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

조 원*, 한대찬*, 최유경†
세종대학교 Robotics and Computer Vision (RCV) 연구실
{jwon, dchan, ykchoi}@rcv.sejong.ac.kr

요 약

본 논문은 NAVER LABS 주관 Mapping & Localization Challenge 실내 트랙에 참가하여 3 등 성적을 거둔 SejongRCV 팀의 방법론에 대해 소개한다. 본 팀은 대회 목적인 영상기반 위치인식(Visual Localization) 문제를 해결하기 위해 Image Retrieval 방법론을 Ensemble 하여 질의 영상(QR)과 시각적으로 유사한 영상 후보 군(Top-K)을 데이터베이스에서 검색한다. 그리고 후보군 내의 영상들을 질의 영상과의 기하학적 정보를 토대로 재정렬(Re-ranking)하고 가장 유사한 후보(Top-1)를 최종 선택한다. 이후, 앞서 선택된 영상의 카메라 정보와 사전에 구축된 3 차원 지도를 함께 이용하여, 질의 영상이 촬영된 카메라의 위치 및 자세 정보(6DoF)를 추정한다. SejongRCV 팀의 방법론은 (위치 오차, 각도 오차)의 기준이 (0.25m, 10.0º) / (0.5m, 10.0º) / (5.0m, 10.0º) 일 때 각 조건 이내 정확도 73.05 / 89.11 / 96.67 의 성능을 보였다. 본 논문에서는 해당 방법론에 대한 자세한 설명과 실험 결과를 소개하며, 실험에 사용한 코드는 아래의 링크에서 확인 가능하다. https://github.com/sejong-rcv/SejongRCV-Indoor

1. 서론

영상기반 위치인식(Visual Localization)은 GPS 사용이 어려운 실내 공간이나, GPS 신호를 신뢰할 수 없는 고층 건물 사이에서 위치 인식을 수행하기 위해 사용되는 기술이다. 이는 영상 정보를 사용하는 기술이며, 위치 인식을 위해 구축된 3 차원 지도를 활용한다. 네이버 랩스에서 주관한 Mapping & Localization Challenge 대회는 앞서 설명한 영상기반 위치인식 문제를 실내/외로 나누어 다룬다. 본 대회의 실내 트랙을 위해 제공된 데이터 셋은 판교 현대 백화점 1 층과 지하 1 층에서 촬영되었으며, 학습을 위해 시퀀스 영상과 각 시퀀스 영상의 카메라 위치 및 자세 정보 그리고 촬영된 공간의 3 차원 지도가 함께 제공되었다.

SejongRCV 팀은 그림 1 과 같은 파이프라인으로 영상기반 위치인식 문제를 해결하였다. 먼저, 영상 검색 단계(Image Retrieval)는 제공된 데이터베이스 내에서 질의 영상(QR)과 유사한 영상의 후보(Top-K)를 찾는 과정이며, 더욱 정확한 후보 선정을 위해 앙상블(Ensemble) 형태의 기술자를 영상 검색 과정에 활용하였다. 재정렬(Re-ranking) 단계는 지도 내 후보 영상과 질의 영상 간 지역 정합(local matching)을 통해 유사도를 재계산 하는 과정이다. 이때, 지역 정합과 RANSAC 을 통해 얻은 인라이어(inlier) 수는 두 영상 사이의 유사도 평가 지표로 활용되었다. 자세 추정(Pose Estimation) 단계는 주어진 3 차원 지도 내에서 질의 영상의 위치 및 자세를 구하는 과정이다. 먼저 가장 유사하다고 판단된 후보 영상(Top-1)과 질의 영상 사이의 3 차원 위치 관계를 계산한다. 그리고 후보 영상과 함께 제공된 위치 및 자세 정보와 추정된 위치 및 자세 정보를 통해 최종적으로 질의 영상의 위치 및 자세 정보를 구하였다.

