동시적 위치 추정 및 지도 작성에서 Variational Autoencoder 를 이용한 루프 폐쇄 검출

신동원, 호요성
광주과학기술원 전기전자컴퓨터공학부
{dongwonshin, hoyo}@gist.ac.kr

Loop Closure Detection Using Variational Autoencoder in Simultaneous Localization and Mapping

Dong-Won Shin and Yo-Sung Ho
Gwangju Institute of Science and Technology

요 약

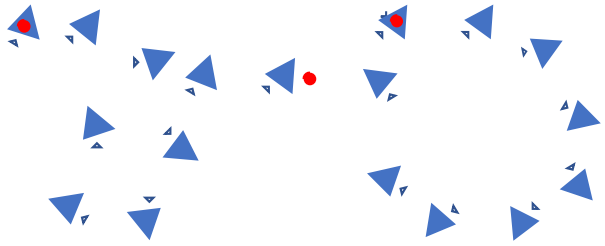
본 논문에서는 동시적 위치 추정 및 지도 작성 (simultaneous localization and mapping)에서 루프 폐쇄 검출을 딥러닝 방법의 일종인 variational autoencoder 를 이용하여 수행하는 방법에 대해 살펴본다. Autoencoder 는 비감독 학습 방법의 일종으로 입력 영상이 신경망을 통과하여 얻은 출력 영상과 동일하도록 신경망을 학습시키는 모델이다. 이 때 autoencoder 중간의 병목 지역을 통과함에도 불구하고 입력과 동일한 영상을 계산해야 하는 제약조건이 있기 때문에 이는 차원 축소나 데이터 추상화의 목적으로 많이 사용된다. 여기서 한 단계 더 발전된 variational autoencoder 는 기존의 autoencoder 가 가진 단점인 입력 변수의 분포와 잠재 변수의 분포 사이에 상관관계가 없다는 단점을 해결하기 위해 Kullback-Leibler divergence 를 활용한 손실 함수를 정의하여 사용했다. 실험결과에서는 루프 폐쇄 검출에서 많이 사용되는 City-Centre 와 New College 데이터 집합을 사용하여 평가하였으며 루프 폐쇄 검출의 결과는 정밀도와 재현율을 계산하여 나타냈다.

1. 서론

동시적 위치 추정 및 지도 작성 (simultaneous localization and mapping, SLAM)이란 카메라와 같은 센서를 가진 로봇의 주변 환경을 3 차원 모델로 복원함과 동시에 3 차원 공간상에서 로봇의 위치를 추정하는 기술을 말한다. 정확한 3 차원 환경 맵과 로봇의 위치를 예측하는 것은 증강현실, 로보틱스, 자율주행과 같은 응용에서 필수적이다. 첫번째로 증강현실이란 가상의 객체를 사용자가 바라보고 있는 실제 공간에 합성하여 실제로 존재하는 것처럼 느끼도록 하는 기술인데 이는 3 차원 환경 맵의 정확한 복원과 사용자가 바라보는 시점의 정확한 예측이 이루어져야 비로소 가능하다. 두번째로 로보틱스에서는 로봇이 특정한 작업(물체 이동, 분류, 수거 등)을 성공적으로 수행하도록 하기 위해서 로봇 주변의 환경 맵과 로봇의 위치가 반드시 필요하다. 세번째로 자율주행에서는 운송수단이 사고없이 안전하게 목적지로 탑승자를 이동시키기 위해 SLAM 기술이 활용될 수 있다. 그 외에도 더 많은 응용에서 활용이 가능하며 전단부에서 입력되는 센서(vision, depth, Lidar, fusion)의 종류에 따라 구분되기도 한다.

전형적인 SLAM 알고리즘은 크게 전단부(front-end)와 후단부(back-end)로 나뉘어진다 [1]. 전단부는 센서로부터 측정된 데이터를 3 차원 점군으로 만들고 정합하는 과정을 담당하며 후단부는 루프 폐쇄 검출, 변형 처리, 환경 맵 최적화 등의 과정을 수행한다. 본 논문에서는 전반적인 SLAM 알고리즘의 후단부에서 중요한 부분을 차지하고 있는 루프 폐쇄 검출에 대해 연구를 수행했다.

