3rd Place Solution to NAVER LABS Mapping & Localization Challenge 2020: Outdoor Track

김지원*, 김태주*, 황유진*, 최유경†
세종대학교 Robotics Computer Vision (RCV) 연구실
{jwkim, tjkim, yjhwang, ykchoi}@rcv.sejong.ac.kr

요 약

본 논문은 NABER LABS 주관 MAPPING & LOCALIZATION CHALLENGE 2020 에서 실외트랙 3 등을 수상한 SejongRCV 팀의 방법론을 소개한다. 테스트 영상(Query) 속 차량의 자세 및 위치 추정을 위해 우선 Retrieval 기반의 Place recognition 을 수행하여 DB 에서 테스트 영상과 유사한 영상 후보군을 선별하고, 지역 특징 매칭(Local Feature Matching)기반 Re-ranking 을 통해 가장 유사한 영상을 선정한다. 최종적으로 두 영상간 매칭된 지역 특징 쌍과 대응되는 3 차원 라이다 포인트의 관계를 RASAC PnP(Perspective-n-Point)로 풀어 카메라 위치 및 자세를 구하고 이를 이용해 차량의 위치 및 자세를 추정하였다. 추가적으로 정확도 향상을 위해 테스트 케이스 중 지역 특징 쌍이 가장 많은 영상의 자세 및 위치를 찾고, Visual Odometry 를 사용해 테스트 케이스의 마지막 영상 속 차량의 자세를 추정하여 챌린지에서 3 등을 수상할 수 있었다. 챌린지에서 사용한 코드는 SejongRCV Github (https://github.com/sejong-rcv/SejongRCV-Outdoor)에서 확인할 수 있다.

1. 서론

자율주행 로봇 및 자동차의 연구에서 정밀한 위치 인식 기술은 필수적이다. 대부분의 경우 GPS 를 이용해 정밀한 위치인식을 수행하지만, 건물의 내부나 빌딩 숲과 같은 환경에서 GPS 는 무용지물이 된다. 따라서 GPS 음영지역에서도 강인한 위치인식을 수행하기 위한 많은 연구가 이뤄졌고, 대표적으로는 Visual Localization 기술이 있다. Visual Localization 기술은 영상 정보로 카메라의 위치와 자세를 추정하는 기술로, 최종적으로는 카메라의 위치와 자세를 이용해 자동차 및 로봇의 자세 및 위치를 추정할 수 있다. 이러한 이유로 Visual Localization 은 자율주행 분야에서 그 기술적 요구가 꾸준히 증가하는 추세이다.

이러한 기술적 요구에 부응하여 NAVER LABS 는 국내 대학에서의 Visual Localization 기반 MAPPING &LOCALIZATION 연구 활성화를 위해 MAPPING& LOCALIZATION CHALLENGE 2020 을 개최하였다. 해당 챌린지에서는 NAVER LABS 에서 구축한 Visual Localization 을 위한 실내외 데이터셋이 공개됐으며, 실내외 트랙 총 187 팀이 참가하였다.

본 논문에서는 실외 트랙에서 3 위를 수상한 SejongRCV 팀의 방법론을 소개한다. Sejong RCV 팀의 테스트 케이스의 마지막 영상 속 차량의 자세 및 위치(6DOF) 추정을 위한 파이프 라인은 그림 1 과 같다.

2. 방법론

2.1 Retrieval based Place Recognition

Image Retrieval 기반의 Place Recognition 은 각각의 영상을 전역 기술자(Global descriptor)로 기술하고, 전역 기술자 간 유사도를 계산해 테스트 영상(쿼리 영상)과 가장 유사도가 높은 지도 내 영상(DB 내 영상)을 선별하는 방법이다.

본 팀은 단일 영상을 전역 기술자로 표현하기 위해 NetVLAD[1]를 사용하였다. NetVLAD 는 Relja Arandjelovic 가 제안한 CNN 기반의 Image Retrieval 네트워크로, Visual appearance 변화에도 강인함이 입증돼 많이 사용된다. 따라서 본 팀도 NetVLAD 를 통해 모든 학습 영상의 전역 기술자를 추출하여 DB 를 구축하였다.

