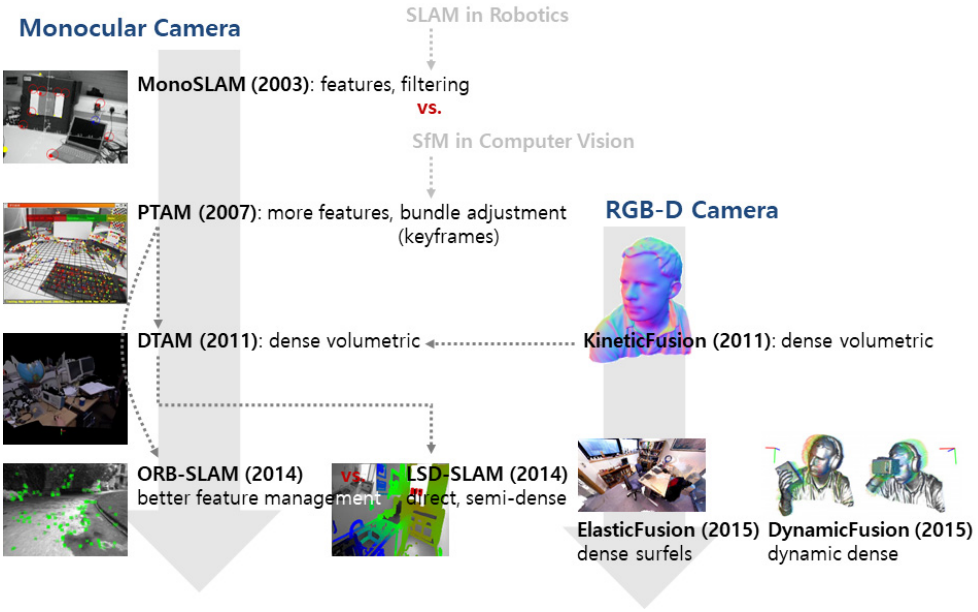
Visual SLAM을 통해 살펴본 SLAM 기술의 변화와 흐름

최성록[1] · 최중용[2] · 김현주[2]
([1]서울과학기술대학교, [2]한국전자통신연구원)

1. 서 론

Simultaneous Localization and Mapping (이하 SLAM) 기술은 로봇이 주어진 공간에서 자신의 위치를 알아내고 동시에 위치 추정에 필요한 지도를 작성하는 기술을 말한다. 위치 추정에 사용되는 지도는 표식점(landmark)이나 특징점(feature)의 집합으로 표현될 수 있고, 공간의 형태나 모양이 위치 추정에 사용되는 경우 격자지도(grid map)나 점군(point cloud)으로 표현될 수도 있다. 위치 추정을 위해서는 이를 위한 환경 정보가 담긴 지도가 필요하고, 다시 지도 작성을 위해 로봇의 위치가 필요하기 때문에 SLAM 문제는 단순한 순차적인 작업이 아니라 '닭과 달걀의 문제'와 같이 인과가 얽혀있는 다소 난해한 문제이다. SLAM 기술은 로봇이 자율주행에 필요한 위치 정보를 제공하고 이를 위한 공간 인프라인 지도 정보를 생성하기 때문에 로봇, 특히 자율주행 분야에서 많은 연구와 개발, 응용이 이뤄져 왔고, 로봇 분야에서 여전히 가장 활발히 연구/개발되는 분야 중의 하나이다.

SLAM 문제와 관련된 연구는 1980년대에도 있었지만, 본격적으로 SLAM이라는 용어가 사용되고 본격적인 연구가 시작된 것은 1991년 John Leonard 교수와 Hugh Durrant-Whyte 교수의 연구[1]부터이다. 이후 많은 그리고 중요한 연구들이 많이 이루어져 왔고, IEEE RAS에서는 몇 차례에 걸쳐 SLAM Summer School (SSS)[2]를 개최하기도 하였다. 또 2016년에는 John Leonard 교수를 비롯한 8명의 저자가 SLAM 기술의 과거와 현재, 그리고 미래를 조망하는 논문[3]을 출간하기도 하였다. SLAM 기술은 다양한 기술들이 한데 어우러진 종합 예술과 같은 연구 분야로 과거의 OpenSLAM.org[4]를 비롯해 현재도 꾸준히 업데이트되고 있는 MRPT[5]와 같은 잘 구현된 양질의 연구 결과들이 오픈소스로 공개되어 있다. 현재 ROS나 Github에 공개된 다양하고 멋



[그림 1] Visual SLAM의 두 가지 패러다임 변화의 관점에서 본 기술 흐름

진 연구 결과들을 손쉽게 자신의 로봇에 적용해 볼 수 있다. 최근에는 성숙된 SLAM 기술을 바탕으로 한 국내외 스타트업들이 많이 생기기도 했다.