루프 폐쇄 검출(loop closure detection)이란 로봇의 이동 궤적상에서 현재의 위치가 이전에 방문했던 위치인지를 판단하는 것으로 검출된 결과를 환경 맵 최적화 단계에서 제약조건으로 활용하도록 하여 SLAM 알고리즘의 로봇 표류 문제를 해결한다. 그림 1 에서 보이는 바와 같이 삼각형이 일련의 로봇 궤적이라고 가정하고 동그란 점이 실제로는 동일한 위치라고 하면 그림 1(a)의 루프 폐쇄 검출을 사용하지 않은 경우 누적된 궤적 오차 때문에 예측된 로봇의 궤적이 심하게 뒤틀리는 것을 확인할 수 있다. 이는 환경 맵의 정확성에도 직접적으로 영향을 미친다. 반대로 그림 1(b)의 루프 폐쇄 검출을 사용한 경우 궤적이 제대로 수정되어 동시에 정확한 환경 맵을 획득할 수 있다.



(a) 루프 폐쇄 검출을 사용하지 않은 경우　(b) 루프 폐쇄 검출을 사용한 경우

그림 1. 루프 폐쇄 검출 예제

2. 관련 연구

기존의 루프 폐쇄 검출 방법은 크게 지역적 영상 특징 기반 방법과 전역적 영상 특징 기반 방법으로 나뉜다. 먼저 지역적 영상 특징 방법은 영상에서 SIFT 또는 SURF 와 같은 지역적 영상 특징을 추출한 다음 이를 기반으로 영상 설명자를 생성한다. 각 영상으로부터 얻은 영상 설명자 간의 거리를 계산하여 일정 문턱치보다 낮으면 현재 위치가 이전에 방문했던 것으로 판단하여 루프 폐쇄를 검출한다.

대표적인 방법으로는 bag-of-visual-words 방법이 있는데 이 방법에서는 추출한 지역적 영상 특징들을 다차원의 특징 공간에서 군집화 하고 군집의 중심을 Visual-Word 라고 정의한다. 다양한 영상들로부터 획득한 Visual-Word 들의 모음을 bag-of-visual-words 라고 정의한다. 이렇게 학습된 bag-of-visual-words 를 활용하여 영상 설명자를 획득하고자 하는 영상이 들어오면 학습 단계에서 사용된 동일한 지역적 영상 특징 검출기로 특징을 추출한다. 그리고 벡터 양자화 단계를 거치고 visual-word 들의 히스토그램을 계산하여 이를 영상 설명자로 사용한다 [2].

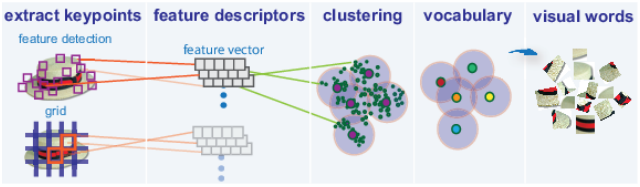


그림 2. Bag-of-Visual-Words 방법

전역적 영상 특징 기반 방법은 영상을 물체나 지역적인 특징과 같은 구성 요소로 분석하는 것이 아니라 전체 영상 그 자체를 활용하여 영상 설명자를 획득하는 방법이다.

최근에는 convolutional neural network (CNN) 방법을 활용한 전역적 영상 특징 기반 루프 폐쇄 검출 방법이 제안되었다 [3]. 이 방법에서는 이미지 분류에 사용되는 ImageNet 신경망 모델을 사용하여 각 레이어에서 얻은 특징을 영상 설명자로 활용하여 루프 폐쇄를 검출한다 [4]. 신경망 모델의 구조는 기존의 ImageNet 을 그대로 사용하되 학습 데이터는 장소 인식에 중점을 둔 Places 데이터 집합을 활용하여 신경망을 학습시킨다 [5]. 그런 다음 각각의 convolution, pooling, fully Connected 레이어에서 추출된 영상 설명자를 이용하여 루프 폐쇄를 검출하는데 [3]의 논문의 실험결과에서는 마지막 Pooling 레이어에서 얻은 영상 설명자가 가장 좋은 루프 폐쇄 검출 결과를 나타낸다는 것을 확인했다.