2. 방법론

2 장에서는 SejongRCV 팀의 방법론을 자세하게 기술한다. 2 장 1 절에서는 질의 영상과 유사한 영상의 후보 군을 찾기 위해 사용한 영상 검색 방법론에 대해 설명한다. 2 장 2 절에서는 유사한 영상의 후보 군과 질의 영상 간의 재정렬을 위한 매칭 및 재정렬 기법에 대해 설명한다. 2 장 3 절에서는 앞서 선택된 가장 유사한 영상의 정보를 이용하여 질의 영상을 촬영한 단일 카메라의 자세 추정 방법에 대해 설명한다.

2.1 영상 검색 (Image Retrieval)

영상 검색(Image Retrieval)은 제공된 데이터베이스 내에서 질의 영상(QR)과 유사한 영상의 후보 군(Top-K)를 찾는 과정이다. 해당 과정은 우선 모든 영상을 전역 기술자(global descriptor)로 기술한다. 이

* 은 공동 저자를 의미하며, † 은 교신 저자를 의미함.

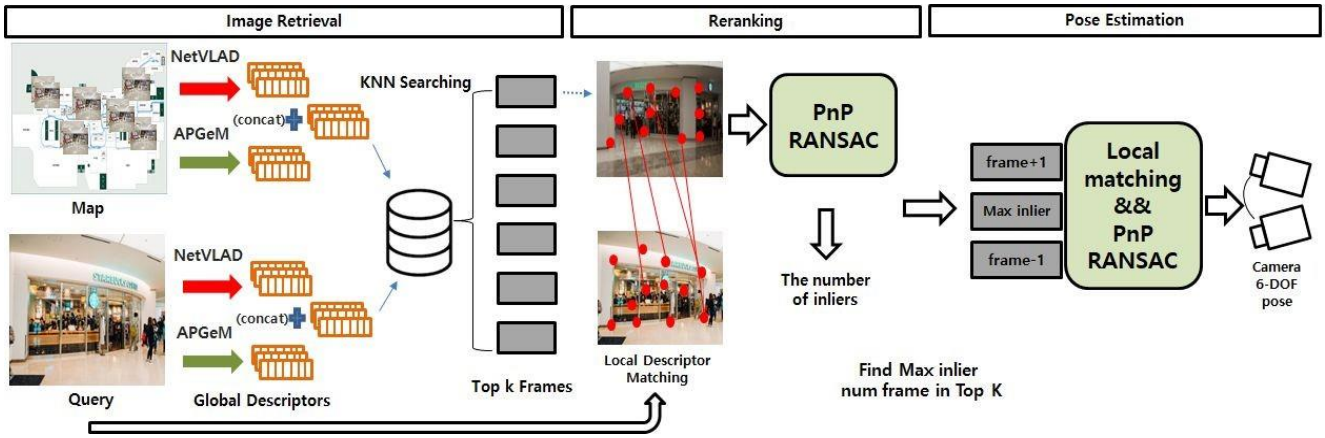


그림 1 SejongRCV 팀의 파이프라인

후, 데이터베이스 내의 영상들과 질의 영상 간의 유사도(L2 similarity)를 계산하여 유사한 영상 후보군을 찾는다. SejongRCV 팀은 영상을 기술하기 위한 전역 기술자로 NetVLAD[1]와 APGeM[2]을 앙상블(Ensemble) 한 전역 기술자를 사용하였으며, 효율적인 기술자 정합을 통한 빠른 검색을 위해 kNN 알고리즘을 사용하였다.

NetVLAD[1] NetVLAD 는 영상 검색 분야에서 기존 전통적인 방식인 VLAD[3]에 딥러닝을 적용한 방법론이다. VLAD 는 지역 기술자(local descriptor)를 추출해 모든 영상에서 대표되는 지역 기술자를 중심점으로 선정하고, 특정 영상의 지역 기술자와 중심점들 간 거리 벡터의 합으로 전역 기술자를 기술하는 방법이다. 이와 달리 NetVLAD 는 백본(Backbone network)을 이용해 영상을 feature 로 기술하고 같은 종류로 분류된 영상들의 feature 와 중심점들 간의 거리 벡터 차이는 최소, 다른 종류로 분류된 영상들의 feature 와 중심점들 간의 거리 벡터 차이는 최대가 되도록 하는 중심점을 학습시켜 더 좋은 표현력을 지닌 전역 기술자를 기술하는 방법이다.

