

# 센서 융합 시스템을 이용한 심층 컨벌루션 신경망 기반 6자유도 위치 재인식

## A Deep Convolutional Neural Network Based 6-DOF Relocalization with Sensor Fusion System

조형기<sup>1</sup>·조해민<sup>1</sup>·이성원<sup>1</sup>·김은태<sup>†</sup>

HyungGi Jo<sup>1</sup>, Hae Min Cho<sup>1</sup>, Seongwon Lee<sup>1</sup>, Euntai Kim<sup>†</sup>

**Abstract:** This paper presents a 6-DOF relocalization using a 3D laser scanner and a monocular camera. A relocalization problem in robotics is to estimate pose of sensor when a robot revisits the area. A deep convolutional neural network (CNN) is designed to regress 6-DOF sensor pose and trained using both RGB image and 3D point cloud information in end-to-end manner. We generate the new input that consists of RGB and range information. After training step, the relocalization system results in the pose of the sensor corresponding to each input when a new input is received. However, most of cases, mobile robot navigation system has successive sensor measurements. In order to improve the localization performance, the output of CNN is used for measurements of the particle filter that smooth the trajectory. We evaluate our relocalization method on real world datasets using a mobile robot platform.

**Keywords:** Relocalization, Sensor Fusion, Convolutional Neural Network, pose regression

### 1. 서 론

로봇의 위치인식은 내비게이션 분야에서 현재 로봇이 어디에 있는지를 추정하는 문제로서 다양한 환경에서 많은 연구가 진행되었다. 위치인식 문제 중 위치 재인식(Relocalization)은 로봇이 이전에 방문했던 지역을 재방문했을 때 현재 절대적인 위치를 인식하며 SLAM (Simultaneous Localization and Mapping)의 루프 클로징(Loop Closing) 검출이나 납치(Kidnapped) 상황에서의 위치인식 문제에도 적용 가능한 필수 요소 기술이다.

비전 센서를 이용한 위치 재인식은 외형 기반 방법과 학습 기반 방법으로 연구가 진행되어 왔다<sup>[1]</sup>. 외형 기반 방법은

SIFT, SURF, BRIEF, ORB와 같은 Hand-crafted 특징점을 영상으로부터 추출하는 것을 기본으로 한다. 매우 많은 영상 데이터를 모아서 추출된 특징점으로 특정 모델을 만들고 재 방문 시 모델의 특징과 매칭하여 현재 센서의 자세(pose)를 추정한다. 이러한 방법 중 널리 알려진 방법은 Structure-from-Motion (SfM)<sup>[2]</sup>을 통해 3차원 지도를 생성하고 새로운 쿼리 이미지가 들어왔을 때, 해당 이미지의 특징점과 모델링 된 특징점과의 비교를 통해 센서 자세를 추정한다. 또는, 특징점으로 미리 설정한 개수만큼의 대표 특징으로 이루어진 코드북(Codebook)을 생성하고 현재 이미지로부터 추출된 코드워드(Codeword)를 이용하여 기존 이미지와 매칭하는 Bag of Words (BoW)<sup>[3]</sup> 방식이 널리 사용되고 있다.

최근 딥러닝이 크게 발전하면서 학습을 기반으로 하여 위치인식을 하는 연구도 진행되고 있다. 학습 기반 방법은 영상 이미지를 이용한 심층 컨벌루션 신경망(Deep Convolutional Neural Network, CNN) 기반 딥러닝 기술로써 학습을 위한 데이터셋을 이용하여 Pose Regression을 수행하는 신경망을 학습시킨다<sup>[4,5]</sup>. CNN은 이미지 프로세싱에서 주로 사용되어 영상에서의 객체 검출, 인식과 같은 분야에서 좋은 성능을 보여

Received : Dec. 8. 2018; Revised : Feb. 1. 2019; Accepted : Feb. 6. 2019

※ This work was supported by the Industrial Convergence Core Technology Development Program(No. 10063172, Development of robot intelligence technology for mobility with learning capability toward robust and seamless indoor and outdoor autonomous navigation) funded by the Ministry of Trade, Industry & Energy (MOTIE), Korea.

1. Researcher, Dept. of Electrical and Electronic Engineering, Yonsei University, Seoul, Korea (hygijo, hm.cho, won4113@yonsei.ac.kr)

† Professor, Corresponding author: Dept. of Electrical and Electronic Engineering, Yonsei University, Seoul, Korea (etkim@yonsei.ac.kr)

주는데<sup>[1]</sup>, 위치 재인식 분야에서 CNN을 사용하는 방법의 가장 큰 장점은 특정한 지도 구조를 갖고 있지 않고 오직 신경망 결과로 위치인식이 가능하기 때문에 필요한 메모리가 지역의 크기와 상관 없다는 점이다.

하지만, 비전 센서만을 사용하는 경우, 특히 실외 환경 데이터셋에서 뚜렷한 위치인식 성능을 보여주지 못한다. 이는 조도, 날씨, 동적 객체 등 다양한 환경 변화에 의해 학습에 사용된 데이터셋을 취득했을 때와 테스트에 사용되는 데이터셋을 취득했을 때 환경이 많이 다를 수 있기 때문이다. 또한, 단일 이미지를 입력으로 사용하기 때문에 이전 시간 단계에서 입력 값과 전혀 독립적으로 결과를 보여준다. 로봇의 주행에 있어서 센서 입력은 항상 연속적이기 때문에 이전 입력을 고려하지 않는 것은 비효율적이며 정밀한 위치인식 수준에는 못 미치게 된다.