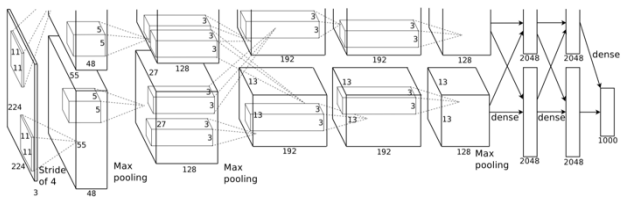


그림 3. ImageNet 신경망 모델

3. 제안하는 방법

제안하는 방법은 비감독 학습법의 일종인 variational autoencoder 를 사용하여 모델을 학습한 다음 영상의 전역적 영상 특징을 추출하고 이를 이용하여 루프 폐쇄를 검출하는 방법을 취한다. 먼저 3.1 절에서 autoencoder 에 대해서 간략하게 설명하고 3.2 절에서 variational autoencoder 에 대해서 소개한다.

3.1 autoencoder

간단한 구조의 autoencoder (AE)는 그림 4 와 같이 입력 레이어와 잠재 레이어, 출력 레이어 간의 완결 연결 구조 (Fully connected)로 구성되어 있다 [6]. 입력 레이어와 잠재 레이어 간의 가중치(W) 연결 구조를 encoder 라고 정의하며 잠재 레이어와 출력 레이어 간의 가중치(W') 연결구조를 decoder 라고 정의한다.

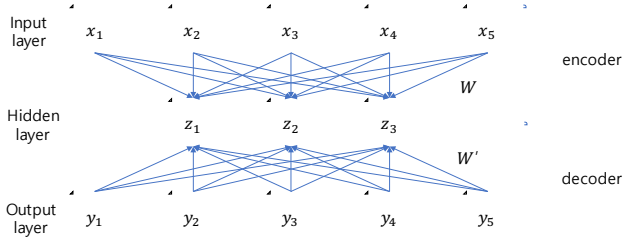


그림 4. Autoencoder 구조

AE 모델의 목표는 입력 변수 x_i와 출력 변수 y_i사이의 차이를 최소화하는 신경망 모델을 학습하는데 있다. 다시 말해 입력 그 자체를 출력으로 내놓는 신경망 모델을 학습하는 것이다. 이 모델이 의미 있는 이유는 중간의 병목 부분인 잠재 레이어의 차원이 입력 레이어의 차원보다 훨씬 적다는데 있다. 다시 말해 입력 변수를 압축하여 잠재 변수 z_i를 만들고 압축된 잠재 변수를 이용하여 원래의 입력과 동일한 출력을 계산하는 신경망을 학습한다. 따라서 AE 는 입력 변수를 압축 및 추상화하는데 탁월한 능력이 있어서 추상화된 잠재 변수를 활용하여 루프 폐쇄 검출을 수행할 수 있다.

3.2 variational autoencoder

기존의 AE 는 입력 변수를 추상화 하는데 좋은 성능을 가지고 있지만 입력 변수의 분포와 잠재 변수의 분포 사이에 상관관계가 없다는 단점이 있다. 우리는 잠재 변수를 비교하여 루프 폐쇄 검출을 수행하는데 예를 들어 잠재 변수 사이의 거리가 가까운데도 불구하고 입력 변수 사이의 거리가 멀다면 루프 폐쇄 검출 능력이 떨어질 것이다. 이 단점을 극복하기 위해 나온 것이 variational autoencoder (VAE)이다 [7].

AE 와 VAE 의 가장 큰 차이점은 손실 함수가 다르다는 점인데 수식 (1)은 VAE 의 손실 함수 $l_i(\theta,\phi)$를 나타낸다. 3.1 절에서 설명한 입력 변수 x_i, 출력 변수 y_i, 잠재 변수 z_i의 표기법을 그대로 사용했으며 $q_\phi(z|x)$와 $p_\theta(x|z)$는 각각 encoder 와 decoder 의 확률분포를 나타낸다. 그리고 ϕ와 θ는 각각 encoder 와 decoder 의 확률분포 모델 파라미터를 나타낸다.

$$l_i(\theta,\phi) = -E_{z\sim q_\phi(z|x_i)}[\log p_\theta(x_i|z)] + KL(q_\phi(z|x_i)||p(z)) \qquad (1)$$

손실 함수의 첫번째 항은 재생성 손실항을 나타내며 decoder 가 생성한 출력이 입력과 다르면 데이터를 잘 재생성하지 못한 것으로 판단하여 손실 값은 높아진다. 손실 함수의 두번째 항은 Kullback-Leibler divergence 를 사용한 정규화항을 나타내며 encoder 의 확률 분포가 잠재 변수의 확률 분포와 유사하도록 제약 조건을 할당한다. 이 때 일반적으로 잠재 변수의 확률 분포는 평균 값이 0 이고 분산 값이 1 인 표준 정규 분포를 가진다. 이 손실 함수를 만족시키는 신경망 모델을 학습시키는 것이 VAE 의 목적이다.

