Visual SLAM algorithms: a survey from 2010 to 2016

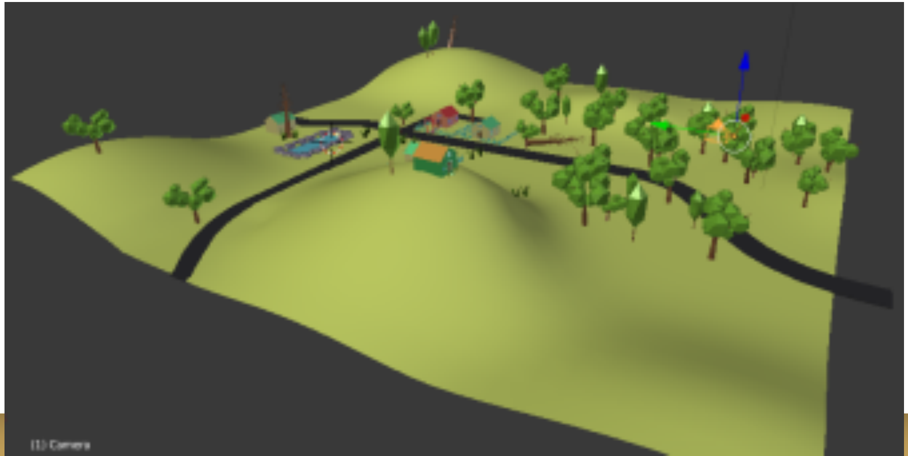
Takafumi Taketomi1*, Hideaki Uchiyama2 and Sei Ikeda3

IPSJ Transactions on Computer Vision and Applications (2017)

What is SLAM ?

Simultaneous Localization And Mapping (SLAM)

- A technique for estimating sensor motion and reconstructing structure in an unknown environment.
- Widely used in computer vision, robotics, and AR.



What is SLAM and Visual SLAM ?

visual SLAM:

- Employ the information from images
- 3 categories:
  - Feature-based approach
  - Direct approach
  - RGB-D approach.

Outline

1. Basic and Additional modules of visual SLAM
2. Related technologies
3. Visual SLAM 3 Categories
4. Open problems
5. Conclusion

Basic and Additional Modules

Basic modules:

1. Initialization
2. Tracking
3. Mapping

Additional modules:

1. Relocalization
2. Global map optimization

Basic Modules

1. Initialization
   - Define the global coordinate system
   - A part of the environment is reconstructed as an initial map
2. Tracking
   - The reconstructed map is tracked in the image to estimate the camera pose of the image with respect to the map
   - 2D–3D correspondences between the image and the map are first obtained from feature matching or feature tracking in the image.

Modules

3. Mapping
    - Expand the map by computing the 3D structure of an environment when the camera observes unknown regions
4. Relocalization
    - Compute the camera pose with respect to the map again when the tracking is failed.
    - As kidnapped robot problems in robotics

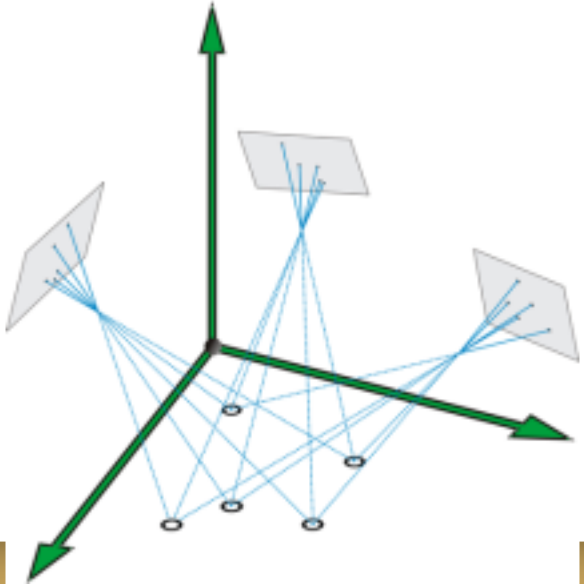
Additional Modules

5. Global map optimization

- To suppress accumulative estimation error
- Loop closing technique
  - When a map is revisited such that a starting region is captured again after some camera movement
- Pose-graph optimization
  - The relationship between camera poses is represented as a graph
  - The consistent graph is built to suppress the error in the optimization
  - (Kümmerle R, Grisetti G, Strasdat H, Konolige K, Burgard W (2011) g2o: A general framework for graph optimization. In: Proceedings of International Conference on Robotics and Automation. pp 3607–3613)

Additional Modules

- Bundle Adjustment (BA)
  - Minimize the reprojection error of the map by optimizing both the map and the camera poses. (Bundle adjustment a modern synthesis. In: Triggs B, McLauchlan PF, Hartley RI, Fitzgibbon AW (eds) (2000) Vision algorithms: theory and practice. pp 298–372)



Related technologies

Visual odometry (VO)

- Estimate the sequential changes of camera positions over time using sensors.
- Visual SLAM = VO + global map optimization
- Geometric consistency of a map is considered only in a small portion of a map
- Only relative camera motion is computed withoutmapping.