또한, 3차원 레이저 스캐너(3D Laser Scanner)를 이용하는 위치인식의 경우 Monte Carlo Localization (MCL)<sup>[6]</sup> 방법이 널리 알려져 있는데 초기 위치를 모를 때 미리 구축된 지도 기반으로 Bayesian update를 통해 전역 위치인식(Global Localization)을 수행한다. 이 방법은 전체 공간에 대한 지도 정보를 갖고 있어야 하므로 넓은 지역에서 동작할 수록 필요한 메모리가 증가하고 파티클이 수렴하는데 많은 시간이 걸릴 수 있어 넓은 실외 환경에서는 부적합하다.

이러한 두 방법의 단점을 극복하기 위해 본 논문에서는 실시간 위치 재인식을 위해 센서 정보의 융합과 알고리즘의 융합이라는 두 가지 측면을 모두 고려한다. 즉, 3차원 레이저 스캐너와 비전 센서를 융합한 센서 시스템을 통한 센서 정보 보완과 학습 기반 및 파티클 필터 기반 위치인식이 동시에 사용되는 방법을 제안한다. 심층 컨벌루션 신경망의 학습에는 카메라 영상의 RGB 정보뿐만 아니라 주변 환경의 거리 정보도 함께 입력으로 사용한다. 또한, 이동 로봇이 주행하면서 얻는 연속적인 센서 입력으로부터 CNN 결과를 얻고 이를 측정값으로 사용하는 파티클 필터(Particle Filter)를 통해 더욱 정밀한 위치인식을 수행한다. 제안된 방법은 로봇 플랫폼을 이용해서 실외에서 취득한 데이터셋을 통해 기존 학습 기반 위치 재인식 방법과 비교되었다.

## 2. 선행 연구 조사

센서 위치 재인식에 대한 연구 초기에는 외형 기반 방법으로 연구가 진행되어 왔다. 위치 재인식은 가본 곳을 재방문하여 기존 데이터 또는 구축된 모델과 비교 한다는 점에서 장소 인식(Place Recognition)<sup>[7]</sup>과 같은 역할을 수행한다. 장소 인식은 주어진 이미지 데이터셋 중에서 입력 이미지가 어떤 이미

지와 일치 하는지를 판단한다. 반면, 위치 재인식은 입력 이미지가 찍혔을 때 카메라의 6-DOF 자세를 추정해야 하므로 추가적으로 기하학적인 변환에 대한 문제 또한 해결해야 한다<sup>[1]</sup>.

위치 재인식을 위한 학습 기반 방법 중 RGBD 센서를 이용하여 각 픽셀의 위치 정보인 Scene coordinate을 추정할 때 Regression Forest를 사용한 Scene Coordinate Regression Forest (SCoRF)<sup>[8]</sup>와 이를 발전시켜 CNN으로 각 픽셀의 Scene coordinate를 추정하고 Scoring function에 CNN을 사용한 DSAC<sup>[9]</sup>이 정확한 위치인식 성능을 보여주었다. R. Li et al.<sup>[11]</sup>은 RGBD 센서로부터 depth를 인코딩한 이미지도 함께 dual-stream CNN 학습에 사용하였다. RGB와 depth 입력이 각각의 CNN에 입력되며 출력된 feature를 concatenate 시켜서 최종 regression에 사용한다. 이러한 연구들은 RGBD 센서를 사용하기 때문에 실내에서만 동작 가능하다는 단점이 있다.

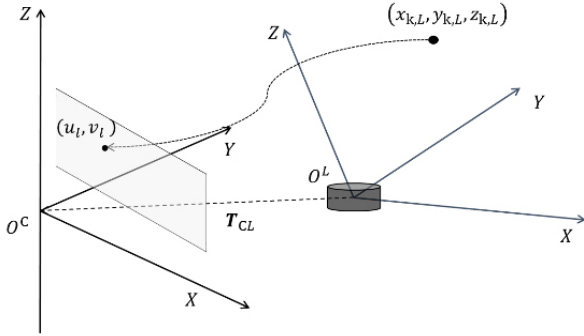
컴퓨터 비전 분야에서 CNN을 이용한 딥러닝 방법이 높은 성능을 보여주었고 위치인식에 있어서는 장소 인식 분야에 Chen et al.<sup>[10]</sup>의 연구가 처음 사용되었다. PoseNet<sup>[4]</sup>은 센서 자세를 end-to-end로 CNN을 학습시켜 광역 공간에서 좋은 위치 인식 결과를 보여준 첫 연구이다. 영상 특징점을 추출하여 사용하는 것보다 CNN이 위치인식에 중요한 특징을 학습을 통해 결정하며 SfM을 통해 학습 데이터셋의 ground truth를 얻는다. PoseNet 이후 위치인식 성능을 높이기 위한 다양한 방법들이 제안되었는데, Kendall et al.<sup>[5]</sup>은 PoseNet의 불확실성을 모델링해서 향상된 버전을 제안하였다. 또한, Clark et al.<sup>[11]</sup>은 CNN과 Bidirectional RNN을 사용한 네트워크로 비디오 클립에서의 카메라 자세를 추정한다. Walch et al.<sup>[12]</sup>은 CNN 출력을 LSTM으로 사용하여 성능을 향상 시켰다.

