Chapter
01

Visual SLAM 기술개발 동향

김정호_한국전자기술연구원 책임연구원

카메라를 이용하여 환경에 대한 3차원 지도를 생성하고, 카메라 위치를 추정하는 기술은 다양한 분야에 적용이 가능하다. 예를 들면, 자율주행차를 위한 차량 위치 및 환경 인지, AR/VR 디바이스를 활용한 가상 정보 증강, 의료 내시경의 네비게이션 등이 있다. 본 고에서는 카메라를 이용한 3차원 환경 복원 기술과 카메라 자세 변화를 추정하는 Visual SLAM 기술 동향에 대해 살펴보고자 한다. 구체적으로 영상으로부터 특징점을 추출하여 3차원 복원 및 측위를 수행하던 기존의 기하학적 방식과 최근 비약적으로 발전하고 있는 딥러닝을 활용한 방식들을 소개한다. 또한, 기하학적 방법과 딥러닝 기반의 방법을 융합한 하이브리드 방식의 Visual SLAM 기술들에 대해서도 소개한다. 마지막으로 SLAM과 관련된 국내외 시장 동향에 대해서 논의하고자 한다.

I. 서론

미지의 환경에서 다양한 센서들을 이용하여 위치를 추정하고 3차원 환경 지도를 생성하는 기술을 SLAM(Simultaneous Localization and Mapping) 또는 SfM(Structure from Motion)이라고 한다. 본 기술은 1990년대부터 꾸준히 연구되고 있으며 컴퓨터 처리 속도가 개선되고 카메라와 라이다 등의 센서 기술이 발전함에 따라 실제 많은 분야에 응용되고 있다. 예를 들면, [그림 1]과 같이 로봇에 장착된 카메라를 이용하여 위치를 추정하고 환경 구조를 파악함으로써 목적지점까지 자율주행이 가능하다[1].

3차원 환경 지도 생성 및 위치 추정을 위한 대표적인 센서로는 카메라, 레이저 센서 등이 있으며, 카메라로는 단안 카메라, 양안 카메라(stereo camera) 및 RGB-D 카메라 등이 있

* 본 내용은 김정호 책임연구원(☎ 031-739-7480, jhkim77@keti.re.kr)에게 문의하시기 바랍니다.
** 본 내용은 필자의 주관적인 의견이며 IITP의 공식적인 입장이 아님을 밝힙니다.

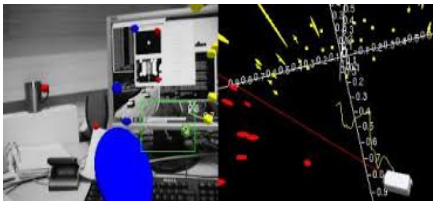
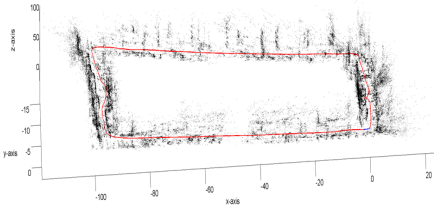
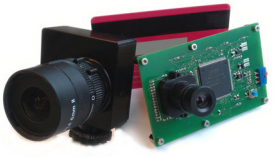
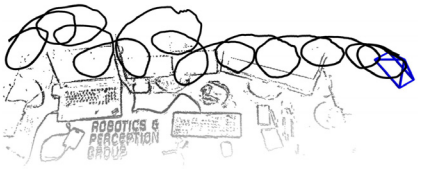
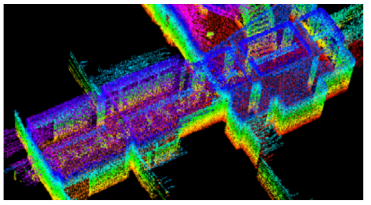
〈자료〉 J. Kim, C. Park, I. Kweon, "Vision-based navigation with efficient scene recognition," JISR, Vol.4, 2011.

[그림 1] SLAM을 이용한 로봇 자율주행 예시

다. 레이저 센서는 수직 해상도(vertical resolution)에 따라서 2D LiDAR와 3D LiDAR로 분류된다. 단일 카메라의 경우 센서의 크기가 작고, 다양한 디바이스에 적용이 가능하지만 단일 영상으로부터 뎁스(depth) 정보를 추정하기 위한 알고리즘이 추가적으로 필요하다. 양안 카메라의 경우 양안 영상으로부터 정합을 통해 거리 정보를 계산할 수 있으나 처리 시간이 필요하며 측정 가능한 거리가 양안 카메라 사이의 거리에 따라서 제한된다. RGB-D 센서는 별도의 계산 없이 뎁스 정보를 바로 획득할 수 있기 때문에 양안 카메라 대비 처리 속도가 빠르고, 조밀한(dense) 지도 데이터를 구축할 수 있는 장점이 있으나 실외 환경에서 적용하기 어렵다. 2D LiDAR 센서는 가격이 저렴하고, 데이터양이 적어서 실시간 처리가 가능하지만 움직임의 자유도(degree-of-freedom)가 높은 경우 적용이 어렵다. 이와는 반대로 3D LiDAR 센서의 경우 많은 수의 3D 포인트들을 제공하기 때문에 조밀한 3차원 복원 및 6자유도의 움직임 추정이 가능하지만 센서가 무겁고 가격이 비싸다는 단점이 있다. [표 1]은 각 센서 유형에 따른 지도 생성 및 위치 추정 결과를 보여준다.