본 기고에서는 저자들의 개인적인 관점에서 SLAM 기술의 변화와 흐름을 가볍고 조금 더 쉬운 이야기로 살펴보고자 한다. 지난 30년 동안 중요한 연구들과 기술들이 많이 있었지만, 본 기고에서는 SLAM 분야의 몇 가지 중요한 패러다임 변화를 이끈 카메라 영상을 이용한 SLAM, 즉 visual SLAM 기술을 바탕으로 SLAM 기술의 변화와 흐름을 살펴보고자 한다. 카메라는 스마트폰이나 CCTV 카메라를 비롯하여 우리 주변에서 가장 쉽게 접할 수 있는 센서이다. 카메라는 조명 변화에 크게 영향을 받는다는 단점을 가지고 있지만, 다른 센서들에 비해 가격대비 빠르고 많은 양의 데이터를 제공해주는 장점이 있다. Visual SLAM 기술은 로봇이나 자율차, 드론의 자율주행뿐만 아니라 공간이나 물체의 3차원 형상 복원(3D reconstruction)이나 VR/AR과 같은 기술의 기반 기술로 널리 사용된다.

Visual SLAM 기술의 진보와 많은 변화 속에서 [그림 1]과 같이 두 가지 지점에서 큰 변곡점, 즉 패러다임의 변화가 있었다. 첫 번째 변곡점은 SLAM 기술 초창기부터 많이 주로 사용되어온 Bayesian filtering 기반의 방법론에서 컴퓨터비전 분야의 bundle adjustment와 같은 그래프 최적화 기반의 방법론으로의 변화이다. 두 번째는 변곡점은 현재 진행 중인데, 기존의 영상에서 특징점을 추출하고 이를 SLAM에 활용하는 특징점 기반(feature-based) 방법론와 특징 추출의 과정 없이 영상을 직접(direct) 사용하는 방법론 사이의 충돌이다. 본 기고에서는 이러한 두 가지 변곡점을 통해 SLAM 기술의 변화와 흐름에 대해 살펴보고자 한다.

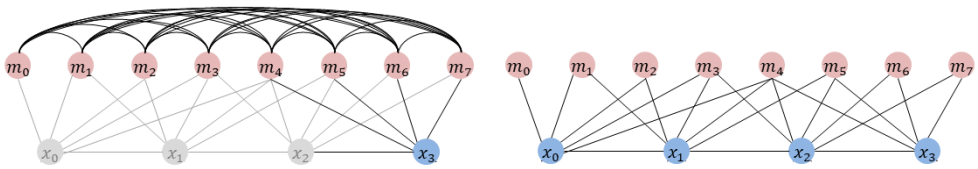
2. 첫 번째 변곡점: Bayesian Filtering vs. Graph Optimization

칼만필터나 확장칼만필터(이하 EKF), 파티클필터로 대표되는 Bayeisan filtering은 시간에 따라 순차적으로 입력되는 시계열 데이터의 추정 문제에 가장 널리 사용되는 방법 중 하나이다. 초창기 SLAM 기술은 대부분 Bayesian filtering을 사용하였고, 2003년에 발표된 MonoSLAM[6]은 카메라 하나만 이용해 drift error 없이 실시간으로 동작하는 첫 번째 성공적인 visual SLAM 기술로 널리 알려졌다. MonoSLAM은 EKF를 이용한 SLAM 기법으로 영상에서 11x11 픽셀 크기의 패치(patch)를 특징점으로 뽑는데, 320x240 크기의 영상에서 일반적으로 약 10개 정도의 특징점을 추출하여 이용하고, 전체 지도에 약 100개 정도의 특징점을 저장하여 사용하는 수준이었다. EKF를 이용한 추정 기술은 $O(n^3)$의 시간 복잡도를 갖고, 특징점의 개수(n)가 늘어나면 너무 많은 연산 시간이 필요하여 많은 수의 특징점을 활용하기에 불리한 구조였다. MonoSLAM의 적용은 당시 좁은 실내 공간에만 한정되었다.