이와 같은 학습 기반 위치 재인식은 외형 기반 방법보다 위치인식 정확도는 조금 떨어지지만 조명, 날씨, 동적 객체, blur, occlusion 등 환경 변화가 심한 경우에 강인하며 빠르다.

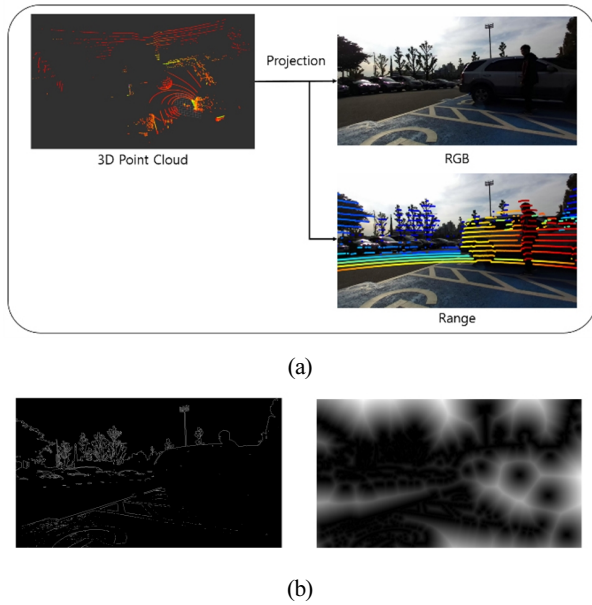
## 3. 심층 컨벌루션 신경망 기반 위치 재인식

### 3.1 센서 융합 시스템

본 연구에서는 3차원 레이저 스캐너 한 대와 스테레오 카메라가 설치된 센서 융합 시스템을 사용한다. 다만, 스테레오 카메라에서 왼쪽 카메라만 알고리즘에 이용한다. 3차원 레이저 스캐너와 카메라 간의 캘리브레이션을 통해 위치 관계를 [Fig. 1]과 같이 파악한다. 이중 센서 간의 캘리브레이션<sup>[13]</sup>은 coarse-to-fine 전략을 사용한다. 정해진 모양과 크기의 인공 특징점을 이중 센서에서 동시에 검출하고 정합시켜 정합 오차를 적게 만드는  $T_{st}$ 을 반복을 통해 도출한다. 이 방법은 sparse한 포인



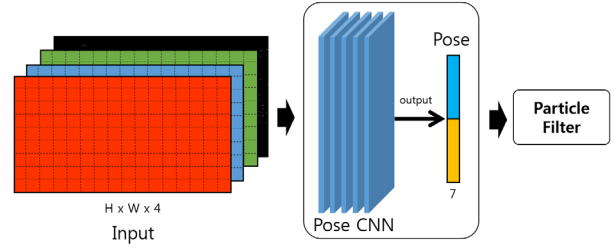
[Fig. 1] Sensor fusion of image and 3D point cloud



[Fig. 2] (a) Input images for pose regression CNN : RGB image with range value, (b) Edge image with Sobel operator and IDT to the RGB image on figure (a)

트 클라우드를 취득하는 레이저 스캐너일수록 부정확 하기 때문에 앞선 방법을 사용한 뒤 수동으로 회전 행렬을 보정하여 정밀한  $T_{SL}$ 을 얻는다. 올바른 정합을 위해 M. Velas et al.<sup>[13]</sup>에서 제안한 edge 기반의 cost 함수를 이용한다. 카메라 이미지를 Sobel operator를 통해 edge만 남아있는 이미지로 변환하고 Inverse Distance Transform (IDT)를 적용하여 각 픽셀 별로 가장 가까운 edge와의 거리에 따라 값을 정한다. 또한 투영된 3차원 레이저 스캐너의 포인트에서 인접한 두 포인트와의 최대 거리를 각 픽셀 값으로 갖는 이미지를 만들고 두 이미지의 비교를 통해 유사도를 판단한다. [Fig. 2(b)]는 Sobel operator와 IDT 적용된 이미지를 보여준다.

3차원 레이저 스캐너에서 측정하는 포인트 클라우드  $\mathbf{Y}_L$ 는 식 (1)<sup>[14]</sup>과 같으며 식 (2)와 캘리브레이션 결과  $\mathbf{T}_{SL}$ 를 통해 카



[Fig. 3] A deep convolutional neural network architecture: GoogLeNet with pose regression

메라 이미지 평면으로 투영시킬 수 있다.  $N_{scan}$ 은 입력 포인트의 수를 나타낸다.

$$\mathbf{Y}_L = [x_k \ y_k \ z_k]^T, \ k = 1, \dots, N_{scan} \quad (1)$$

$$s \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = P \cdot \mathbf{T}_{SL} \cdot \begin{bmatrix} x_k \\ y_k \\ z_k \\ 1 \end{bmatrix}, \mathbf{T}_{SL} = \begin{bmatrix} R_{SL} & \mathbf{t}_{SL} \\ 0 & 1 \end{bmatrix} \quad (2)$$

$P$ 는 프로젝션 행렬(Projection matrix),  $\mathbf{T}_{SL} \in SE(3)$ 은 두 센서 간의 강체 변환 행렬(rigid-body transformation matrix)를 나타낸다. 3차원 공간 포인트의 homogeneous coordinate  $\mathbf{Y}_L = [x_k \ y_k \ z_k \ 1]^T$ 은 이미지 평면의 한 점  $\mathbf{Y}_I = [u_l \ v_l \ 1]^T$ 로 매칭된다. [Fig. 2(a)]는 센서 융합을 통해 3차원 포인트 클라우드를 영상에 투영시켜 얻은 RGB와 Range 입력 영상을 보여준다.