3.3 variational autoencoder 모델 구조

본 논문에서 사용한 신경망 모델의 구조는 그림 5 와 같다. 입력으로 들어가는 영상의 크기는 128×128×3 (가로 × 세로 × 채널)이다. Encoder 에서는 4 개의 convolutional layer 와 1 개의 fully connected layer 가 포함되어 있고 decoder 에서는 4 개의 deconvolutional layer 와 1 개의 fully connected layer 가 포함되어 있다. Convolutional 필터와 deconvolutional 필터의 크기는 모두 5×5이며 strides 는 2×2이다. 정중앙의 잠재 레이어에서 노드의 갯수는 50 개로 정의했다. 학습을 위해 구글에서 제공하는 딥러닝 라이브러리인 Tensorflow 를 활용했다 [8].

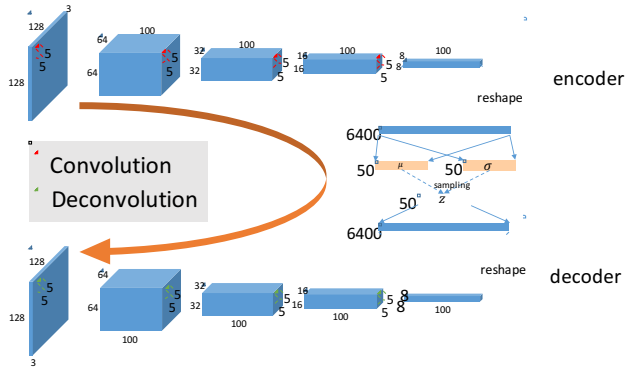


그림 5. 제안하는 variational autoencoder 모델 구조

3.4 학습 데이터 집합

신경망의 학습에는 학습 데이터의 종류와 양도 중요하게 고려된다. 본 연구는 루프 폐쇄 검출을 위한 장소 인식에 목적이 있기 때문에 장소와 관련한 데이터가 주를 이루는 Places 데이터 집합을 사용했다 [9]. Places 데이터 집합에는 약 140 만 장의 장소에 관한 다양한 사진이 있고 장소의 종류별로 분류가 되어있다. 그림 5 는 Places 데이터 집합에서 얻은 영상의 일부를 보여준다.

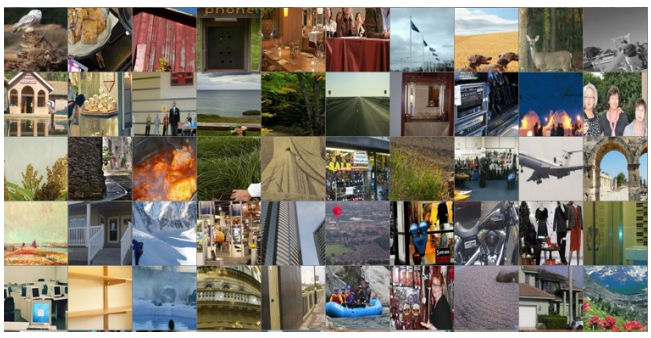


그림 6. Places 학습 데이터 집합

4. 실험

4.1 실험 방법

루프 폐쇄 검출의 실험에는 같은 장소를 다른 위치 또는 다른 시간대에 촬영한 데이터가 필요하다. 본 논문에서는 루프 폐쇄 검출 실험에 많이 사용되는 City Centre 와 New College 데이터를 사용했다 [10]. 이 데이터 집합에는 스테레오 카메라가 장착된 로봇이 특정한 경로를 지나가면서 일정한 시간 간격마다 촬영된 사진들이 포함되어 있다. 영상의 갯수는 City Centre 데이터 집합에 모두 2474 장이 있고 New College 데이터 집합은 모두 2146 장의 사진이 포함되어 있다. 해당 경로에는 방문했던 곳을 다시 방문하는 루프 폐쇄 부분도 포함되어 있고 루프 폐쇄 정보에 대한 Ground-truth 정보도 이진 행렬 형태로 제공하므로 루프 폐쇄 검출 능력을 평가하기에 적합하다.