APGeM[2] GeM Pooling(Generalized Mean Pooling)은 기존 feature 의 차원을 줄이기 위해 사용하였던 Max Pooling 과 Average Pooling 을 기반하여 학습 가능한 파라미터를 추가하고 재설계한 방법론이다. 전역 기술자를 기술하는 해당 방법론은 학습 가능한 파라미터에 의해 학습한 데이터에 따라 Max Pooling 에 가까워질 수도 있고 Average Pooling 에 가까워질 수도 있다. 이러한 GeM Pooling 을 학습시킬 때 기존 성능 평가 시 사용되던 mAP(mean Average Precision)라는 지표 자체를 손실 함수로 설계한 방법론을 APGeM 이라 한다.

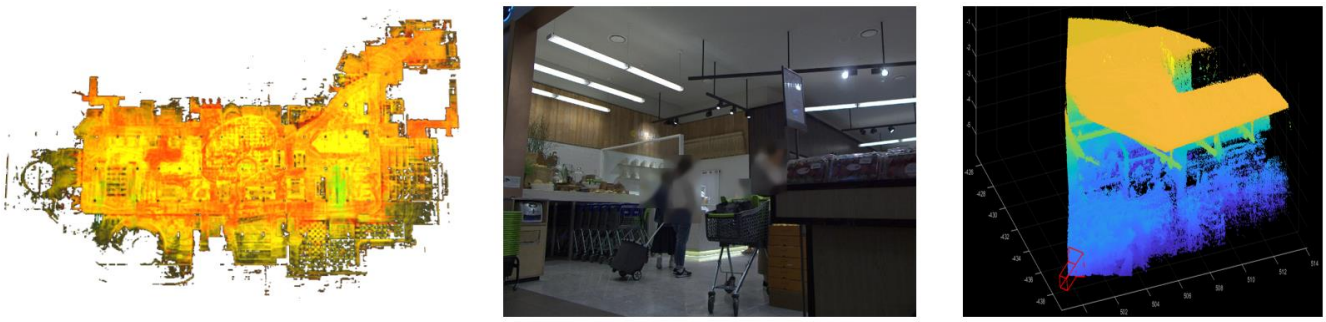
Ensemble 특정 방법론들로 예측할 때 생기는 오류를 줄이고 데이터에 대한 과적합 현상을 줄이기 위해 여러 방법론들의 예측 결과를 동시에 사용하는 앙상블 기법이 사용되어왔다. SejongRCV 팀도 영상 검색의 성능을 높이기 위해 NetVLAD 와 APGeM 을 앙상블 하였다. 앙상블 할 때 NetVLAD 의 전역 기술자는 PCA 를 이용하여 4,096 차원으로 축소시켰고, 축소된 NetVLAD 의 전역 기술자와 APGeM 의 전역 기술자를 이어 붙인 후 L2 정규화를 적용하였다.

2.2 재정렬 (Re-ranking)

영상 검색에 사용되는 전역 기술자(global descriptor)는 인코딩 과정에서 영상 내 지역 정보를 상실한다. 즉, 전역 기술자는 분류(classification) 수준의 문제에는 적합하나 식별(identification) 수준 문제에서는 어려움이 따른다는 것을 의미한다. 영상 내 지역 정보 기반으로 식별 수준에 가까운 문제를 해결하기 위해 본 팀은 재정렬이라는 방법을 사용하였다. 재정렬(Re-ranking)이란 질의 영상과 이에 유사한 영상 후보 군(Top-K) 간의 지역 특징 정합(local feature matching)을 통해 영상 간 유사도를 재계산하는 과정을 말한다. SejongRCV 팀은 재정렬 방식으로 PnPR(Perspective-n-Point Re-ranking)을 사용하였으며, 해당 방법론 사용 시 필요한 지역 특징 정합으로는 실험을 통해 D2-Net[4]과 SuperPoint[5]를 활용한 SuperGlue[6] 중 SuperGlue 를 사용하였다.