### 3.2 네트워크 구조

본 연구에서 기본적으로 사용되는 심층 컨벌루션 신경망의 구조는 GoogLeNet<sup>[15]</sup>이다. GoogLeNet은 PoseNet<sup>[4]</sup>에서 자세 추정에 사용되는 pre-trained 네트워크 모델이며 22개의 컨벌루션 레이어와 9개의 inception module로 구성된다. GoogLeNet은 영상 인식과 같은 분류(Classification) 작업 수행에 많이 사용되는데 위치 재인식을 수행하기 위해 다음과 같은 변화가 존재한다. 분류를 하는 Fully-connected 레이어를 regression을 수행하는 레이어로 수정한다. 기존 GoogLeNet에 존재하는 3개의 softmax 분류기를 affine regressor로 대체한다. 학습 단계에서 이런 중간 단계의 affine regressor에 대한 loss를 포함하며 테스트 단계에서는 제일 마지막의 regressor만 사용한다.

RGB 영상만을 사용하는 기존 다른 연구와는 달리 3차원 포인트를 이용한 거리 정보도 사용하기 때문에 [Fig. 3]과 같이 입력은 네 개의 채널이 된다. 입력을 위해 취득 이미지 데이터

를 455x256 픽셀로 resize 시키고 가운데에서 224x224 픽셀을 crop하여 사용한다. 추정하고자 하는 자세  $\mathbf{P} = [\mathbf{x} \ \mathbf{q}]^T$ 는 3차원 위치(position)  $\mathbf{x}$ 와 방향(orientation) 정보를 나타내는 쿼터니언(Quaternion)  $\mathbf{q}$ 로 이루어져 있다. 이미지에 맞는 pose regressor를 학습시키기 위해 식 (3)의 Euclidean loss를 사용한다<sup>[4]</sup>.

$$L_I = \|\hat{\mathbf{x}} - \mathbf{x}\|_2 + \beta \left\| \hat{\mathbf{q}} - \frac{\mathbf{q}}{\|\mathbf{q}\|} \right\|_2 \quad (3)$$

$\beta$ 는 위치 오차와 방향 오차의 중요도를 결정시키는 scale factor로서 위치와 방향 중 학습에 영향을 끼치는 정도를 결정한다<sup>[4]</sup>.  $\beta$ 가 큰 값일수록 방향 성분 오차의 중요도가 증가한다. 이렇게 학습된 네트워크에 테스트 영상을 입력하면 pose regressor의 결과로 센서의 절대 pose를 나타내는  $\hat{\mathbf{p}}_t$ 가 생성된다.  $t$ 는 시간 index를 나타낸다.

### 3.3 위치 회귀 모델 및 파티클 필터 기반 위치인식

CNN 구조를 사용하는 학습 기반 위치 재인식은 현재영상을 입력으로 받아 해당 영상이 취득된 센서의 자세를 출력으로 제공해준다. 따라서, 현재 영상의 추정에 이전 결과값이 사용되지 않아 종종 큰 오차를 보여주기도 한다. 이런 문제를 해결하고자 이전 시간에서의 결과를 포함하는 파티클 필터를 사용하여 위치를 보정한다. 파티클 필터는 motion model에 따라 다음 스텝에서의 파티클의 상태를 결정하는 prediction 단계, 측정값에 따라 파티클의 중요도를 결정하는 correction 단계, 그리고 resampling 단계로 이루어져 있다<sup>[16]</sup>. Resampling은 파티클의 중요도에 따라 다음 단계의 파티클을 확률적으로 뽑는다. Bayesian framework에서의 상태 추정은 식 (4)와 같다.

$$p(\mathbf{x}_t | \mathbf{u}_t, \mathbf{z}_{t-1}) = \int p(\mathbf{x}_t | \mathbf{u}_t, \mathbf{x}_{t-1}) p(\mathbf{x}_{t-1} | \mathbf{z}_{t-1}, \mathbf{u}_{t-1}) d\mathbf{x}_{t-1} \quad (4)$$

$$p(\mathbf{x}_t | \mathbf{z}_t) = \frac{1}{\alpha} p(\mathbf{z}_t | \mathbf{x}_t) p(\mathbf{x}_t | \mathbf{u}_t, \mathbf{z}_{t-1})$$