2007년 발표된 PTAM (Parallel Tracking and Mapping)[7]은 컴퓨터비전 분야에서 structure-from-motion[1] (이하 SfM), 즉 카메라의 촬영 위치와 영상 특징점의 위치를 동시에 추정하는 분야에서 사용되는 비선형 최적화 기법인 bundle adjustment (이하 BA)를 이용한 첫 번째 실시간 visual SLAM 기술이다. SfM에 사용되는 BA는 복잡한 비선형 에러 함수를 최소화하는 기법으로 많은 시간이 소모되어 실시간 SLAM에 거의 활용되지 않았다. PTAM은 위치 추정과 지도 작성 문제를 각각 실시간으로 동작하는 motion-only BA, 많은 계산 시간이 필요하지만 가끔 동작하는 motion-and-structure BA로 나누어 별도의 병렬적인 스레드로 동작하게 하였다. 이를 통해 위치 추정에 의한 결과는 실시간으로 제공되고, 이를 위한 지도는 충분한 연산을 통해 개선된다. 그동안 Bayesian filtering과 같이 하나의 프레임워크로 동시 추정되었던 로봇의 위치와 지도가 다시 분리되는 순간이었다. 단순히 병렬처리로 실시간 동작이 가능했던 것은 아니고, 키프레임(keyframe)이라는 개념을 도입하여 과거의 전체 영상을 모두 이용하는 것이 아니라 꼭 필요한 영상만 추려서 BA에 활용하여 계산 시간을 획기적으로 줄였다.

'Why filter?'라는 별칭의 2010년에 발표된 논문[8]은 Bayesian filtering과 BA 사이의 성능 차이를 분명하게 결정지었다. [그림 2]는 EKF와 BA를 이용한 SLAM을 그래프 형태로 가시화한 것이다. 그래프에서 x_i는 i번째 로봇의 위치, m_j는 j번째 특징점의 위치이다. EKF의 경우 마르코프(Markov) 가정하에서 직전의 위치를 기준으로 현재의 위치인 x_3만을 추정하는 반면, BA의 경우 모든 과거의 위치도 최적화에 포함하여 과거의 위치도 현재 관찰된 특징점으로 개선되고 결국 보다 정확한 현재의 위치를 얻게 된다. 또한 Bayesian filtering을 이용한 방법은 위치 추정과 지도 작성 사이에 의존성(highly-coupled)이 크고, 특히 EKF의 경우 $O(n^3)$의 계산복잡도로 많은 특징점을 이용하는데 제약이 있었지만, BA를 이용한 방법은 위치 추정과 지도 작성의 분리가 용

1) SLAM과 SfM은 단어와 그 순서만 다를 뿐 근본적으로 같은 문제를 풀고 있다. (localization ~ motion, map ~ structure)



(a) Bayesian filtering (b) Bundle adjustment

[그림 2] Bayesian filtering과 bundle adjustment를 이용한 SLAM의 그래프 표현

이하고 보다 많은 특징점을 활용할 수 있어 속도 개선과 정확도 개선에 이점이 있다. 현재 개발되는 대부분의 SLAM 기술은 BA, 즉 그래프 최적화에 기반을 두어 개발되고 있다. 현재 널리 사용되고 ORB-SLAM[9]은 PTAM의 아이디어를 보다 넓은 공간에 보다 장시간 동작 가능하도록 만든, 즉 완성도를 매우 높인 확장판이다. ORB-SLAM은 이후 꾸준히 개선되어 스테레오 카메라와 RGB-D 카메라를 지원[10]하고, 최근 관성센서와 다중 분할 지도를 지원[11]하도록 버전업되었다.

3. 두 번째 변곡점: Feature-based Methods vs. Direct Methods

Microsoft는 2010년 Xbox 게임기를 위한 RGB-D 카메라인 Kinect를 발매하였고, 이후 학계에서는 이를 이용한 인상적인 RGB-D SLAM 기술들이 다수 개발되었다. RGB-D 카메라는 RGB 영상 외에 각 픽셀의 물리적 거리(depth) 값이 추가로 제공된다. 따라서 초기 RGB-D SLAM 기술들은 기존 visual SLAM 기술보다 scan matching과 유사한 방법론들이 많이 사용되었다. 즉, 영상에서 특징점을 추출하여 활용하기보다 scan matching 중 하나인 ICP (iterative closest points)와 같이 두 RGB-D 데이터 사이의 카메라 작은 크기의 자세 변화에 대한 정합(registration)을 추정하고 이를 SLAM에 이용하는 방법이다. 이러한 RGB-D SLAM은 영상에서 특징점을 추출하여 이용하지 않고, 직접(direct) 또 영상의 대부분을 고밀도(dense)로 이용하는 것이 큰 특징이다.