Related technologies

Structure from motion (SfM)

- Estimate camera motion and 3D structure of the environment in a batch manner. (Agarwal S, Furukawa Y, Snavely N, Simon I, Curless B, Seitz SM, Szeliski R (2011) Building rome in a day. Commun ACM 54(10):105–112)
- No need to be real-time
- Computer vision
- Goal: Reconstruction

Visual SLAM

- Robotics
- Goal: Real-time nevigation

Visual SLAM 3 Categories

Feature Point-based Approach

- Employ handcrafted feature detectors and descriptors
- Provide stable estimation results in rich textured environments.
- Difficult to handle curved edges and other complex cues by using such handcrafted features.

MonoSLAM

MonoSLAM (Davison AJ (2003) Real-time simultaneous localisation and mapping with a single camera. In: Proceedings of International Conference on Computer Vision. pp 1403–1410)

- Map initialization is done by using a known object
- Camera motion and 3D structure of an unknown environment are simultaneously estimated using an extended Kalman filter (EKF)
- Computational cost that increases in proportion to the size of an environment.

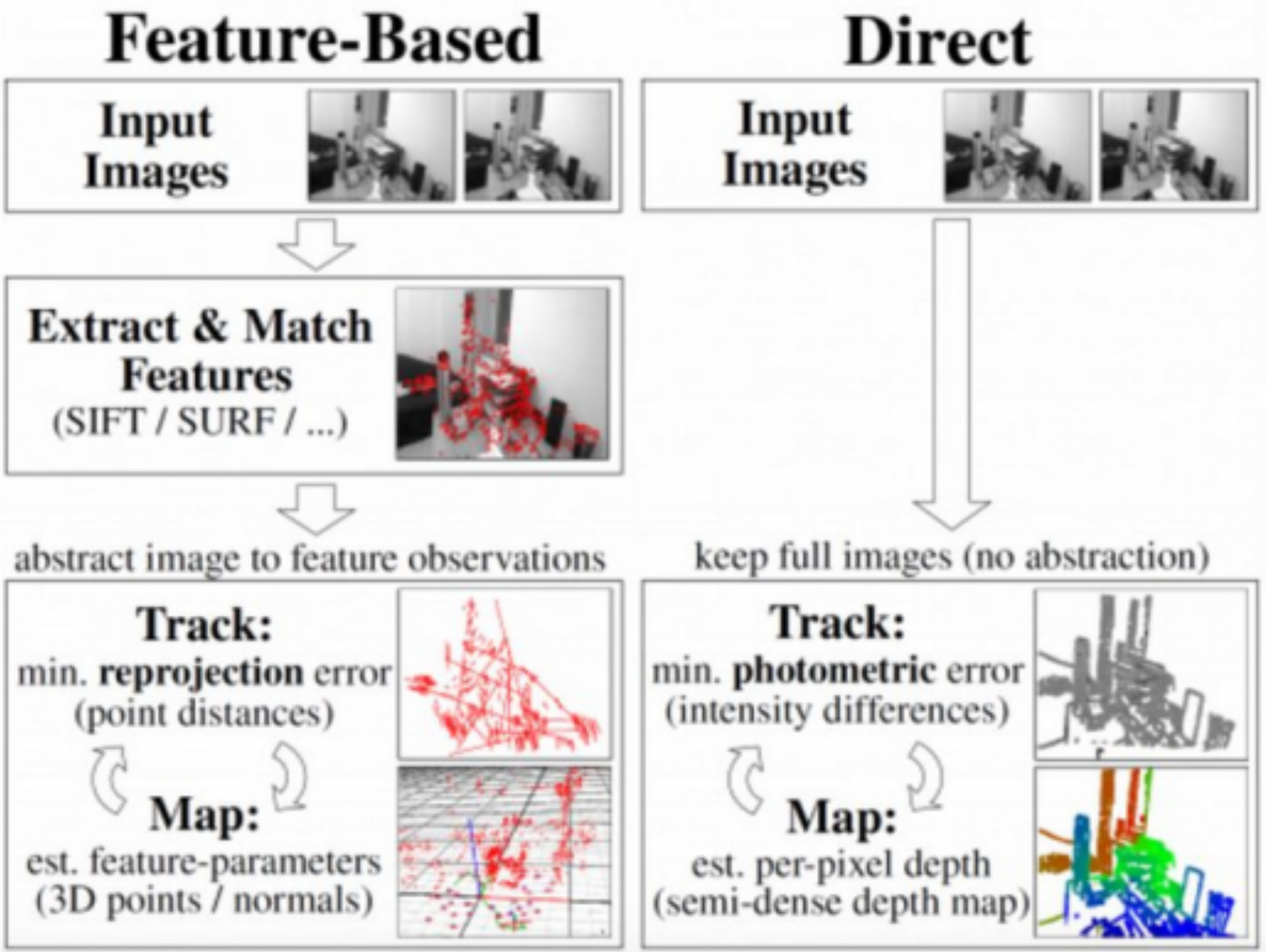
PTAM

PTAM (Klein G, Murray DW (2007) Parallel tracking and mapping for small AR workspaces. In: Proceedngs of International Symposium on Mixed and Augmented Reality. pp 225–234)

- Split the tracking and the mapping into different threads on CPU and executed in parallel.
- Camera poses are estimated from matched feature points between map points and the input image.
- 3D positions of feature points are estimated by triangulation, and estimated 3D positions are optimized by BA.