D2-Net[4] 일반적으로 지역 불변 특징량은 영상 내에서 분별력 있는 좌표를 탐지(detect)한 뒤, 해당 좌표를 기준으로 주변 정보를 기술(descript)한다. 그러나 탐지를 위한 알고리즘의 공통적인 특징은 밤과 낮 같은 급격한 환경 변화 시에 반복성(repeatability)이 강인하지 못하다는 단점이 존재한다. 이러한 문제를 해결하기 위한 D2-Net 은 영상 전체의 좌표에 대해 기술한 뒤, 이들 중 분별력 있는 좌표를 탐지하도록 설계된 방법론에 해당한다. 이는 단일 크기의 영상을 사용하는 D2-Net SS(Single-scale)과 영상 크기 변화에 따른 강인성을 보이기 위해 기술 및 탐지 과정을 여러 장의 영상 피라미드를 두어 사용하는 D2-Net MS(Multi-scale)로 나뉜다.

SuperGlue[6] SuperGlue 는 자기지도 학습기반(self-supervised learning)으로 설계된 지역 기술자인 SuperPoint 를 활용하며, 두 영상에서 지역 기술자



(a) (b) (c)

그림 2 고정밀 지도와 지도 내 영상의 위치 정보를 이용해 획득한 3 차원 정보 예시,
(a)는 지하 1 층의 고정밀 3 차원 지도 (b)는 Top-1 영상 (c)는 Top-1 영상과 쌍을 이루는 3 차원 정보

간의 연관성과 각 영상 내 지역 기술자 간의 연관성을 활용하여 Attentional GNN(Graph Neural Network) 설계 및 학습시키는 방법론이다. 해당 방법론은 두 영상 간의 지역 정합 시, 각 영상마다의 지역 기술자 분포 및 관계 정보도 활용하기에 좋은 표현력으로 지역 정합이 가능하다.

PnPR PnPR 은 영상 후보 군의 재정렬을 위해 도입한 방법론으로 각 영상 내의 지역 정합으로 얻은 대응점과 영상 후보 군 내에 각각 할당된 3 차원 지도 좌표를 이용하여 PnP 알고리즘의 해를 찾고, 이를 RANSAC 으로 최적화 후 생기는 인라이어(inlier)의 개수로 재정렬하는 방법론이다. 이를 위한 PnP 알고리즘으로는 3 개의 3 차원-2 차원 좌표 쌍만으로 높은 계산 효율성과 높은 정확도를 보이는 AP3P[7]를 사용하였다.

2.3 자세 추정(Pose Estimation)

3D-2D Corresponding points 영상을 촬영한 카메라의 자세 정보 추정을 위해서는 3 차원-2 차원 좌표 쌍으로 회전 성분과 이동 성분을 계산하는 PnP 알고리즘이 주로 사용된다. SejongRCV 팀 또한 이와 같은 방법을 사용하였으며, 이를 위해 우선 가장 유사한 영상(Top-1)과 질의 영상(QR)간의 지역 정합으로 대응점(corresponding point) 좌표를 찾아내었다. 그리고 계산 효율성을 목적으로 그림 2 와 같이 가장 유사한 영상의 자세 정보와 캘리브레이션 파라미터(Intrinsic Parameter)를 통해 고정밀 3 차원 지도 상의 좌표를 가장 유사한 영상으로 사영(projection)하여 해당 영상과 쌍을 이루는 3 차원 정보를 추출하였다. 이와 같이 추출된 3 차원 정보를 2 차원 지역 특징량이 추출된 영상에 사영하여 3 차원-2 차원 좌표 쌍을 만들었으며, 좌표 쌍을 생성을 위한 3 차원 데이터 사영 시, 캘리브레이션 파라미터의 오차 및 3 차원 정밀 지도의 오차로 인해 2 차원 지역 특징량 좌표에 정확하게 할당되지 못하는 경우가 발생한다. 이는 3 차원 데이터에 가장 가까운 지역 특징량 좌표를 할당하는 방식으로 해결하였다.

6D Pose Estimation 앞선 방식으로 구한 3 차원-2 차원 좌표 쌍을 입력으로 PnP 알고리즘을 최적화하여 질의 영상의 자세 정보를 추정하였다. PnP 알고리즘으로는 AP3P 를 사용하였으며, 추정 정확도를 높이기 위해 보다 밀집된 3 차원 지도 상의 좌표를 활용하는 Covisibility 방식을 사용하였다.