여기서  $\alpha$ 는 normalized coefficient이며 식 (4)는 두 개의 확률 모델을 포함하고 있다.  $p(\mathbf{z}_{t+1} | \mathbf{x}_{t+1})$ 는 measurement 모델,  $p(\mathbf{x}_{t+1} | \mathbf{u}_t, \mathbf{x}_t)$ 는 motion 모델을 나타낸다. Motion 모델은 control 입력  $\mathbf{u}_t$ 이 주어졌을 때, 현재 상태  $\mathbf{x}_t$ 의 확률을 이전 스텝의 상태  $\mathbf{x}_{t-1}$ 로부터 표현한다. Motion 모델에서 사용되는  $\mathbf{u}_t$ 는 constant velocity motion으로 가정했다. 파티클 필터에서 상태  $\mathbf{x}_t$ 는 파티클  $\mathbf{x}_t^{(j)} = \langle \mathbf{x}_t^{(j)}, \omega_t^{(j)} \rangle$ ,  $j=1, \dots, N_p$ 로 표현되며 앞서 설명한 pose regression 결과를 measurement로 사용한다. measurement 모델에 따른 파티클 업데이트는 식 (5)-(7)와 같다.

[Algorithm 1] Proposed relocalization algorithm

1:	<b>Relocalization</b> $\left( \left\{ \mathbf{x}_{t-1}^{(j)} \right\}_{j=1:N_p}, \mathbf{y}_{L,t}, \mathbf{I}_t \right)$ :
2:	Generate RGBD image using (2)
3:	<b>if</b> $t == 0$
4:	Initialize $\mathbf{x}_0 = \hat{\mathbf{p}}_0$
5:	<b>else</b>
6:	Calculate $\hat{\mathbf{p}}_t$
7:	Update Particle Filter (5)-(7)
8:	<b>return</b> $\left\{ \mathbf{x}_t^{(j)} \right\}$

$$\mathbf{z}_t = \hat{\mathbf{P}}_t + v_t \quad (5)$$

$$\omega_t^{(j)} \propto p(\mathbf{z}_t | \mathbf{x}_t^{(j)}) \cdot \omega_{t-1}^{(j)} \quad (6)$$

$$\propto \exp\left(-\frac{\|\hat{\mathbf{P}}_t - \mathbf{x}_t^{(j)}\|^2}{2\sigma^2}\right) \cdot \omega_{t-1}^{(j)}$$

$$\mathbf{x}_t^{(j)} \leftarrow \text{draw } i \propto \omega_t^{(j)} / \sum_j \omega_t^{(j)} \quad (7)$$

여기서  $\hat{\mathbf{P}}_t$ 는 pose regression의 추정 결과값,  $v_t$ 는 measurement noise를 나타내며, 식 (7)은 resampling 과정이다.

[Algorithm 1]은 본 논문에서 제안된 위치 재인식 알고리즘이며 라인 1을 통해 학습된 CNN에 입력될 영상을 생성한다. 입력 영상을 이용하여 CNN 결과값을 얻고 파티클 필터의 measurement로 사용하여 업데이트 한다.

## 4. 실험 및 결과

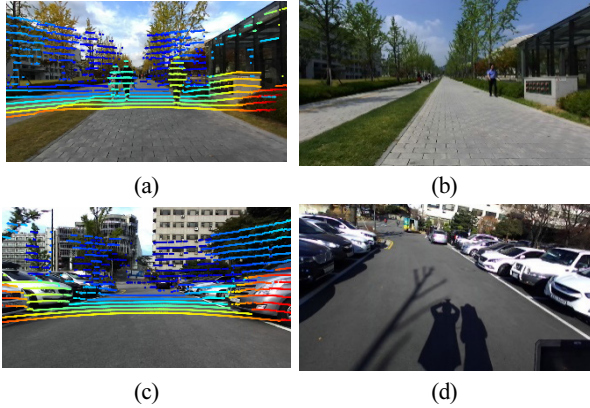
본 장에서는 제안한 위치인식 기술의 성능 평가를 위해 실외용 로봇 플랫폼에서 다양한 환경의 데이터를 취득하였다. 본 논문에서 사용된 GoogLeNet 기반 심층 신경망과 제안된 알고리즘은 텐서플로우(Tensorflow)로 구현되었다. 실험에 사용된 컴퓨팅 환경은 Intel Core i7-7700HQ @2.8Ghz의 CPU와 NVIDIA Geforce GTX 1070 8G의 GPU가 탑재된 노트북이다.

### 4.1 데이터셋

실험에 사용된 데이터셋은 서로 다른 두 환경에서 취득하였다. 두 환경 모두 연세대학교 캠퍼스 내 실외 환경에서 로봇 플랫폼을 이용하여 취득하였으며 사용된 3차원 레이저 스캐너는 Velodyne VLP-16, 이미지는 ZED stereo 카메라의 왼쪽 영상을 사용하였다. Velodyne VLP-16 센서는 초당 약 300,000 points를 측정하며 측정 최대 거리는 약 100 m이다. ZED stereo

[Table 1] Dataset description

Dataset	Frames			Spatial Extent (m)
	Train	Test1	Test2	
Baekyang	5526	4912	4621	40 x 160
Eng. Bldg.	5329	3432	3679	80 x 80



[Fig. 4] Sensor fusion of a 3D laser scanner and camera: calibrated sample data. (a) Baekyang dataset (Test1), (b) Baekyang dataset (Test2), (c) Engineering Building dataset (Test1), (d) Engineering Building dataset (Test2)

카메라의 경우 1280 x 720 해상도와 30 fps 모드로 사용하였다. 투영 단계에서 이미지 평면 앞쪽에 있는 포인트들은 걸러내어 평면에서 현재 카메라 영상에 투영될 수 있는 포인트만 사용한다.