Direct Approach

- Directly use an input image without any abstraction using handcrafted feature detectors and descriptors.
- Using photometric consistency as an error measurement (Geometric consistency such as positions of feature points in an image is used in feature-based methods)


Feature-Based
Direct
Input Images
Input Images
Extract & Match Features
(SIFT / SURF / ...)
abstract image to feature observations
keep full images (no abstraction)
Track:
min. reprojection error
(point distances)
Map:
est. feature-parameters
(3D points / normals)
Track:
min. photometric error
(intensity differences)
Map:
est. per-pixel depth
(semi-dense depth map)

DTAM (Newcombe RA, Lovegrove SJ, Davison AJ (2011) DTAM: dense tracking and mapping in real-time. In: Proceedings of International Conference on Computer Vision. pp 2320–2327)

- The tracking is done by comparing the input image with synthetic view images generated from the reconstructed map.
- Camera motion is estimated by synthetic view generation from the reconstructed map.
- Depth information is estimated for every pixels by using multi-baseline stereo, and then, it is optimized by considering space continuity

LSD-SLAM (Engel J, Sturm J, Cremers D (2013) Semi-dense visual odometry for a monocular camera. In: Proceedings of International Conference on Computer Vision. pp 1449–1456)

- Random values are set as an initial depth value for each pixel.
- Camera motion is estimated by synthetic view generation from the reconstructed map.
- Reconstructed areas are limited to high-intensity gradient areas

RGB-D Approach

- By using RGB-D cameras, 3D structure of the environment with its texture information can be obtained directly
- Suitable for indoor environment
- Iterative Closest Point (ICP) algorithm have widely been used to estimate camera motion.
- The 3D structure of the environment is reconstructed by combining multiple depth maps.

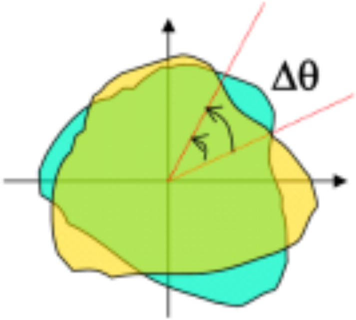
KinectFusion (Newcombe RA, Izadi S, Hilliges O, Molyneaux D, Kim D, Davison AJ, Kohi P, Shotton J, Hodges S, Fitzgibbon A (2011) KinectFusion: real-time dense surface mapping and tracking. In: Proceedngs of International Symposium on Mixed and Augmented Reality. pp 127–136)

- The 3D structure of the environment is reconstructed by combining obtained depth maps in the voxel space
- Camera motion is estimated by the ICP algorithm using an estimated 3D structure and the input depth map, which is depth-based vSLAM.
- KinectFusion is implemented on GPU to achieve real-time processing.

SLAM++ (Salas-Moreno RF, Newcombe RA, Strasdat H, Kelly PHJ, Davison AJ (2013) SLAM++: simultaneous localisation and mapping at the level of objects. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. pp 1352–1359)

- Several 3D objects are registered into the database in advance
- These objects are recognized in an online process.
- By recognizing 3D objects, the estimated map is refined, and 3D points are replaced by 3D objects to reduce the amount of data.

Open problems



- Pure rotation
  - Disparities cannot be observed during purely rotational motion with monocular vSLAM.
  - Not a problem in RGB-D vSLAM since tracking and mapping processes can be done by using obtained depth maps.
- Map initialization
  - Reference objects such as fiducial markers and known 3D objects have been used to get a global coordinate system
  - Initial camera poses are estimated by tracking reference objects.

Open problems

- Estimating intrinsic camera parameters
  - Camera calibration should be done before
  - Intrinsic camera parameters should be fixed during vSLAM estimation process.
- Rolling shutter distortion
  - Most vSLAM algorithms assume a global shutter, and estimate one camera pose for each frame.
  - However, most consumer cameras employ rolling shutter due to its cost.
  - Each row of a captured image is taken by different camera poses

Open problems

- Scale ambiguity
    - Absolute scale information is needed in some vSLAM applications with monocular vSLAM.
    - To obtain absolute scale information, user’s body is used in the literature

Conclusion

SLAM includes 5 modules:

- Initialization, tracking, mapping, relocalization and global map optimization

Related technologies:

- Structure from Motion (SfM) and Visual odometry (VO)

SLAM methods can be classified into 3 categories:

- Feature-based, direct, RGB-D methods.

Open problems:

Pure rotation, map initialization, estimating intrinsic camera parameters, rolling shutter distortion, scale ambiguity

Q&A