Covisibility 전체 영역의 3 차원 지도에서 추출된 가장 유사한 영상이 나타내는 3 차원 지도에는 일부 좌표의 오차가 존재한다. 이는 가장 유사한 영상과 시간 상 연속된 영상의 3 차원 지도를 사용하는 방법인 Covisibility 방법을 사용해 사영 오차를 줄일 수 있었으며 질의 영상의 보다 정확한 자세 정보 추정이 가능하다.

3. 실험 결과 및 분석

3.1 Dataset

본 대회에서 제공된 데이터 셋은 판교 현대백화점 1 층과 지하 1 층에서 촬영되었으며, 위치 정보를 포함하는 학습 데이터와 알고리즘을 통해 영상의 위치 정보를 추정해야하는 평가 데이터로 나뉘어져 진다. 학습 베이스에서 제공되는 영상은 Basler 카메라 혹은 Samsung Galaxy S9 으로 촬영되었으며, 각 카메라들의 캘리브레이션 파라미터(Intrinsic Parameter)도 함께 제공되었다. Velodyne Puck sensor 로부터 취득된 LIDAR 데이터는 촬영된 영상과 쌍을 이루어 제공되었고, 네이버 랩스에서 자체적으로 구성한 고정밀 3 차원 지도도 함께 제공되었다. 평가 데이터의 경우 학습 데이터와 서로 다른 날 촬영되었으며, 3 차원 라이다 센서 정보 없이 카메라 영상과 해당 카메라의 캘리브레이션 정보만 제공되었다. 실험에 사용된 평가 지표는 (위치 오차, 각도 오차)의 기준이 (0.25m, 10.0º) / (0.5m, 10.0º) / (5.0m, 10.0º) 일 때 각 조건 이내의 성능을 측정하는 방식으로 정확도를 계산하였다.

3.2 Implementation details

Pre-processing 본 대회의 데이터 셋 일부는 렌즈 왜곡이 제거되지 않은 상태(distorted image) 로 제공되었으며, SejongRCV 팀은 렌즈 왜곡이 지역 특징량 정합(Local feature matching) 성능 저하의 원인이 되므로 영상 검색 알고리즘 적용 전 사전 처리 과정을 통해 영상 왜곡을 제거하였다.

Image Retrieval NetVLAD 의 경우, 전이 학습(transfer learning)을 위해 VGG16[8]을 백본(Backbone network)으로 선정하였고, 표 1 의 비교 실험을 위해 D2-Net[4]에서 MegaDepth 데이터 셋[9]으로 학습된 사전학습모델(D2_NetVLAD)과 Pittsburgh 데이터 셋[10]으로 학습된 사전학습모델(Pitts_NetVLAD)을 사용하였다. 이때 NetVLAD 의 중심점 수는 64 이며, 입력 영상 크기는 (512, 512) 이다. 학습 시, 영상 분류 및 검색에 적합한 표현력을 학습하기 위해 Metric Learning 기법을 사용하며, 본 팀은 다음과 같은 방법으로 Positive/Negative 셋을 구성하였다. 기분영상이 촬영된 카메라의 위치/자세와 비교하여 오차가 (0.5m, 10.0º) 이내인 학습 데이터 셋을 Positive 셋으로 사용하였으며, 조건을 만족하는 영상이 없는 경우에는 오차허용범위를 늘려 (1.0m, 15.0º) 이내의 오차를 갖는 영상을 Positive 셋으로 사용하였다. Negative 셋의 경우, 기준 영상이 촬영된 카메라 위치 및 자세를 비교해 x 축, y 축으로 각각 10.0m 이상 떨어진 영상을 랜덤으로 선택하여 사용하였다.

APGeM 의 경우, 추가적인 학습 없이 ImageNet 데이터 셋으로 학습된 사전학습모델(APGeM)과 Google_Landmarks 데이터 셋으로 학습된 사전학습모델(APGeM_LM18)을 사용하였다.