평가 과정에서는 NetVLAD 를 통해 테스트 영상의 전역 기술자를 기술하고, 최근접 이웃 탐색 알고리즘(k-NN search)을 수행하여 DB 에서 테스트 영상과 전역기술자의 유사도(L2 similarity)가 높은 10 개의 후보군 영상(top-10)을 선별하였다.

이후 후보군 영상에 대해 Re-ranking 과정을 거쳐 테스트 영상과 영상 기하학적으로 가장 유사도가 높은 영상(top-1)을 선별한다. Re-ranking 에는 Daniel DeTone 이 제안한 딥러닝 기반 지역 특징 추출 네트워크인 Super Point[2]와 Paul-Edouarad Sarlin 이 제안한 딥러닝 기반 지역 특징 매칭 네트워크인 Super-Glue[3]를 사용하였다.

* 은 공동 저자를 의미하며, † 는 교신 저자를 의미함

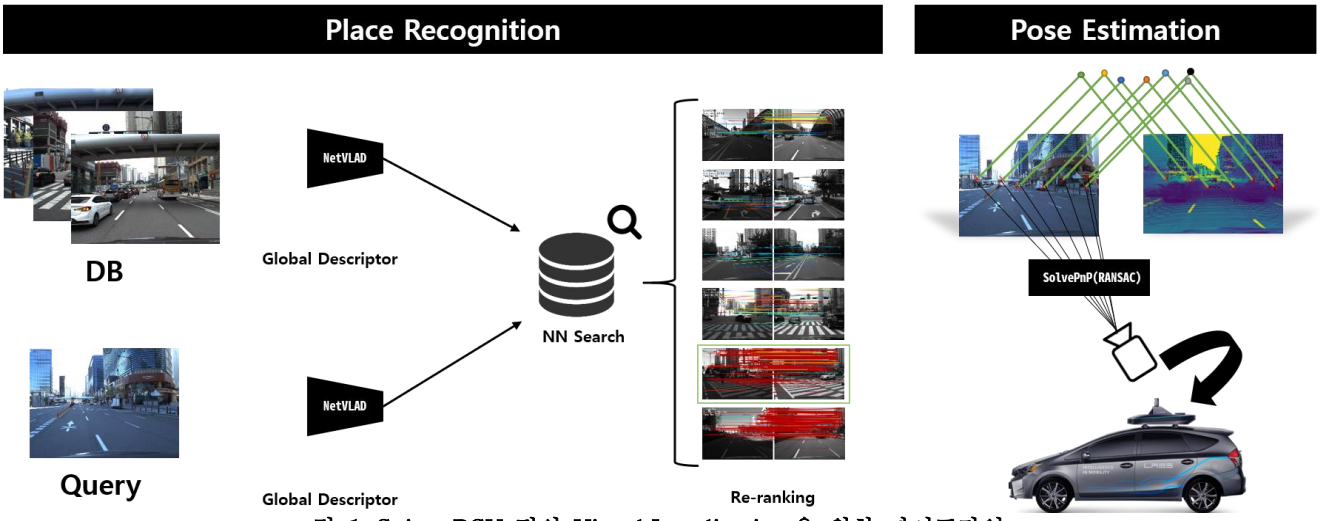

그림 1. SejongRCV 팀의 Visual Localization 을 위한 파이프라인

2.2 Pose Estimation

Place Recognition 과정을 통해 테스트 영상과 가장 유사도가 높은 DB 내 레퍼런스 영상을 선정하였다. Pose Estimation 단계에서는 테스트 영상과 레퍼런스 영상의 지역 특징 매칭 그리고 레퍼런스 영상이 촬영된 시점에 수집한 3 차원 라이다 포인트를 이용해 테스트 영상에서의 차량의 위치(X,Y,Z)와 자세(qw,qx,qy,qz)를 추정한다.

테스트 영상의 위치 및 자세 추정 과정은 다음과 같다. 첫번째, 라이다와 카메라의 관계를 이용해 3 차원 라이다 포인터를 레퍼런스 영상에 투영(Projection)한다. 두번째, 레퍼런스 영상에 투영된 3 차원 라이다 포인트와 대응되는 테스트 영상과 매칭된 지역 특징점(u,v)에 구한다. 세번째, 테스트 영상에서 매칭된 지역 특징점과 이에 대응되는 3 차원 라이다 포인트를 함께 활용하여 RANSAC PnP 의 해를 계산하며, 최종적으로 이를 통해 차량의 자세 및 위치를 추정한다. 이때, RANSAC PnP 의 해를 구하는 과정에는 Tong Ke 가 제안한 AP3P[4]방법을 사용하였다.