첫 번째 데이터셋은 연세대학교 백양로(Baekyang)에서 취득하였으며 보행자가 다수 존재하는 환경이다. 학습에 사용된 영상은 총 5526 프레임이다. 테스트에 사용된 데이터셋은 총 2개이며 환경변화에 강한 특성을 확인하기 위해 각각 가을과 여름에 취득했다. 첫 번째 테스트에 사용된 영상은 총 4912 프레임이고 두 번째 테스트에 사용된 영상은 총 4621 프레임이며 시험 공간의 크기는 약 40 m x 160 m이다. 두 번째 데이터셋은 연세대학교 공학관(Eng. Bldg.)에서 취득하였고 비교적 동적 객체가 적은 환경이다. 학습에는 총 5329 프레임이 사용되었으며 첫 번째 테스트는 3432 프레임, 두 번째 테스트는 3679 프레임이 사용되었다. 첫 번째 테스트 데이터셋은 가을, 두 번째 테스트 데이터셋은 겨울에 취득하였다. 시험 공간의 크기는 약 80 x 80 m이며 데이터셋에 대한 정보는 [Table 1]에 정리했다.

[Fig. 4(a)]와 [Fig. 4(c)]는 센서 융합 시스템으로부터 취득된 입력 센서 정보를 보여주며, 3차원 포인트 클라우드를 해당 이미지로 투영시켜 거리 별로 표현한 결과이다. [Fig. 4(b)]와 [Fig. 4(d)]는 각각 다른 계절에 취득된 데이터로서 해당 계절과 다른 계절의 데이터셋으로 학습된 CNN의 테스트에 사용되었다. [Fig. 4(a)]와 [Fig. 4(b)], [Fig. 4(c)]와 [Fig. 4(d)]는 같은 위치에서의 샘플 테스트 이미지를 보여준다. 학습에 사용

된 데이터셋의 ground truth는 고정밀 IMU, wheel odometry, 3차원 레이저스캐너와 고정밀 지도 기반의 위치추정 알고리즘을 사용하여 얻는다. 학습에 사용된 식 (3)의  $\beta$ 는 500, 파티클 필터에서 사용된  $\sigma$ 는 0.1로 실험을 진행하였다. 학습에 걸리는 시간은 iteration 당 약 0.55초가 소요되며, 테스트 단계에서는 약 0.0059초가 소요된다.

#### 4.2 위치인식 결과

[Fig. 5]는 학습 및 테스트 데이터셋과 함께 PoseNet 및 제안된 방법의 위치인식 결과를 나타낸 것이다. 학습 및 테스트 데이터셋의 센서 경로는 [Fig. 5]의 왼쪽 그림과 같으며 오른쪽은 PoseNet과 제안된 방법의 위치인식 결과를 위성사진과 함께 보여준다.