각각의 데이터 집합 $D \in \{City\ Centre, New\ College\}$ 의 영상을 학습된 VAE 에 입력하여 얻은 잠재 변수를 영상 설명자로 사용한다. 이 때 $i \in D$ 번째 영상과 $j \in D$ 번째 영상을 입력하여 얻은 영상 설명자 사이의 L2-norm 값을 계산해서 특정한 문턱치 값 보다 작으면 루프 폐쇄로 판단하도록 했다. 본 실험에서는 문턱치 값을 다양하게 조절하여 제안하는 방법으로 찾은 결과 중에서 실제 루프 폐쇄의 숫자를 측정했다.

4.2 실험 결과

첫번째로 City Centre 실험 데이터 집합에 대해 얻은 결과를 표 1 에 나타냈다. 문턱치 값 (threshold)은 {5, 5.5, 6, 6.5, 7} 의 값을 적용했고 루프 폐쇄 인식 결과는 인식 분야에서 많이 사용되는 true positive (TP), false positive (FP), false negative (FN)로 나타냈다. 간단히 설명을 하자면 True Positive 는 실제 루프 폐쇄인 영상을 루프 폐쇄인 것으로 잘 판단한 것이고 false positive 는 실제 루프 폐쇄가 아닌데 루프 폐쇄인 것으로 잘못 판단한 것이다. 마지막으로 false negative 는 실제 루프 폐쇄인데 루프 폐쇄가 아닌 것으로 잘못 판단한 것이다.

표 1. City Centre 실험 데이터 집합에서의 인식 결과

threshold	5	5.5	6	6.5	7
true positive	42	149	465	1088	2103
false positive	6877	22159	60049	134196	254352
false negative	16938	16831	16515	15892	14877

두번째로 New College 실험 데이터 집합에 대해 얻은 결과를 표 2 에 나타냈다.

표 2. New College 실험 데이터 집합에서의 인식 결과

threshold	5	5.5	6	6.5	7
true positive	23	106	315	737	1535
false positive	5467	17414	46502	105251	205979
false negative	16957	16874	16665	16243	15445

이렇게 얻은 TP, FP, FN 값을 이용하여 정밀도 (Precision)와 재현율 (Recall) 값을 표 3 과 표 4 에 나타냈다. 정밀도는 알고리즘으로부터 검출된 결과들 중에서 실제 루프 폐쇄의 비율이고 재현율은 실제 루프 폐쇄들 중에서 알고리즘이 루프 폐쇄로 검출한 결과의 비율이다. 정밀도의 계산식은 precision = TP/(TP+ FP)이고 재현율의 계산식은 recall = TP/(TP+ FN) 이다.

표 3. City Centre **데이터 집합에 대한 정밀도와 재현율**

Threshold	5	5.5	6	6.5	7
Precision	0.00607	0.00667	0.00768	0.00804	0.00820
Recall	0.00247	0.00877	0.02738	0.06407	0.12385

표 4. New College **데이터 집합에 대한 정밀도와 재현율**

Threshold	5	5.5	6	6.5	7
Precision	0.00419	0.00605	0.00673	0.00695	0.00740
Recall	0.00135	0.00624	0.01855	0.04340	0.09040

4.3 실험 결과 분석 및 논의

우리는 실험 결과로부터 다음의 내용을 도출할 수 있다. 첫번째로 정밀도는 100%, 즉 1 의 값에 가까울수록 좋은 알고리즘이라고 할 수 있다. 하지만 실험 결과를 살펴보면 1 의 값과는 상당히 거리가 떨어져 있음을 알 수 있다. 문턱치 값을 조절함에 의해서 정밀도를 높일 수 있지만 여전히 좋지 않은 결과를 보인다. 두번째로 재현율도 1 의 값에 가까울수록 좋은 알고리즘이라고 할 수 있다. 재현율 측면에서도 좋지 않은 결과를 보인다.

이러한 실험 결과로부터 우리는 2 가지 시사점을 파악할 수 있다. 첫번째로 VAE 모델 구조와 파라미터의 변경을 통해 인식 성능을 높일 수 있다. 두번째로 전체 영상으로부터 VAE 를 통해 학습한 잠재 변수를 영상 설명자로 사용하기 보다는 지역적 영상 특징 기반 방법과의 결합을 이용해 영상으로부터 지역적 영상 특징을 추출하고 특징 주변의 영상 패치를 획득하여 패치들에 대해 학습한 모델을 활용하여 루프 폐쇄 검출을 수행할 수 있다 [11].