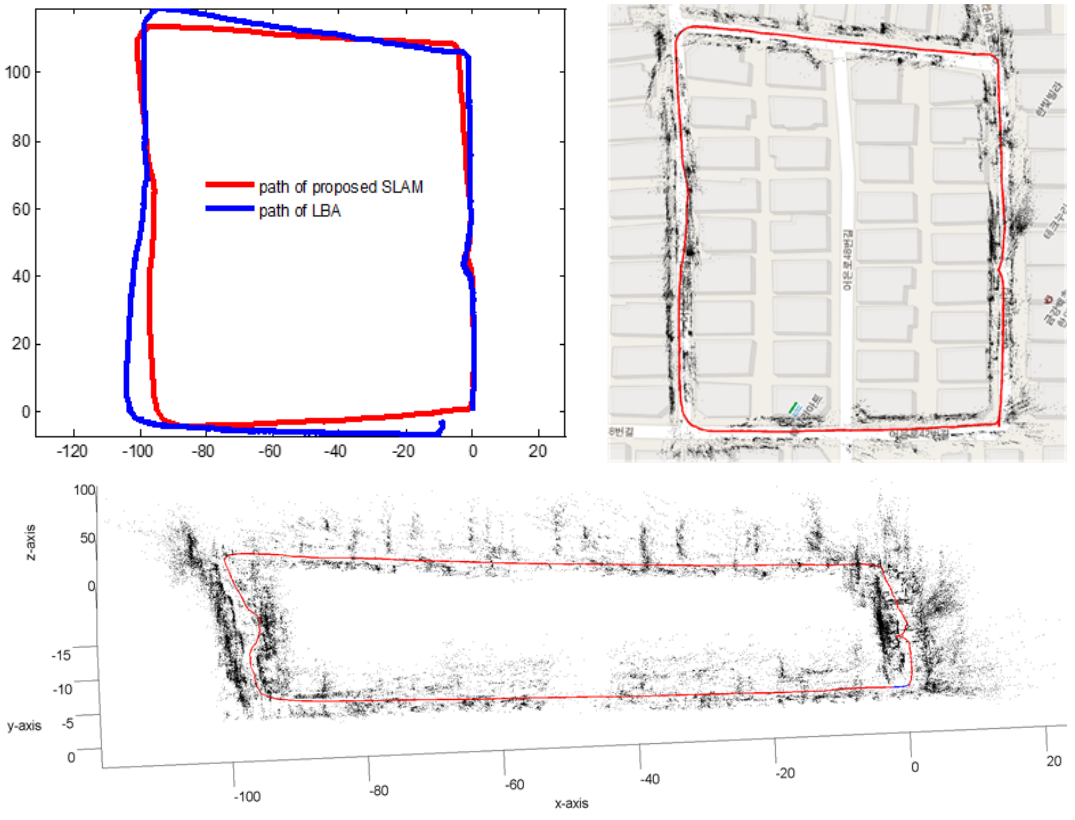
카메라를 이용하여 SLAM을 수행할 경우 시간이 지남에 따라서 오차가 누적되므로 최종 생성된 지도와 위치 추정의 오차가 매우 커지는 경우가 있다. 오차가 발생하는 원인은 주변 환경과 조도 변화에 따라서 센서 관측치(observation)에 대한 노이즈 또는 모호성(ambiguity)이 존재하기 때문이다. 이러한 문제점을 극복하기 위한 Visual SLAM 기술은 크게 기하학적 방법, 학습을 이용한 방법 그리고 학습과 기하학적 방법을 융합한 하이브리드 방법으로 나눌 수 있다. 본 고에서는 오차가 누적되는 문제를 해결하기 위한 개발 동향에 대해서 분석하고, 실제 응용되고 있는 분야와 시장 동향에 대해서 살펴보고자 한다.