이후 RGB-D SLAM의 영향을 받은 direct SLAM 기술들이 제안되었고, DTAM (Dense Tracking and Mapping)[12], LSD-SLAM (Large-scale Direct Monocular SLAM)[13], SVO (Semi-direct Monocular Visual Odometry)[14], DSO (Direct Sparse Odometry)[15]가 대표적인 예들이다. 이러한 direct method의 가장 큰 특징은 RGB-D SLAM과 마찬가지로 특징점 추출의 단계가 없이 영상의 픽셀의 값을 그대로 이용한다는 점이다. 조금 더 쉽게 direct method와 기존의 feature-based method를 비교하기 위해 [그림 3]과 같은 영상 정합 (image stitching)의 문제를 살펴보자. 영상 정합에서는 (SLAM에서 영상의 위치와 방향각과 마찬가지로) 영상의 병진 이동 와 회전 이동 θ를 정확하게 추정하는 것이 두 영상을 합치는데 핵심이다. 기존의 feature-based method의 경우, 영상에서 특징점을 추출하고 이를 매칭하여 유사한 특징점의 쌍(pair)을 찾는다. 이후 특징점의 위치에 병진과 회전 이동을 적용하여 특징점 쌍 사이의 거리, 즉 기하학적 에러(geometric

$$\arg\min_{\mathbf{t},\theta} \sum_i \left\| \mathrm{R}(\theta)\mathbf{x}_i + \mathbf{t} - \mathbf{x}'_i \right\|^2$$

(a) Feature-based method
(minimizing geometric errors)

$$\arg\min_{\mathbf{t},\theta} \sum_{u,v} \left\| I(\mathrm{R}_\theta \begin{bmatrix} u \\ v \end{bmatrix} + \mathbf{t}) - I'(\begin{bmatrix} u \\ v \end{bmatrix}) \right\|^2$$

(b) Direct method
(minimizing photometric errors)

[그림 3] Feature-based 기법과 direct 기법의 비교

error)를 최소로 만드는 병진/회전 이동을 찾는다. 반면 direct method에서는 특징점 추출과 매칭의 과정 없이 주어진 영상에 병진 이동과 회전 이동을 적용하며, 현재 정합된 두 영상의 각 픽셀의 값의 차, 즉 광량적 에러(photometric error)를 최소로 만드는 병진/회전 이동을 찾는다.

Visual SLAM을 위한 feature-based method와 direct method의 우열은 아직 명확히 결론 나지 않았다. 다만 direct method의 경우 최근에 많은 관심을 두고 본격적으로 연구되기 시작한 방법론이기 때문에 앞으로 더욱 발전될 수 있겠다는 기대감이 있는 것은 사실이다. Direct method의 경우 생성된 지도만으로 공간의 윤곽을 쉽게 파악할 수 있을 만큼 밀도가 높은 지도를 생성하고, 특징점 추출과 매칭 단계가 없기 때문에 상대적으로 계산 시간을 줄일 여지가 있다. 하지만 지도의 높은 데이터 밀도로 인해 현재는 feature-based method보다 조금 더 느린 것이 사실이고, 특히 광량적 에러를 사용하기 때문에 카메라의 광량적 보정(photometric calibration)이 필요하다. 또 같은 연장선에서 롤링 셔터(rolling shutter), 자동 노출 조정(auto-exposure) 등과 같은 동적인 카메라 특성이나 움직이는 물체와 같은 동적인 환경에서 성능이 크게 떨어진다. 2015년에 ICCV에서 'The Future of Real-time SLAM'이라는 이름의 워크샵[16]이 있었고, 두 방법론을 대표하는 연구자들의 불꽃 튀기는 발표과 설전이 있었다. 당시 라이브 데모에서 feature-based method의 속도와 성능이 조금 더 나아 보였는데, 꽤 많은 시간이 지난 지금도 두 방법론의 우열은 쉽게 결론 나지 않았다.

4. 결 론

SLAM 기술이 본격적으로 연구된 것은 약 30년 정도, visual SLAM에 대한 연구는 MonoSLAM 이후 약 20년 정도이다. 그동안 괄목할 만한 많은 연구들이 있었고, 몇 번의 큰 변곡점이 있기도 하였다. 본 기고에서는 visual SLAM 관점에서 중요한 몇 가지 연구들을 살펴보고, 그들이 갖는 두 가지 패러다임 변화를 간략히 살펴보았다. Graph-based SLAM이 최근 SLAM의 기본적인 프레임워크가 된 것과 같이 이러한 패러다임 변화는 비단 visual SLAM 뿐만 아니라 LiDAR

SLAM, GPS-based SLAM 등의 다른 센서들을 활용한 SLAM에도 큰 영향을 미쳤다. 최근 딥러닝에 대한 연구가 지식의 표현과 학습, 인식, 계획, 제어 분야뿐만 아니라 SLAM을 비롯한 3차원 컴퓨터비전 분야에도 폭넓게 활용되고 적용되고 있다. 일부는 학습된 데이터 내에서만 좋은 결과를 갖지만, 이를 벗어나 다른 기하학적 속성의 카메라도 적용 가능한 알고리즘들도 제안되고 있다. 이러한 최근 흐름이 2번째 패러다임의 변화를 끝낼 수도 있고, 아니면 3번째 패러다임의 시작이 될 수도 있다.