2.3 Visual Odometry

테스트 영상과 DB 내 레퍼런스 영상이 실제 유사하다면 정확한 자세 및 위치 추정이 가능하다. 하지만 테스트 영상과 일치하는 영상이 DB 에 없다면 자세 추정을 위해 충분한 지역 특징 매칭이 이뤄지지 않거나, 오매칭이 일어나 잘못된 자세를 추정하는 경우가 발생한다.

본 팀은 이러한 상황에서의 정확도를 높이기 위해 Stereo Odometry 알고리즘을 통해 추정한 Visual Odometry 정보를 사용하였다. 이때 Visual Odometry 란 각 이웃한 영상 프레임간의 위치 변화 추정을 의미한다. 챌린지에서는 실제 자세를 추정해야 하는 테스트 영상 외에 테스트 영상이 촬영되기 이전 49 장의 시퀀스 영상을 함께 제공한다. 따라서 이전 영상들에 대해서도 Place Recognition 을 동일하게 수행하여 50 장의 시퀀스 영상 중 매칭된 지역 특징 쌍이 가장 많은 테스트-레퍼런스 영상 쌍을 선택해 자세를 추정하였다. 최종 매칭이 일어난 영상이 테스트 케이스의 마지막 영상이라면 자세 추정한 결과를 그대로 사용하였으며, 이전 영상에서 최적 지역 특징 매칭 쌍이 선택됐다면, Visual Odometry 를 통해 최종 테스트 영상까지의 위치 및 자세를 추정하였다. 해당 논문에서 사용하는 Visual Odometry 의 파이프라인은 그림 2 와 같다.

3. 실험 결과 및 분석

3.1 데이터셋

NAVER LABS 주관 MAPPING & LOCALIZATION CHALLENGE 실외트랙 도전자들에게 제공된 데이터셋은 MMS(Mobile Mapping System) 차량인 R1 으로 판교와 여의도 지역을 주행하며 수집되었다. 해당 데이터 셋은 스테레오 영상과 3 차원 라이다 포인트로 구성됐으며, 자체 개발한 고정밀 측위 기술을 이용해 정확하고 정밀한 3 차원 라이다 포인트와 자세정보를 제공한다는 특징이 있다. 데이터 셋은 판교와 여의도에서 수집한 48,107 장, 86,275 장의 학습 데이터와 판교, 여의도 각각 50 개의 테스트 케이스로 구성됐다. 이때 테스트 케이스는 50 개의 연속된 스테레오 영상으로 이뤄졌다. 이외에도 데이터 셋을 수집한 각 시점의 차량 자세 정보(6DOF), 각 센서의 캘리브레이션 파라미터, 차량-센서의 상대적인 위치 관계, 각 이웃한 프레임 간의 변환 정보가 함께 제공된다.

챌린지에서 테스트 케이스에 대한 정답은 공개되지 않아 SejongRCV 팀은 실험과정에서 시퀀스 학습 데이터셋에서 10 프레임 당 1 프레임을 검증용 데이터셋으로 추출하여 평가를 진행하였다.