[표 1] 다양한 센서들을 이용한 SLAM 결과

센서 종류	센서 이미지	지도 생성 예시
단일 카메라		
스테레오 카메라		
RGB-D 카메라		
이벤트 카메라		
2D LiDAR		
3D LiDAR		

II. 기하학적 Visual SLAM 방법론

임의의 환경에서 획득한 영상 시퀀스로부터 기하학적 계산을 통해 카메라의 위치 및 3차원 지도를 생성하면서 발생하는 오차 누적을 줄이기 위한 방법은 크게 필터링(filtering)과 최적화(optimization) 기법으로 나눌 수 있다. 필터링 기반의 접근 방식은 카메라 자세와 3차원 지도에 대한 확률분포를 영상 데이터를 획득할 때마다 새롭게 갱신하는 방법으로 확률분포의 정의에 따라서 칼만 필터(Kalman filter)[2]와 입자 필터(particle filter)[3] 기반의 방법으로 나눌 수 있다. 칼만 필터는 카메라의 자세와 지도를 구성하는 3차원 랜드마크(landmark)의 위치를 가우시안 분포로 가정하고 영상 데이터를 획득할 때마다 가우시안 분포를 표현하는 평균 벡터와 공분산을 새롭게 갱신하면서 SLAM을 수행한다. 비선형 시스



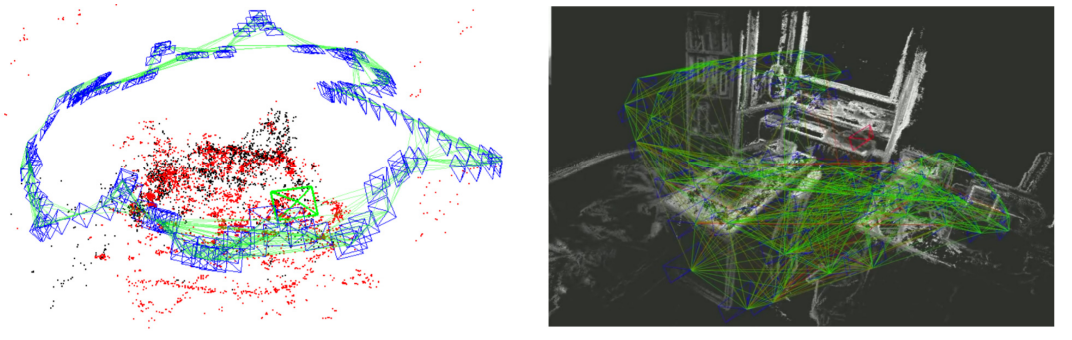
〈자료〉 J. Kim, K. Yoon, I. Kweon, "Bayesian filtering for Keyframe-based Visual SLAM," IJRR, Vol.34, 2015.

[그림 2] Rao-Blackwellized 입자 필터를 이용한 SLAM 결과

템에서 칼만 필터의 정확도 개선을 위한 무향 칼만 필터(unscented Kalman filter) 기반의 SLAM 방법론도 제안되었다[4]. 입자 필터 기반 SLAM 방법의 대표적인 예로 [그림 2]와 같이 Rao-Blackwellized 입자 필터와 키프레임 기법을 이용하여 성능과 처리 속도를 개선한 방법도 제안되었다[5].

최적화 기반의 방법은 크게 두 가지로 나눌 수 있다. 첫 번째 방법은 영상으로부터 특징점을 추출하고 이를 영상 시퀀스에서 추적하여 초기 카메라의 위치를 계산하고 3차원 지도를 생성한다. 그리고 3차원 지도를 구성하는 랜드마크의 위치들을 카메라의 추정된 자세로 재투영(re-projection)시켜서 영상으로부터 추적된 특징점의 좌표와의 거리를 최소화하도록 갱신한다[6]. 또 다른 방법으로는 두 장의 영상으로부터 카메라의 움직임과 환경에 대한 3차원 정보를 획득하기 위해서 첫 번째 영상을 두 번째 위치에서의 영상으로 변환하였을 때 실제 획득한 두 번째 영상과의 밝기 차이를 최소화하도록 최적화를 수행하여 개선하는 Direct SLAM 방법이 있다[7].

특징점 기반 방법의 경우 영상에서 특정 화소들을 이용하여 카메라 위치를 추정하고 3차원 지도를 생성함으로써 처리속도가 빠르다는 장점이 있다. 그리고 특징점의 오정합(false matching)으로 발생하는 문제들을 RANSAC(Random Sample Consensus) 기반의 방법을 활용하여 제거하는 것이 가능하다. Direct SLAM 방법의 경우 처리속도가 느리나 환경을 조밀하게 모델링하는 것이 가능하며 특징점이 없는 균질한(homogeneous) 환경에서 성능이 우수하다. [그림 3]은 특징점 기반과 Direct SLAM의 비교 결과를 보여준다.



[Feature-Based SLAM] [Direct SLAM]

〈자료〉 R, Artal, “Should we still do sparse-feature based SLAM?”, ICCV Workshop, 2015.