최종적으로, 유사 영상 검색의 성능향상을 위해 여러 기술자를 함께 사용하는 앙상블(Ensemble) 기법을 사용하였다. 판교 현대백화점 데이터 셋으로 전이 학습한 D2_NetVLAD 와 사전학습모델인 APGeM, APGeM_LM18 을 이어 붙이고, 이를 L2 정규화 하여 사용하였다. 이때, D2_NetVLAD 의 경우 PCA 를 이용하여 4,096 차원으로 차원 축소 후 사용하였다.

Re-ranking 본 팀은 지역 특징량 정합(Local feature matching)을 위해 MegaDepth 데이터 셋으로 학습된 D2-Net 을 사용하였다. SuperGlue 역시 추가학습과정 없이 ScanNet 데이터 셋으로 학습된 사전학습모델을 사용하였고, 입력 영상의 크기만 (1024, 960)으로 변경해 사용하였다.

3.3 Experiments

본 절에서는 정량적 성능 검증(표 1~4)을 위해 평가용 데이터 셋이 아닌 검증용 데이터 셋을 사용한다. 검증용 데이터 셋은 다음과 같이 구성하였다. 학습 데이터 셋(지하 1 층) 중 2019-08-20 에 촬영된 데이터의 일부를 검증용 데이터로, 2019-04-16 에 두 번에 걸쳐 촬영된 데이터 전체를 학습용 데이터로 사용하였다. 검증용 데이터는 일부 샘플링 하였고, 전체 시퀀스 영상을 10 스텝당 1 장을 선택하여 총 405 장으로 구성하였다.

Image Retrieval 표 1 에서 train X 는 사전학습모델을 이용한 정량적 평가를 의미하며, train O 는 본 대회에서 제공된 데이터로 학습된 전이학습모델을 이용한 평가를 의미한다. 성능 평가 결과 앙상블(Ensemble) 방법론이 가장 좋은 성능을 보였으며, 이후 실험 (표 2~4)에서는 앙상블 기반의 영상 검색 기술 결과를 고정하여 실험하였다.

Local Matching 다양한 지역 기술자(local descriptor) 중 본 대회 환경에 가장 적합한 지역 기술자를 선정하기 위해, 검증 영상과 재정렬하기 이전 가장 유사한 영상(Top-1)으로 자세 추정 실험을 진행하였다. 표 2 의 실험 결과, 전통적인 알고리즘인 SIFT[11], Root SIFT[12]와 비교했을 때 D2-Net 과 SuperGlue 는 월등히 우수한 성능을 보였으며, D2-Net 이 가장 높은 성능을 보였다. 그러나 SuperGlue 의 처리 속도는 D2-Net 에 비해 2 배정도 빠르다는 장점이 있다. 표 2 에서 D2SS 는 D2-Net SS, D2MS 는 D2-Net MS 를 의미한다.

표 1 검증용 데이터 셋에서
Image Retrieval 방법론에 따른 Top-K 정확도

Method	Top-1	Top-10
SIFT+VLAD	2.22/9.38/40.00	4.68/17.53/64.20
Root SIFT+VLAD	1.98/8.89/40.00	4.44/17.78/66.42
D2_NetVLAD (train X)	0.99/3.70/17.53	1.48/4.94/29.63
D2_NetVLAD (train O)	1.98/7.41/46.91	4.94/20.25/76.30
D2_NetVLAD (train X, PCA)	1.73/7.65/37.28	3.46/17.28/68.40
D2_NetVLAD (train O, PCA)	1.73/8.89/45.68	4.44/19.75/75.31
Pitts_NetVLAD (train X)	1.23/5.43/29.88	3.21/14.07/56.79
Pitts_NetVLAD (train O)	1.23/8.40/45.93	4.20/19.75/71.36
Pitts_NetVLAD (train X, PCA)	1.73/7.65/34.57	4.20/17.28/67.90
Pitts _NetVLAD (train O, PCA)	1.98/9.63/45.68	4.69/20.25/75.80
APGeM	1.73/6.42/32.10	2.47/11.60/59.26
APGeM_LM18	1.23/7.41/34.32	2.47/14.32/64.94
Ensemble	**1.98/9.88/46.42**	**4.44/20.00/76.05**