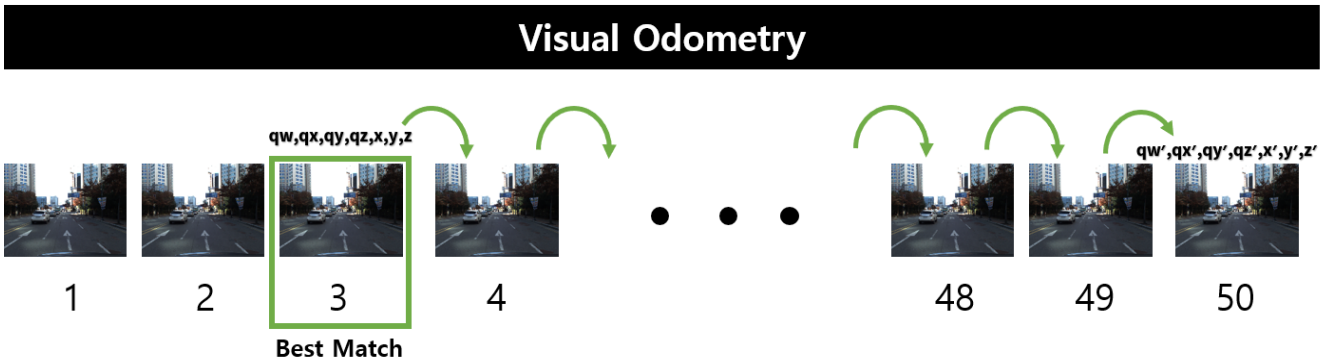

그림 2. SejongRCV 팀의 Visual Odometry 파이프라인

3.2 평가 지표

NAVER LABS MAPPING & LOCALIZATION CHALLENGE 에서는 각 테스트 케이스의 마지막 영상 속 차량의 위치 및 자세를 예측하면 이를 정답(Ground Truth)과 위치, 각도 오차 계산해 평가를 수행한다. 이때 위치와 각도 정확도 기준을 (0.5m/2°), (1m/5°), (5m/10°) 3 단계로 나누어 각 단계에서의 오차가 기준 이하인 비율을 계산한다.

3.3 Retrieval based Place Recognition

NetVLAD 를 이용한 Place Recognition 을 검증용 데이터셋을 제외한 학습 데이터셋을 DB 로 사용하여 이를 본 챌린지에서 사용되는 평가지표로 검증용 데이터셋을 평가한 결과는 표 1 과 같다.

표 1. Place Recognition 을 통한 검증용 데이터셋 성능

DB	Recall	0.5m/2°	1.0m/5°	5.0m/10°
판교	Top 1	28.42%	96.40%	100.00%
	Top 5	34.88%	97.83%	100.00%
	Top 10	35.70%	97.88%	100.00%
여의도	Top 1	32.08%	86.60%	100.00%
	Top 5	36.64%	89.11%	100.00%
	Top 10	39.88%	90.16%	100.00%

3.4 Re-ranking

Place Recognition 을 수행한 이후 선별된 후보군에서 레퍼런스 영상을 선택하기위해 Re-ranking 과정을 수행한다. Re-ranking 과정에는 Super-point 와 Super-Glue 를 사용하였으며, 검증용 데이터셋에 Re-ranking 을 적용한 레퍼런스 영상에 대한 정량적 평가는 표 2 와 같다. Re-ranking 적용 유무에 따른 성능결과는 판교 데이터셋에서 약 1.5%, 여의도 데이터셋에서 약 0.5%의 성능 향상이 나타남을 확인할 수 있다. 그림 3 은 Re-ranking 과정을 통해 실제 유사도가 높은 레퍼런스 영상이 Top-1 으로 선정되는 사례에 해당한다.

표 2. Re-ranking 을 적용한 검증용 데이터셋 성능

DB	0.5m/2°	1.0m/5°	5.0m/10°
판교	29.97%	96.74%	100.00%
여의도	32.51%	87.25%	100.00%

3.5 Pose Estimation

레퍼런스 영상이 선택되면 동시간에 수집된 3 차원 라이다 포인트를 레퍼런스 영상에 투영시킨다. 그리고 레퍼런스 영상과 테스트 영상간 매칭된 지역 특징 쌍에 투영된 3 차원 라이다 포인트를 구해 Pose Estimation 을 수행한다 (그림 4). 이를 통해 정확한 자세를 추정할 수 있으며, 표 3 은 검증용 데이터셋에 대해서 Pose Estimation 을 수행한 결과이다.

표 3 Pose Estimation 을 통한 검증용 데이터셋 성능

DB	0.5m/2°	1.0m/5°	5.0m/10°
판교	99.79%	100.00%	100.00%
여의도	98.71%	99.51%	100.00%

표 3 의 결과를 통해서 검증용 데이터셋에 대해서 높은 정확도로 위치 및 자세 추정이 이뤄짐을 확인할 수 있다. 이를 이러한 과정을 실제 테스트 영상에 적용하여 챌린지에 제출하였고, 챌린지에서의 최종 성능은 표 4 와 같으며 이를 NAVER LABS 에서 공개한 베이스라인 (NetVLAD+ Root SIFT[5])와 비교하여 나타냈다.