[그림 3] Feature-based SLAM과 Direct SLAM의 결과

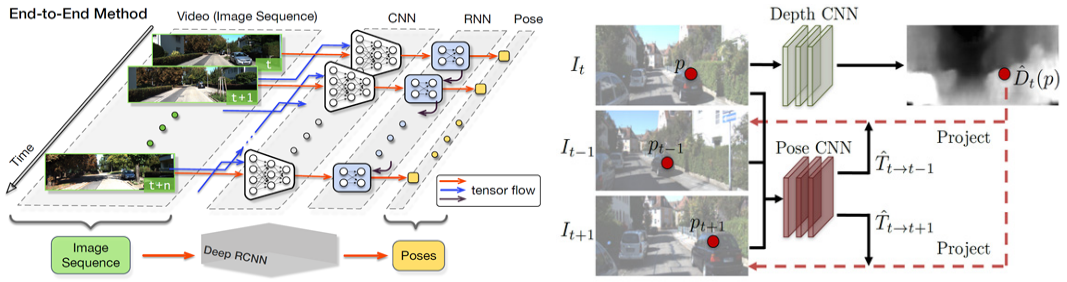
III. 딥러닝 기반 Visual SLAM 방법론

최근 딥러닝 기술의 비약적인 발전으로 기존의 hand-designed 모델 기반 알고리즘의 한계를 극복하며 다양한 분야에 적용되고 있다. 기존의 알고리즘의 경우 특정 환경 및 응용 도메인에 맞도록 개발자들이 보정 또는 수정해서 사용하지만 딥러닝 기반 알고리즘의 경우 학습된 지식을 통해 응용 분야에 맞게끔 모델을 자동으로 설계한다. 학습 기반 방법의 장점은 복잡한 모델과 고차원의 특징 정보들을 사람의 정의 없이 학습 데이터로부터 자동으로 계산할 수 있으므로 다양한 환경 변화 및 특징 정보가 부족한 환경에서도 강인성을 확보할 수 있다. 최근 방대한 양의 SLAM 관련 학습 데이터의 사용이 가능해지고, 복잡한 모델을 학습할 수 있는 최적화 기술들이 개발됨에 따라 SLAM의 다양한 문제들을 딥러닝 학습을 통해 해결하는 것이 가능해지고 있다.

딥러닝 기반의 SLAM 기술들은 크게 오도메트리(odometry) 추정과 매핑(mapping)으로 분류된다[8]. 오도메트리 추정은 두 영상 사이의 상대적인 자세 변화를 추정하는 기술이고, 매핑은 주변 환경에 대한 공간 모델을 생성하는 것을 의미한다.

1. 오도메트리 추정

딥러닝 기반의 오도메트리 추정 기술은 지도 학습과 비지도 학습으로 분류된다. 지도 학습은 연속적인 영상과 그에 대응하는 카메라 자세 변화에 대한 학습데이터가 존재하는 경우 입력 영상에 대한 자세 변화의 출력을 제공하는 종단간(end-to-end) 딥러닝 기술이다. 지도 학습 기반의 방법으로 영상에서 특징 정보를 추출하기 위한 CNN(Convolutional Neural Network)과 순차적 자세 변화 추정을 위한 RCNN(Recurrent Convolutional Neural Network)을 이용하여 입력 영상 시퀀스에 대한 카메라 자세를 출력하는 기술이 제안되었다[9]. 비지도 학습 기반의 방법은 주어진 영상 시퀀스에 대한 자세 학습데이터가 없는 경우 뎁스 정보를 추출하고 자세 변화를 추정하기 위한 딥러닝 기술로서 계산된 자세 변화와 뎁스로부터 다른 시점의 영상을 합성하여 그 시점의 실영상과 비교를 통해 손실함수를 정의하고 학습한다[10]. [그림 4]는 지도 학습과 비지도 학습 기반의 오도메트리 추정 기술의 예를 보여준다.

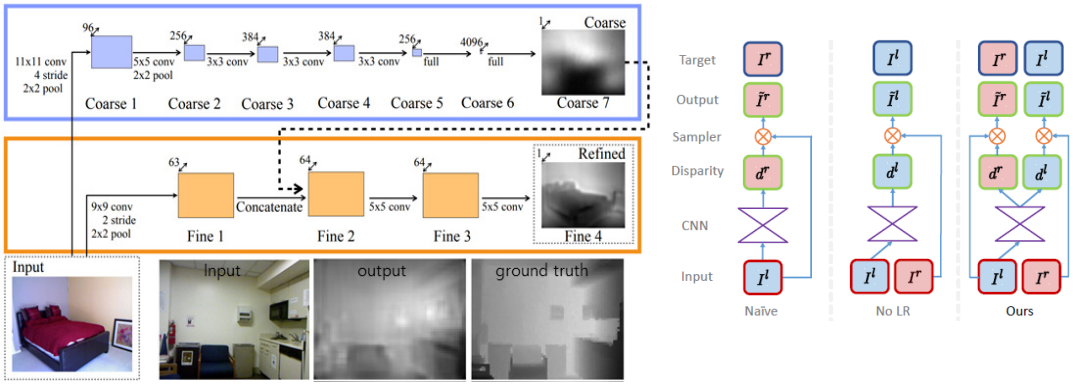


〈자료〉 S. Wang, R. Clark, H. Wen, and N. Trigoni, "DeepVO : Towards End-to-End Visual Odometry with Deep Recurrent Convolutional Neural Networks," ICRA 2017, T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised Learning of Depth and Ego-Motion from Video," CVPR, 2017.

