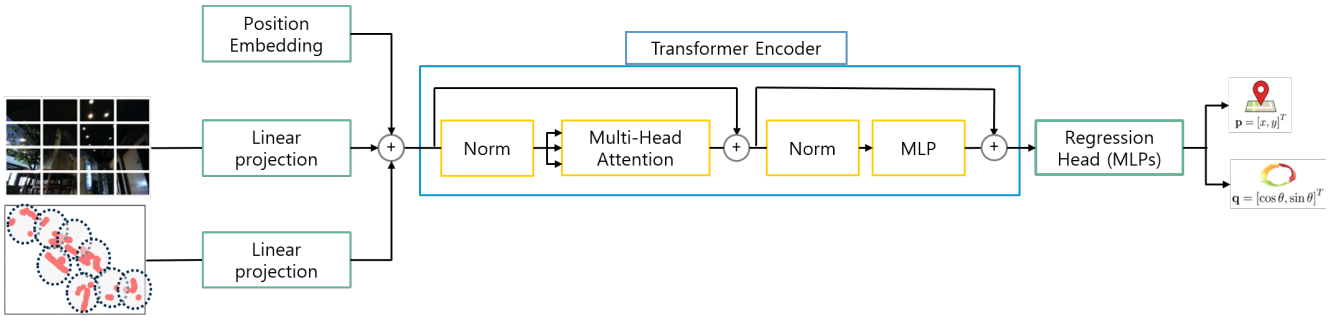

그림 1. 트랜스포머 기반 카메라-라이다 융합 모델 구조

표 1. 각 모델의 실험 결과 (평균 위치 오차 및 평균 방향 오차)

모델 / 평균 오차	비전 트랜스포머	CNN 기반 융합	트랜스포머 기반 융합
위치, 방향	1.56 m, 41.69°	1.07 m, 11.50°	0.99 m, 7.79°