표 4 리더보드에 제출된 최종 성능

DB	0.5m/2°	1.0m/5°	5.0m/10°
판교(baseline)	24.00%	42.00%	52.00%
판교(our)	**44.00%**	**58.00%**	**66.00%**
여의도(baseline)	30.00%	54.00%	72.00%
여의도(our)	**82.00%**	**86.00%**	**92.00%**

그림 3. 테스트 영상(QR)에 대해서 Re-ranking 을 적용해
영상 기하학적 유사도가 높은 레퍼런스 영상이 선택됨을 확인할 수 있다.

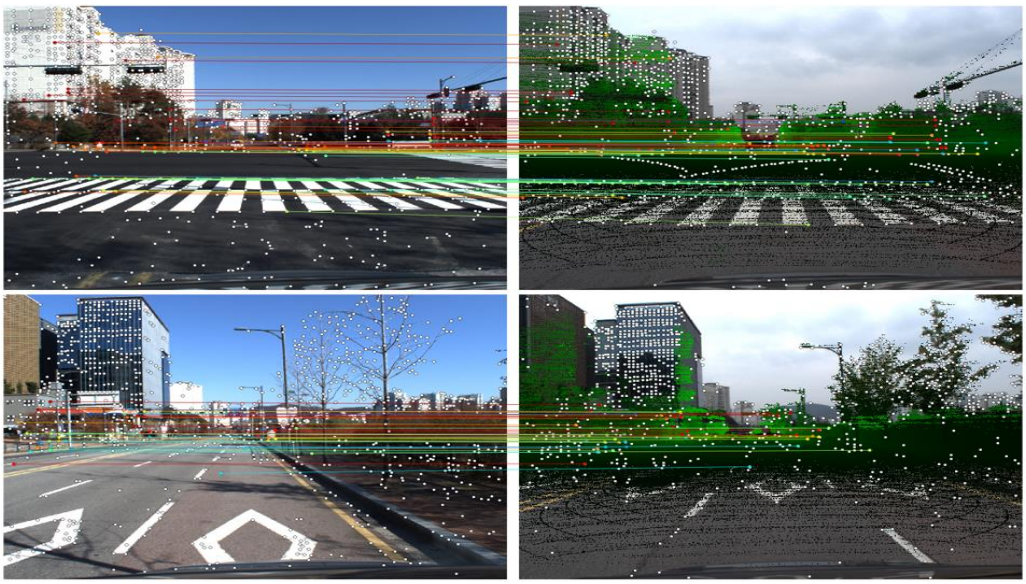

그림 4. 테스트 영상(좌)과 3 차원 라이다가 투영된 레퍼런스 영상(우)의 지역 특징 매칭 결과

4. 결론

본 논문에서는 NABER LABS 주관 MAPPING & LOCALIZATION CHALLENGE 실외 부문에 참가한 Sejong RCV 팀의 방법을 설명하였다. 일부 DB 에 존재하지 않는 테스트 영상에 대해 부정확한 자세가 추정됐지만, 베이스라인보다 높은 정확도로 Visual Localization 을 수행하여 3 등의 성적을 거두었다.

감사의 글

이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2020R1F1A1076987). Visual Localization 연구 분야를 이끌어나갈 인재들이 서로의 경험을 공유할 수 있는 장을 마련해준 네이버랩스에 감사드립니다.

참고문헌

[1] Arandjelovic, Relja, et al. "NetVLAD: CNN architecture for weakly supervised place recognition." *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 2016.

[2] DeTone, Daniel, Tomasz Malisiewicz, and Andrew Rabinovich. "Superpoint: Self-supervised interest point detection and description." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2018.

[3] Sarlin, Paul-Edouard, et al. "Superglue: Learning feature matching with graph neural networks." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020.

[4] Ke, Tong, and Stergios I. Roumeliotis. "An efficient algebraic solution to the perspective-three-point problem." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* .2017

[5] Arandjelović, Relja, and Andrew Zisserman. "Three things everyone should know to improve object retrieval." *IEEE. Conference on Computer Vision and Pattern Recognition*. 2012