[그림 4] 지도 학습과 비지도 학습 기반의 오도메트리 추정 기술

2. 매핑

매핑은 센서 데이터를 이용하여 주변 환경에 대한 3차원 형상 또는 구조를 표현하는 기술로서 지도를 구성하는 기본 요소에 따라서 뎁스(depth), 포인트(point), 메쉬(mesh), 복셀(voxel) 등으로 나뉜다. 영상정보로부터 거리를 획득하는 방법은 양안 영상을 사용하거나 영상 시퀀스를 이용한 방법이 일반적이었으나 최근 단안 영상으로부터 뎁스 정보를 추출할 수 있는 딥러닝 기술이 활발히 연구되고 있다. 뎁스 생성 기술은 [그림 5]와 같이 지도 학습과



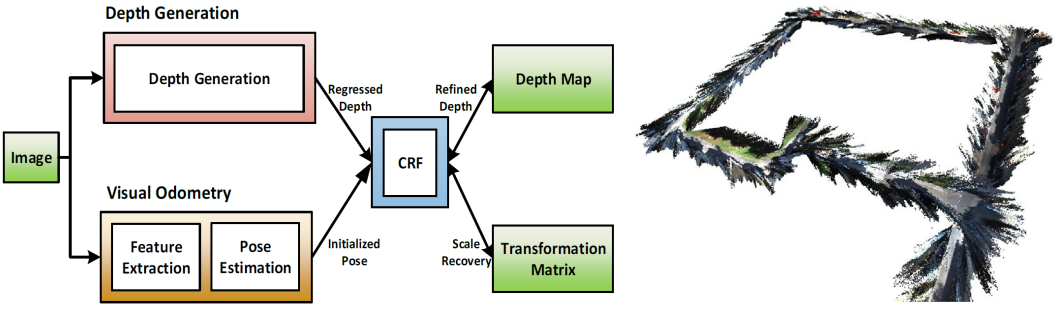
〈자료〉 D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," NIPS, 2014, C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," CVPR, 2017.

[그림 5] 지도 학습과 비지도 학습을 이용한 매핑 기술

비지도 학습 방법으로 나눌 수 있다. 지도 학습 방법은 방대한 양의 영상과 해당 뎁스 데이터를 학습하여 입력영상으로부터 바로 뎁스를 예측하는 기술로 전역과 지역적으로 뎁스를 예측하는 두 개의 네트워크를 이용하여 정확도를 개선하는 기술이 제안되었다[11]. 하지만 학습을 위한 영상과 그에 대응하는 정확한 뎁스 영상을 확보하는 것은 어려운 일이다. 이 문제를 해결하기 위해 비지도 학습 방법에서는 뎁스 영상 대신에 양안 카메라로부터 획득한 영상을 학습 데이터로 활용한다. 구체적으로는 왼쪽 영상을 오른쪽 영상으로 변환하기 위한 시차(disparity)와 오른쪽 영상을 왼쪽 영상으로 변환하는 시차를 계산하고 이를 왼쪽-오른쪽 일관성(left-right consistency) 제약조건을 이용하여 네트워크를 구성함으로써 향상된 뎁스를 생성하는 기술을 제안하였다[12].