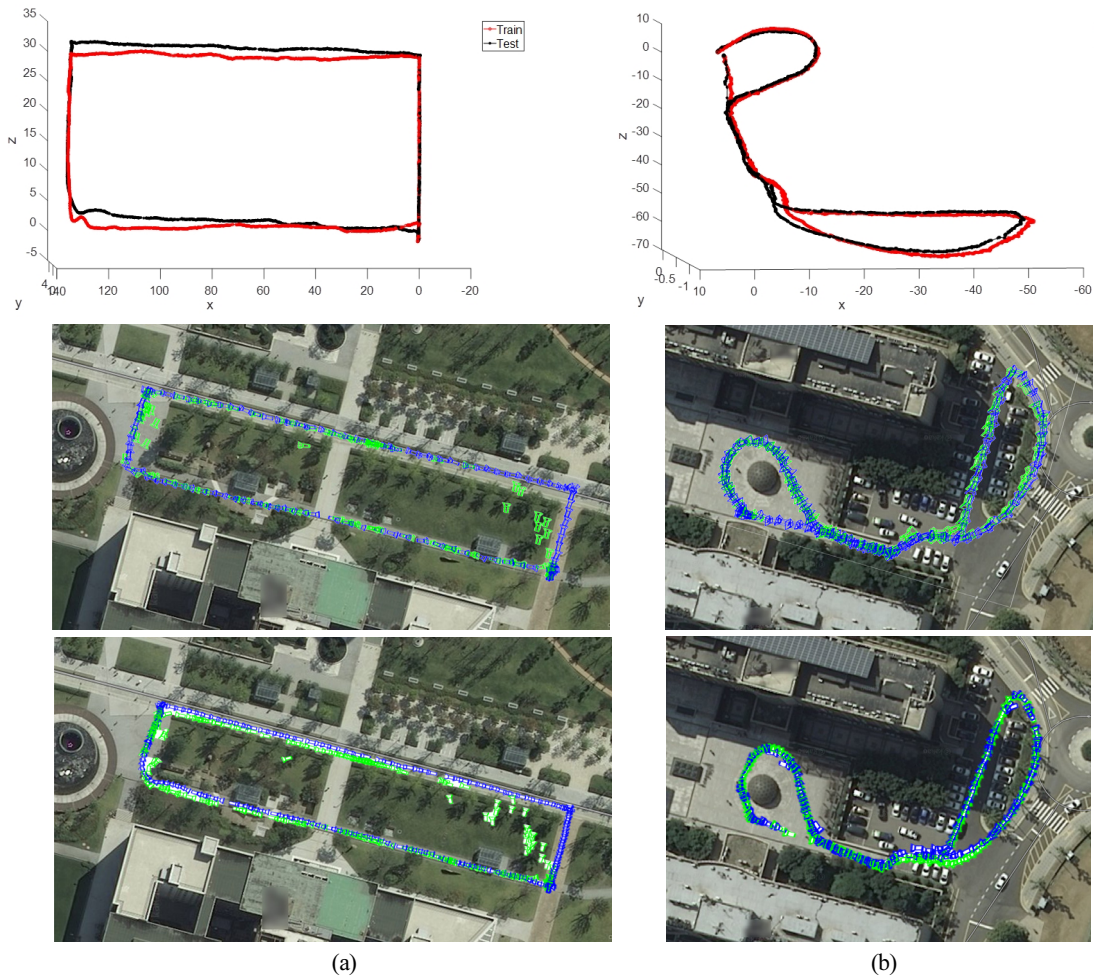
정량적인 위치인식 결과는 [Table 2]와 같다. [Table 2]에서 PoseNet과 본 논문에서 성능 향상을 위해 추가한 Depth 정보 사용, 파티클 필터에 대해 개별적으로 사용했을 때의 성능 또한 포함하고 있다. 상대적으로 depth 정보를 사용할 때는 rotation, 파티클 필터를 사용할 때 translation 오차를 감소 효과가 큰 것을 알 수 있다. 기존 이미지만을 사용한 PoseNet은 방향 성분보다 위치 성분에서 오차가 크게 발생한다. 이는 지속적으로 비슷한 영상이 취득되는 직선 주행 중 이미지 정보만으로는 위치를 정확히 추정할 수 없기 때문이라고 해석 가능하다. 제안된 방법의 경우 거리 정보를 포함하기 때문에 위치 오차를 크게 줄여주는 역할을 해줄 수 있다. 또한 [Fig. 5]에서 볼 수 있듯이 PoseNet의 경우 연속된 센서 취득 정보를 사용하지 않기 때문에 이전 결과값과 많이 차이는 경우가 자주 발생한다. 반면, 제안된 학습 결과 기반 파티클 필터를 통해 이러한 문제를 해결하였다.

각 알고리즘 모두 동적 객체(보행자)가 많은 환경(Baekyang)에서 위치인식 오차가 3.08 m로 매우 큰 것을 알 수 있다. 반면, PoseNet 보다 제안된 방법이 위치인식 오차가 적으며 특히 방향 성분보다 위치 성분의 오차 감소 비율이 더 크다. 이는 앞서 언급된 반복되는 지형에 대한 위치인식 오차를 거리 센서 데

[Table 2] Localization results

	Baekyang Test 1		Baekyang Test 2		EngC Test 1		EngC Test 2	
	Error XYZ	Error Q	Error XYZ	Error Q	Error XYZ	Error Q	Error XYZ	Error Q
PoseNet	3.08	1.74	3.49	1.79	1.06	1.73	1.87	1.89
PoseNet+Depth	2.97	1.66	3.34	1.66	1.02	1.69	1.55	1.87
PoseNet+PF	2.02	1.62	1.78	1.69	0.99	1.71	1.54	1.87
Proposed	<b>1.25</b>	<b>1.54</b>	<b>1.56</b>	<b>1.61</b>	<b>0.87</b>	<b>1.68</b>	<b>1.26</b>	<b>1.78</b>





[Fig. 5] Relocalization results using our proposed method. The first row shows trajectory of training (red dot) and test (black asterisk) datasets. The second and third row show trajectory of PoseNet (green) and proposed method (blue) of Test1 and Test2 respectively. (a) Baekyang dataset, (b) Engineering Building dataset

이터를 이용하여 상쇄시킬 수 있기 때문이다. 또한, 데이터셋 별로 위치인식 결과를 보면 학습 데이터와 다른 계절에 취득한 데이터의 경우(Test 2) 전체적으로 오차가 크다. RGB 영상을 사용하는 visual relocalization 방식을 사용하면서 이를 보완하기 위해 추후 연구에 효율적인 depth 정보 encoding 방법을 사용하려 한다.

## 5. 결 론

본 논문에서는 3차원 레이저 스캐너와 비전 센서를 융합한 센서 시스템을 통해 심층 컨벌루션 신경망의 학습 및 파티클 필터 기반 위치인식 방법을 제안하였다. 심층 컨벌루션 신경망의 학습에는 카메라 영상의 RGB 정보뿐만 아니라 주변 환경의 거리 정보도 포함한다. 센서 융합을 통해 생성된 입력으로부터 센서의 자세를 regression하는 CNN 구조를 end-to-end로 학습시킨다. 또한, 이동 로봇이 주행하면서 얻는 연속적인

센서 입력을 모두 사용하는 파티클 필터(Particle Filter)를 통해 더욱 정밀한 위치인식을 수행한다. 제안된 방법은 실외 환경에서 로봇 플랫폼으로부터 취득된 데이터셋으로 검증하였다.