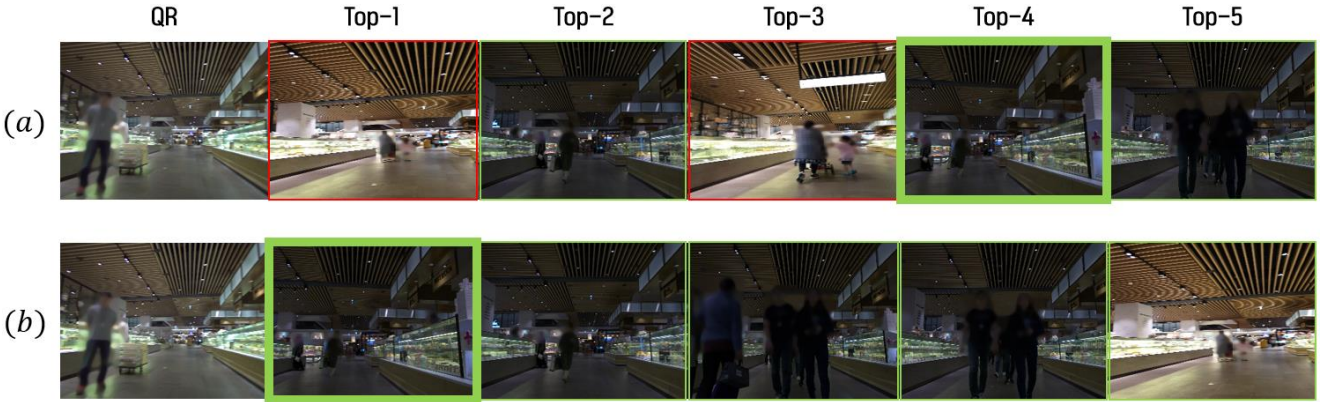


그림 3 SejongRCV 팀의 최종 파이프라인에서 영상 검색 및 재정렬 시 정성적 결과,
(a)는 영상 검색 (b)는 Top-10 까지 후보 군 재정렬 후 Top-5 까지의 정성적 결과 표현,
녹색 테두리는 Positive, 적색 테두리는 Negative 의미

표 2 검증용 데이터 셋에서 Ensemble 방식으로
검증 영상과 가장 유사한 영상 검색 후,
두 영상을 이용한 자세 추정 성능

Method	Top-1
Ensemble + SIFT	17.28/29.63/54.32
Ensemble + Root SIFT	18.77/32.10/58.52
Ensemble + D2SS	**41.23/52.84/71.60**
Ensemble + D2MS	**39.75/53.58/70.86**
Ensemble + SuperGlue	**37.78/51.36/69.88**

Re-ranking 영상 검색(Image Retrieval)의 결과인 검증 영상과 유사한 영상의 후보 군(Top-K)을 PnPR 방법으로 재정렬(Re-ranking)한 후, 자세 정보를 추정하였다. 재정렬 후, 가장 유사한 영상(Top-1)으로 자세 추정 결과, 표 3 에서 볼 수 있듯이, 모든 방법론의 성능이 크게 향상되었다. 재정렬하기 위한 영상 후보 군의 크기는 K=10 이며, 정성적인 결과는 그림 3 에서 확인 가능하다.

Covisibility 검증용 데이터 셋에서 앙상블(Ensemble) 방식으로 얻은 영상 후보 군을 PnPR 방식으로 재정렬하고 검증 영상과 가장 유사한 영상으로 자세 추정 시, 해당 영상과 일정 간격 인접한 영상의 3 차원 좌표를 활용하는 Covisibility 에 대한 성능 비교 실험을 진행하였다. 표 4 에서 Covis1 은 검증 영상 앞뒤 각 1 개 영상의 3 차원 좌표 활용을 의미하며, Covisibility 방식을 추가했을 때 대부분의 방법론에서 성능이 향상되었다. 이로써 밀집된 3 차원 좌표를 사용하는 것이 3 차원-2 차원 좌표 대응쌍 할당에서 사영 오차를 줄여 보다 정밀한 자세 추정을 가능하게 하는 것을 확인하였다.