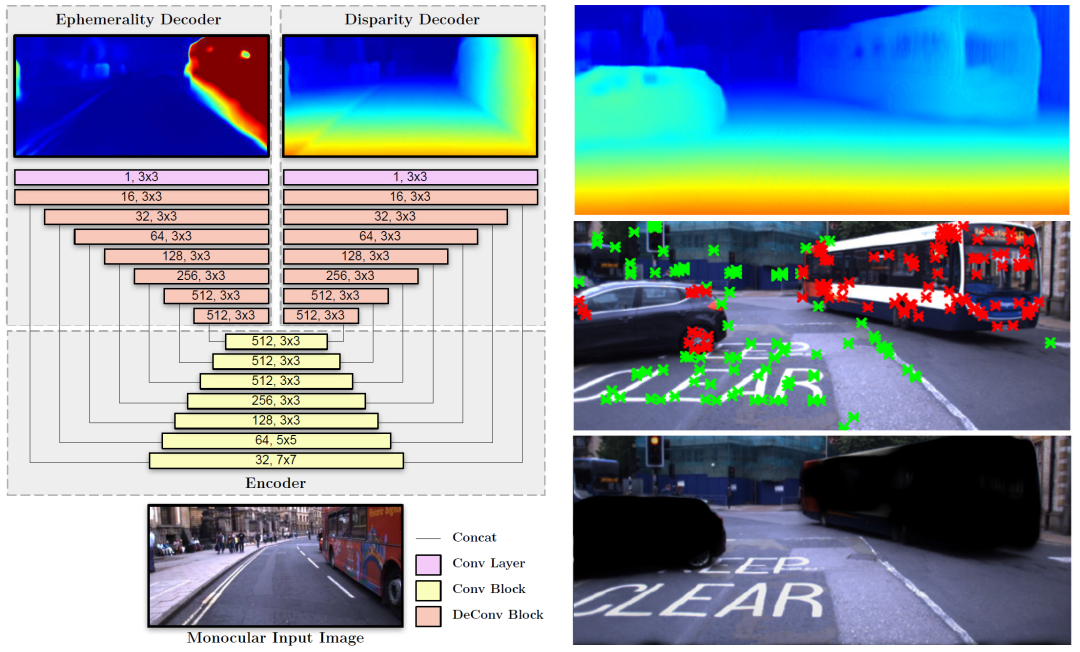
IV. 하이브리드 Visual SLAM 방법론

하이브리드 방식은 Visual SLAM을 구성하고 있는 여러 단계 중 일부를 딥러닝 방법으로 계산하고 다른 일부는 고전적인 기하학적 방법을 활용한다. 딥러닝 기반의 방법은 영상에서 특징 정보를 추출하기 어려운 환경에서 더 나은 결과를 제공하지만, 특징 정보가 풍부한 환경에서는 고전적인 방법의 성능이 우수하다. 예를 들면, [그림 6]과 같이 단일 영상으로 뎁스를 예측하기 위해 딥러닝 기반의 방법을 사용하였고, 입력 영상으로부터 자세 변화를 추정하는 부분은 기학적 오도메트리 추정 방법을 채택하였다[13]. 생성된 초기 뎁스와 카메라 자세는 CRF(Conditional Random Field)를 이용하여 정확도를 향상시켰다.



〈자료〉 X. Yin, X. Wang, X. Du, and Q. Chen, “Scale recovery for monocular visual odometry using depth estimated with deep convolutional neural fields,” ICCV, 2017.

[그림 6] 딥러닝 기반의 뎁스 생성과 VO를 융합한 하이브리드 Visual SLAM 기술



〈자료〉 D. Barnes, W. Maddern, G. Pascoe, and I. Posner, "Driven to distraction: Self-supervised distractor learning for robust monocular visual odometry in urban environments," ICRA, 2018.

[그림 7] 동적 환경에 강인한 하이브리드 SLAM 기술

또 다른 하이브리드 방법으로는 뎁스와 광류(optical flow)를 딥러닝 기반의 학습으로부터 계산하고, 이 결과물을 기하학적 오도메트리 알고리즘에 적용하여 카메라의 자세 변화를 추정하였다[14]. [그림 7]과 같이 딥러닝 기술로부터 움직이거나 또는 변화가 있는 부분을 검출하여 기존 특징점 기반과 Direct SLAM의 성능을 개선하는 방법도 제안되었다[15].

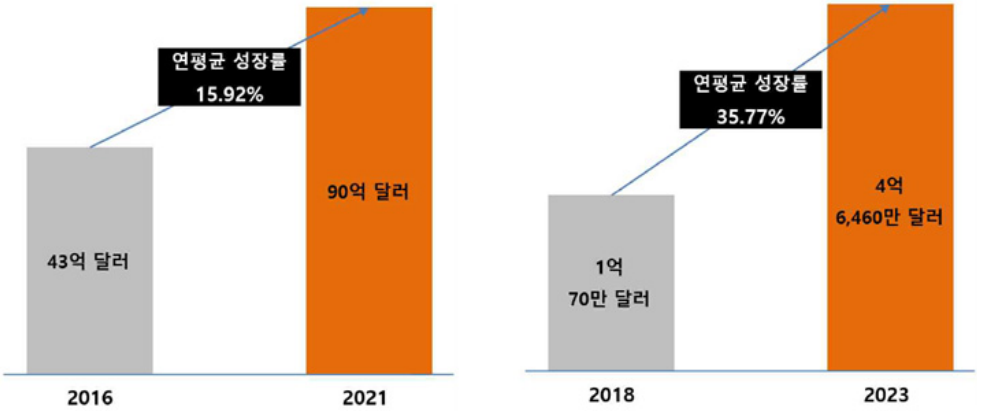
최근에는 딥러닝 기술과 기존 필터링 기반의 방법이 융합된 SLAM 기술로서 칼만 필터와 입자 필터를 딥러닝으로 학습하는 기술들이 개발되었고, 이는 카메라의 자세 변화 추정 기술에 적용되어 성능이 개선됨을 보여 주었다[16],[17].