본 기고에서 설명된 기술들 외에 그동안 국내 연구자들에 의해 연구된 인상적인 SLAM 분야의 연구 결과들도 많았다. 다음 기회에는 전술한 SLAM의 흐름 속에서 국내 연구자들이 제안한 멋진 SLAM 기술들을 소개하고 싶다.

사 사

본 연구는 문화재청 및 국립문화재연구소의 2021년도 '문화유산 스마트 보존·활용 기술 개발' 사업으로 수행되었습니다. (과제명: 초고해상도 기가픽셀 3D 데이터 생성 기술 개발, 과제번호: 2021A02P02-001)

참고문헌

[1] J. Leonard and H. Durrant Whyte, "Simultaneous Map Building and Localization for an Autonomous Mobile Robot," IROS, 1991

[2] "SLAM Summer School 2006", https://www.robots.ox.ac.uk/~SSS06/

[3] C. Cadena et al., "Past, Present, and Future of Simultaneous Localization and Mapping: Toward the Robust-Perception Age," T-RO, vol 32, no. 6, 2016

[4] "OpenSLAM.org", https://openslam-org.github.io/

[5] "Mobile Robot Programming Toolkit (MRPT)", https://www.mrpt.org/

[6] Andrew Davison, "Real-Time Simultaneous Localisation and Mapping with a Single Camera," ICCV, 2003

[7] Georg Klein and David Murray, "Parallel Tracking and Mapping for Small AR Workspaces," ISMAR, 2007

[8] H. Strasdat et al., "Real-time monocular SLAM: Why filter?," ICRA, 2010

[9] R. Mur-Artal et al., "ORB-SLAM: A Versatile and Accurate Monocular SLAM System," T-RO, 2015

[10] R. Mur-Artal and J. Tardos, "ORB-SLAM2: an Open-Source SLAM System for Monocular, Stereo and RGB-D Cameras," T-RO, 2017

[11] C. Campos et al., “ORB-SLAM3: An Accurate Open-Source Library for Visual, Visual –Inertial, and Multimap SLAM,” T-RO, 2021

[12] R. Newcombe et al, “DTAM: Dense Tracking and Mapping in Real-time,” ICCV, 2011

[13] J. Engel et al., “LSD-SLAM: Large-Scale Direct Monocular SLAM,” ECCV, 2014

[14] C. Foster et al., “SVO: Fast Semi-Direct Monocular Visual Odometry,” ICRA, 2014

[15] J. Engel et al., “Direct Sparse Odometry,” T-PAMI, 2018

[16] “The Future of Real-Time SLAM: Sensors, Processors, Representations, and Algorithms,” ICCV Workshop, 2015, http://wp.doc.ic.ac.uk/thefutureofslam/

최성록

2006 서울대학교 기계항공공학과 (공학사)
2008 KAIST 로봇공학학제전공 (공학석사)
2019 KAIST 로봇공학학제전공 (공학박사)
2008~2020 ETRI 지능로보틱스연구본부(선임연구원)
2021~현재 서울과학기술대학교 컴퓨터공학과 (조교수)
관심분야 : 로봇 주행, 3차원 컴퓨터비전
E-mail : sunglok@seoultech.ac.kr

최중용

2006 삼육대학교 컴퓨터공학과 (공학사)
2006~2007 LPA 소프트웨어연구소
2007~2008 LBS Plus GIS연구소
2008~2012 팅크웨어 GIS연구소
2012~현재 ETRI 콘텐츠연구본부 (선임기술원)
관심분야 : 3D Reconstruction, GIS, 컴퓨터비전
E-mail : choijy725@etri.re.kr

김현주

2000 성균관대학교 정보공학과 (공학사)
2002 성균관대학교 컴퓨터공학전공 (공학석사)
2016 성균관대학교 컴퓨터공학전공 (공학박사)
2002~현재 ETRI 콘텐츠연구본부 (선임연구원)
관심분야 : 컴퓨터비전, SLAM, 포토메트리, 시각화기술
E-mail : hjookim@etri.re.kr