5. 결론

본 논문에서는 SLAM 시스템에서의 루프 폐쇄 검출을 위해 variational autoencoder 를 활용하는 방법에 대해서 연구했다. 장소 인식에 중점을 둔 Places 데이터 집합을 제안하는 VAE 신경망 모델에 입력하여 학습시켰다. 실험에서는 루프 폐쇄 성능의 평가를 위해 많이 사용되는 City Centre 와 New College 데이터 집합을 사용했고 인식 결과로부터 정밀도와 재현율을 측정했다.

제안하는 알고리즘은 정밀도와 재현율의 측면에서 좋지 않은 성능을 보였지만 다음의 2 가지 시사점을 제공한다. 첫번째로 VAE 모델 구조와 파라미터의 변경을 통해 인식 성능을 높일 수 있으며 두번째로 전체 영상으로부터 VAE 를 통해 학습된 잠재 변수를 영상 설명자로 사용하기 보다는 지역적 영상 특징 기반 방법과의 결합을 통해 영상으로부터 지역적 영상 특징을 추출하고 특징 주변의 영상 패치를 획득하여 패치들에 대해 학습한 모델을 활용하여 루프 폐쇄 검출을 수행할 수 있다.

루프 폐쇄 검출에서 장소 인식을 잘 수행하기 위해서는 인식 알고리즘의 생성 능력(generative capability)에 주목해야한다. 같은 장소일지라도 다른 시점, 다른 시간, 다른 조명 아래에서 촬영된 사진은 변화된 정보를 내포하고 있기 때문에 인식 알고리즘이 이러한 변화를 잘 감지하여 같은 장소로 인식해야하기 때문이다. 알고리즘의 성능 향상을 위해 향후 연구 주제로 지역적 영상 특징 기반 방법과의 결합하는 방법을 수행할 것이다.

감사의 글

이 논문은 2016 년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (No. 2011-0030079)

참고문헌

[1] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. D. Reid, and J. J. Leonard, “Past, Present, and Future of Simultaneous Localization And Mapping: Towards the Robust-Perception Age,” *IEEE Trans. Robot.*, vol. 32, no. 6, pp. 1309–1332, 2016.

[2] D. Gálvez-López and J. D. Tardós, “Bags of binary words for fast place recognition in image sequences,” *IEEE Trans. Robot.*, vol. 28, no. 5, pp. 1188–1197, 2012.

[3] Y. Hou, H. Zhang, and S. Zhou, “Convolutional Neural N etwork-Based Image Representation for Visual Loop Closure Detection,” in *IEEE International Conference on Information and Automation*, 2015.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” in *ImageNet Classification with Deep Convolutional Neural Networks*, 2012, pp. 1097–1105.

[5] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. O liva, “Learning Deep Features for Scene Recognition using P laces Database,” *Adv. Neural Inf. Process. Syst. 27*, pp. 487–495, 2014.

[6] G. E. Hinton and R. R. Salakhutdinov, “Reducing the Dimensionality of Data with Neural Networks,” *Science*, vol. 313, no. 5786, pp. 504–507, Jul. 2006.

[7] D. P. Kingma and M. Welling, “Auto-Encoding Variational Bayes,” *ICLR*, no. 1. pp. 1–14, 2014.

[8] GoogleResearch, “TensorFlow: Large-scale machine learning on heterogeneous systems,” in *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation*, 2016, pp. 265–283.

[9] B. Zhou, A. Khosla, A. Lapedriza, A. Torralba, and A. Oliva, “Places: An Image Database for Deep Scene Understanding,” *ArXiv*, 2016. [Online]. Available: https://arxiv.org/pdf/1610.02055.pdf. [Accessed: 16-Mar-2017].

[10] M. Cummins and P. Newman, “FAB-MAP: Probabilistic Localization and Mapping in the Space of Appearance,” *Int. J. Rob. Res.*, vol. 27, no. 647, pp. 647–665, 2008.

[11] X. Gao and T. Zhang, “Unsupervised learning to detect loops using deep neural networks for visual SLAM system,” *Auton. Robots*, vol. 41, no. 1, pp. 1–18, 2017.