V. Visual SLAM 기술 시장 동향

SLAM 기술의 경우 다양한 분야에 적용이 가능하다. 현재는 디지털트윈과 같은 증강현실 및 가상현실 콘텐츠 생성과 AR 디바이스의 자세 트래킹 등이 대표적인 적용 사례이다. 뿐만

아니라 제조, 의료, 교육 등 산업별 특화된 고부가가치 AR 기술 수요의 급증이 전망되고 있으며, COVID-19로 인한 비대면 문화 확산으로 AR/VR 응용 기술들이 주목 받고 있다. 그리고 SLAM 기술은 지능형 로봇 또는 자동차의 자율 주행을 위한 제품에 적용 가능하며 특히 기존 맵이 없는 실내 응용 분야에 대한 수요가 증가함에 따라서 개발 필요성이 대두되고 있다.

[그림 8]과 같이 전 세계 디지털 맵 시장은 2016년 43억 달러에서 연평균 15.92% 성장하여 2021년에는 90억 달러에 이를 것으로 전망되고 있으며, 전 세계 SLAM 시장은 2018년 1억 70만 달러에서 연평균 35.77% 성장하여, 2023년에는 4억 6,460만 달러에 이를 것으로 전망되고 있다. 또한, 국내 SLAM 시장은 2018년 480만 달러에서 연평균 41.19% 성장하여, 2023년에는 2,690만 달러에 이를 것으로 전망된다.



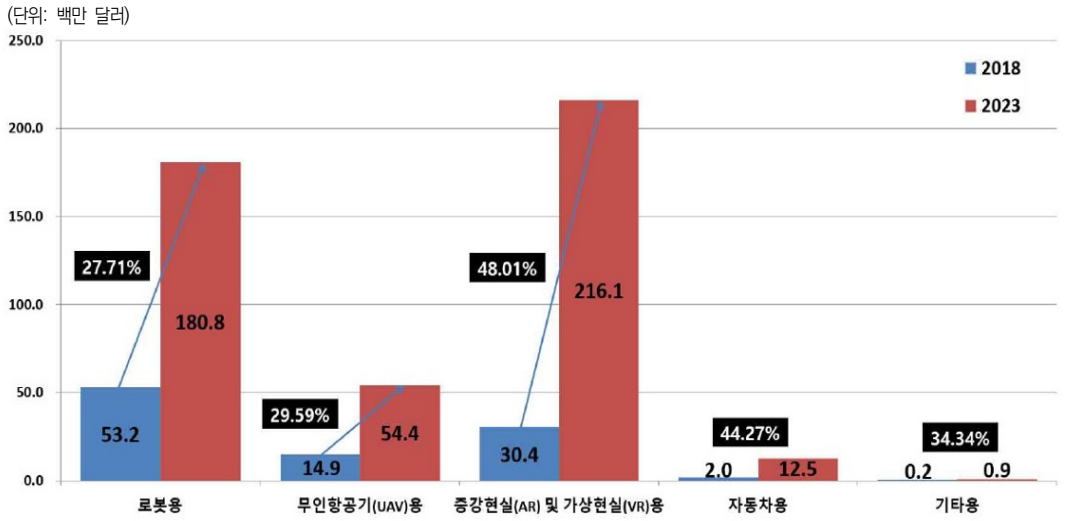
[글로벌 디지털 맵 시장 규모 및 전망]

[글로벌 SLAM 시장 규모 및 전망]

〈자료〉 연구개발특구진흥재단, SLAM 시장, 2020.

[그림 8] 글로벌 디지털 맵과 SLAM 시장의 규모 및 전망

[그림 9]와 같이 전 세계 SLAM 시장은 용도에 따라서 로봇용, 무인항공기(UAV)용, AR/VR용, 자동차용, 기타용으로 분류되며, 2018년부터 2023년까지 연평균 성장률이 로봇용은 27.71%, 무인항공기용은 29.59%, AR/VR용은 48.01% 그리고 자동차용은 44.27%에 각각 이를 것으로 전망되고 있다.



〈자료〉 연구개발특구진흥재단, SLAM 시장, 2020.