표 3 검증용 데이터 셋에서 Ensemble 방식으로
검증 영상과 가장 유사한 영상 후보군 Top-10 검색 후,
PnPR 방식으로 재정렬하여 선정된 검증 영상과
가장 유사한 영상을 이용한 자세 추정 성능

Method	Top-1
Ensemble + D2SS	41.23/52.84/71.60
Ensemble + D2SS + PnPR	**51.85/67.41/85.68**
Ensemble + D2MS	39.75/53.58/70.86
Ensemble + D2MS + PnPR	**51.36/68.40/84.94**
Ensemble + SuperGlue	37.78/51.36/69.88
Ensemble + SuperGlue + PnPR	**51.60/64.69/85.19**

표 4 검증용 데이터 셋에서 Ensemble 방식으로
검증 영상과 가장 유사한 영상 후보군 Top-10 검색 후,
PnPR 방식으로 재정렬하여 선정된 검증 영상과
가장 유사한 영상 및 연속된 일정 간격 영상의
3 차원 좌표를 이용한 자세 추정 성능

Method	Top-1
Ensemble + D2SS + PnPR	51.85/67.41/85.68
Ensemble + D2SS + PnPR + Covis1	**53.83/71.36/87.65**
Ensemble + D2MS + PnPR	51.36/68.40/84.94
Ensemble + D2MS + PnPR + Covis1	**54.32/72.35/87.65**
Ensemble + SuperGlue + PnPR	51.60/64.69/85.19
Ensemble + SuperGlue + PnPR + Covis1	**50.62/68.40/86.42**

4. 결론

SejongRCV 팀은 네이버 랩스가 주관한 본 대회 실내 트랙에 참가하여 검증용 데이터 셋을 이용한 다양한 실험 결과 중 (Ensemble + SuperGlue + PnPR + Covis1)을 최종 파이프라인으로 선정하였다. 선정된 파이프라인을 이용한 평가 데이터 셋 결과는 (위치 오차, 각도 오차) 기준 (0.25m, 10.0º) / (0.5m, 10.0º) / (5.0m, 10.0º) 에서 각각 73.05 / 89.11 / 96.67 의 정확도 성능을 보였다. 해당 성능은 베이스라인 보다 높은 정확도로 실내 트랙 리더보드에서 3 등의 성적을 기록하였다.

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 네이버 랩스에 감사드립니다.

참고 문헌

[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] Revaud, Jerome, et al. "Learning with average precision: Training image retrieval with a listwise loss." Proceedings of the IEEE International Conference on Computer Vision. 2019.

[3] H. Jegou, M. Douze, C. Schmid, and P. P ´ erez, “Aggregating local ´ descriptors into a compact image representation,” in CVPR, June 2010

[4] Dusmanu, Mihai, et al. "D2-net: A trainable cnn for joint detection and description of local features." arXiv preprint arXiv:1905.03561 (2019).

[5] D. DeTone, T. Malisiewicz and A. Rabinovich, "SuperPoint: Self-Supervised Interest Point Detection and Description," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Salt Lake City, UT, 2018, pp. 337-33712, doi: 10.1109/CVPRW.2018.00060.

[6] 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.

[7] Tong Ke and Stergios Roumeliotis. An efficient algebraic solution to the perspective-three-point problem. In Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on. IEEE, 2017.

[8] S. Liu and W. Deng, "Very deep convolutional neural network based image classification using small training sample size," 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR), Kuala Lumpur, 2015, pp. 730-734, doi: 10.1109/ACPR.2015.7486599.

[9] Zhengqi Li and Noah Snavely. MegaDepth: Learning singleview depth prediction from internet photos. In Proc. CVPR, 2018.

[10] A. Torii, J. Sivic, T. Pajdla, and M. Okutomi. Visual place recognition with repetitive structures. In Proc. CVPR, 2013.

[11] David G Lowe, “Distinctive Image Features from Scale-Invariant Keypoints,” International Journal of Computer Vision, vol.50, No. 2, 2004, pp.91-110.

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