추후 연구로는 향상된 네트워크 구조를 학습하여 조명, 날씨 등 다양한 환경 변화가 존재하는 환경에서도 동작 가능한 위치 재인식 방법에 대한 연구를 수행하고자 한다. 또한, 네트워크 학습에 필요한 파라미터를 검색하는 방식을 도입하여 위치인식 성능을 높이고, 파티클 필터를 통한 위치인식이 심하게 실패한 경우 오차를 극복하는 연구를 진행하고자 한다.

## References

- [1] R. Li, Q. Liu, J. Gui, D. Gu, and H. Hu, "Indoor Relocalization in Challenging Environments With Dual-Stream Convolutional Neural Networks," *IEEE Transaction on Automation Science and Engineering*, vol. 15, no. 2, pp. 651-662, Apr. 2018.

- [2] J. L. Schonberger and J.-M. Frahm, "Structure-from-Motion Revisited," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4104-4113, Jun. 2016.
- [3] D. Nistér and H. Stewénus, "Scalable recognition with a vocabulary tree," *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, New York, NY, USA, 2006, DOI: 10.1109/CVPR.2006.264.
- [4] 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 (ICCV)*, Santiago, Chile, pp. 2938-2946, 2015.
- [5] A. Kendall and R. Cipolla, "Modelling uncertainty in deep learning for camera relocalization," *2016 IEEE International Conference on Robotics and Automation (ICRA)*, Stockholm, Sweden, pp. 4762-4769, 2016.
- [6] S. Thrun, D. Fox, W. Burgard, and F. Dellaert, "Robust Monte Carlo localization for mobile robots," *Artificial Intelligence*, vol. 128, no. 1-2, pp. 99-141, May. 2001.
- [7] R. Arandjelović, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN Architecture for Weakly Supervised Place Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 6, pp. 1437-1451, Jun., 2018.
- [8] J. Shotton, B. Glocker, C. Zach, S. Izadi, A. Criminisi, and A. Fitzgibbon, "Scene coordinate regression forests for camera relocalization in RGB-D images," *2013 IEEE Conference on Computer Vision and Pattern Recognition*, Portland, OR, USA, pp. 2930-2937, 2013.
- [9] E. Brachmann, A. Krull, S. Nowozin, J. Shotton, F. Michel, S. Gumhold, and C. Rother, "DSAC - Differentiable RANSAC for Camera Localization," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017, DOI: 10.1109/CVPR.2017.267.
- [10] Z. Chen, O. Lam, A. Jacobson, and M. Milford, "Convolutional neural network-based place recognition," *arXiv:1411.1509 [cs.CV]*, 2014.
- [11] R. Clark, S. Wang, A. Markham, N. Trigoni, and H. Wen, "VidLoc: A Deep Spatio-Temporal Model for 6-DoF Video-Clip Relocalization," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, pp. 6856-6864, 2017.
- [12] F. Walch, C. Hazirbas, L. Leal-Taixé, T. Sattler, S. Hilsenbeck, and D. Cremers, "Image-based localization using LSTMs for structured feature correlation," *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, pp. 627-637, 2017.
- [13] M. Velas, M. Spanel, Z. Materna, and A. Herout, "Calibration of RGB Camera With Velodyne LiDAR," *WSCG 2014*, pp. 135-144, 2014.
- [14] H. Jo, H. M. Cho, S. Lee, and E. Kim, "Multi-Resolution Point Cloud Generation Based on Heterogeneous Sensor Fusion System," *2017 14th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI)*, Jeju, South Korea, pp. 886-888, 2017.

- [15] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, 2015, DOI: 10.1109/CVPR.2015.7298594.
- [16] H. Jo, H. M. Cho, S. Jo, and E. Kim, "Efficient Grid-Based Rao-Blackwellized Particle Filter SLAM With Interparticle Map Sharing," *IEEE Transactions on Mechatronics*, vol. 23, no. 2, pp. 714-724, Apr. 2018.



### 조 형 기

2012 연세대학교 전기전자공학부(학사)

2012~현재 연세대학교 전기전자공학부 박사과정 재학중

관심분야: 이동로봇 자율주행, Visual SLAM, Sensor Fusion

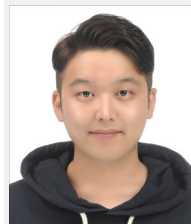


### 조 해 민

2015 연세대학교 전기전자공학부(학사)

2015~현재 연세대학교 전기전자공학부 박사과정 재학중

관심분야: Visual SLAM, 로봇 인공지능



### 이 성 원

2016 연세대학교 전기전자공학부(학사)

2016~현재 연세대학교 전기전자공학부 박사과정 재학중

관심분야: 로봇 인공지능, Place Recognition



### 김 은 태

1992 연세대학교 전기전자공학과(공학사)

1994 연세대학교 전기전자공학과(공학석사)

1999 연세대학교 전기전자공학과(공학박사)

2002~현재 연세대학교 전기전자공학부 교수

관심분야: Computational Intelligence, 지능형 로봇, Deep Learning, Autonomous Vehicle