[그림 9] 글로벌 SLAM 시장의 용도별 시장 규모 및 전망

VI. 결론

Visual SLAM 기술은 AR/VR을 위한 콘텐츠 생성 및 지능형 로봇 또는 자동차의 자율주행 등 다양한 분야에 적용되는 핵심 기술이다. 하지만 카메라와 환경 조건으로부터 야기되는 다양한 문제들로 인해 아직 완벽한 솔루션이 없는 것이 현실이다. 예를 들면, 실내 가정환경에서 특징정보가 없는 벽으로부터 SLAM 적용 시 큰 오차가 발생하는 문제 또는 급격한 움직임 변화와 저조도로 인한 오차가 크게 발생하는 등 해결해야 할 많은 문제들이 있다. 그리고 SLAM을 위한 알고리즘은 복잡한 연산이 필요하기 때문에 저사양 임베디드 또는 AR 디바이스 내에 탑재되어 있는 처리 장치를 사용할 경우 실시간 운용이 어렵다.

기존 기하학적 방법들은 영상에서 특징정보를 추출하고 영상 간 매칭을 통해 카메라의 자세 변화를 추정하고 이로부터 점진적으로 3차원 지도를 생성한다. 오차 누적을 방지하기 위해서 확률적인 필터링 기법을 적용하거나 또는 획득한 모든 센서 데이터를 활용하여 비선형 최적화 기법을 사용한다. 최근 딥러닝 기술의 비약적인 발전에 힘입어 입력 영상으로부터 카메라의 자세 변화와 3차원 구조에 대한 정보를 학습을 통해 직접 획득할 수 있는 방법들이

제안되었다. 그리고 두 방법의 장점을 융합한 하이브리드 방식의 SLAM 방법들이 제안되고 있다.

지난 수십 년간 Visual SLAM의 기술들은 발전하였고, 최근 딥러닝과 융합된 기술들이 개발되고 있으므로 저가의 카메라를 이용한 로봇 및 자동차 자율주행을 위한 기술 개발이 머지않아 완성될 것으로 예상되며, 이를 통해 의료, 콘텐츠, 스마트 팩토리 등 다양한 산업에 응용될 것으로 예상된다.

참고문헌

[1] J. Kim, C. Park, I. Kweon, "Vision-based navigation with efficient scene recognition," Journal of Intelligent Service Robotics, Vol.4, 2011.

[2] A. Davison, "MonoSLAM: Real-Time Single Camera SLAM," IEEE Trans. Pattern Analysis and Machine Intelligence, Vol.29, 2007.

[3] I. Rekleitis, "A particle filter tutorial for mobile robot localization," McGill University, TR-CIM-04-02, 2004.

[4] D. Chekhlov, M. Pupilli, W. Mayol, A. Calway., "Robust Real-Time Visual SLAM Using Scale Prediction and Exemplar Based Feature Description," IEEE International Conference on Computer Vision and Pattern Recognition, 2007.

[5] J. Kim, K. Yoon, I. Kweon,, "Bayesian filtering for Keyframe-based visual SLAM," The Intenational Journal of Robotics Research, Vol.34, 2015.

[6] R. Mur-Artal, J. M. M. Montiel and J. D. Tardos., "ORB-SLAM: A Versatile and Accurate Monocular SLAM System," IEEE Transactions on Robotics, Vol.31, 2015.

[7] J. Engel, V. Koltun and D. Cremers., "Direct Sparse Odometry," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.40, 2018.

[8] C. Chen, B. Wang, C. Lu, A. Trigoni, A. Markham., "A Survey on Deep Learning for Localization and Mapping: Towards the Age of Spatial Machine Intelligence," arXiv:2006.12567, 2020.

[9] S. Wang, R. Clark, H. Wen, and N. Trigoni, "DeepVO: Towards End-to-End Visual Odometry with Deep Recurrent Convolutional Neural Networks," IEEE International Conference on Robotics and Automation, 2017.

[10] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised Learning of Depth and Ego-Motion from Video," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[11] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in Advances in Neural Information Processing Systems, 2014.

[12] C. Godard, O. Mac Aodha, and G. J. Brostow, “Unsupervised monocular depth estimation with left-right consistency,” IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp.270-279.
[13] X. Yin, X. Wang, X. Du, and Q. Chen, “Scale recovery for monocular visual odometry using depth estimated with deep convolutional neural fields,” IEEE International Conference on Computer Vision, 2017.
[14] H. Zhan, C. S. Weerasekera, J. Bian, and I. Reid, “Visual odometry revisited: What should be learnt?,” The International Conference on Robotics and Automation, 2020.
[15] D. Barnes, W. Maddern, G. Pascoe, and I. Posner, “Driven to distraction: Self-supervised distractor learning for robust monocular visual odometry in urban environments,” IEEE International Conference on Robotics and Automation, 2018.
[16] T. Haarnoja, A. Ajay, S. Levine, and P. Abbeel, “Backprop kf: Learning discriminative deterministic state estimators,” in Advances in Neural Information Processing Systems, 2016.
[17] P. Karkus, S. Cai and D. Hsu, "Differentiable SLAM-net: Learning Particle SLAM for Visual Navigation," IEEE Conference on Computer Vision and Pattern Recognition, 2021.