S-PTAM stereo parallel tracking and mapping

Sunday, December 27, 2020 8:07 AM

Notation:

SE(3) transformation:

$$T = \begin{bmatrix} R & t \\ 0 & 1 \end{bmatrix}$$

E^{cw}: transformation representing a camera pose that transforms a point in world frame

$$x^w = [x^w \ y^w \ z^w \ 1]^T$$

to a point in camera frame

$$x^c = [x^c \ y^c \ z^c \ 1]^T$$

that is

$$x^c = E^{cw} x^w$$

Motion matrix M: a 4x4 matrix $\in SE(3)$ representing the changes in camera pose by left-multiplication:

$$E^{cw} = M^c E^{cw}_{prev}$$

In Lie Groups, the motion matrix M can be represented by a six-vector

$$\mu = (t_x, t_y, t_z, \theta^{roll}, \theta^{pitch}, \theta^{yaw})$$

μ, M are related by

$$M = \exp(\mu) = e^{\sum_{j=1}^{6} \mu_j G_j}$$

$G_j, j=1\dots6$ is the group generator matrix.

Measurement $z = \begin{bmatrix} u \\ v \end{bmatrix}$, is the true 2D position that matches with the projected 3D point on the image plane.

Map point: p, is ordered pair (x^w, d), contains
- 3D point x^w
- its descriptor d

Stereo keyframe k is a stereo pair of images with the associated stereo camera pose.

Map is defined as the set of map points and the set of stereo keyframes.

A point in camera frame x^c projects into the image as

$$\begin{bmatrix} \hat{u} \\ \hat{v} \end{bmatrix} = P(x^c)$$

$$P(x^c) = \begin{bmatrix} f_u & 0 & u_0 \\ 0 & f_v & v_0 \end{bmatrix} \begin{bmatrix} \frac{x^c}{z^c} \\ \frac{y^c}{z^c} \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{f_u x^c}{z^c} + u_0 \\ \frac{f_v y^c}{z^c} + v_0 \end{bmatrix}$$

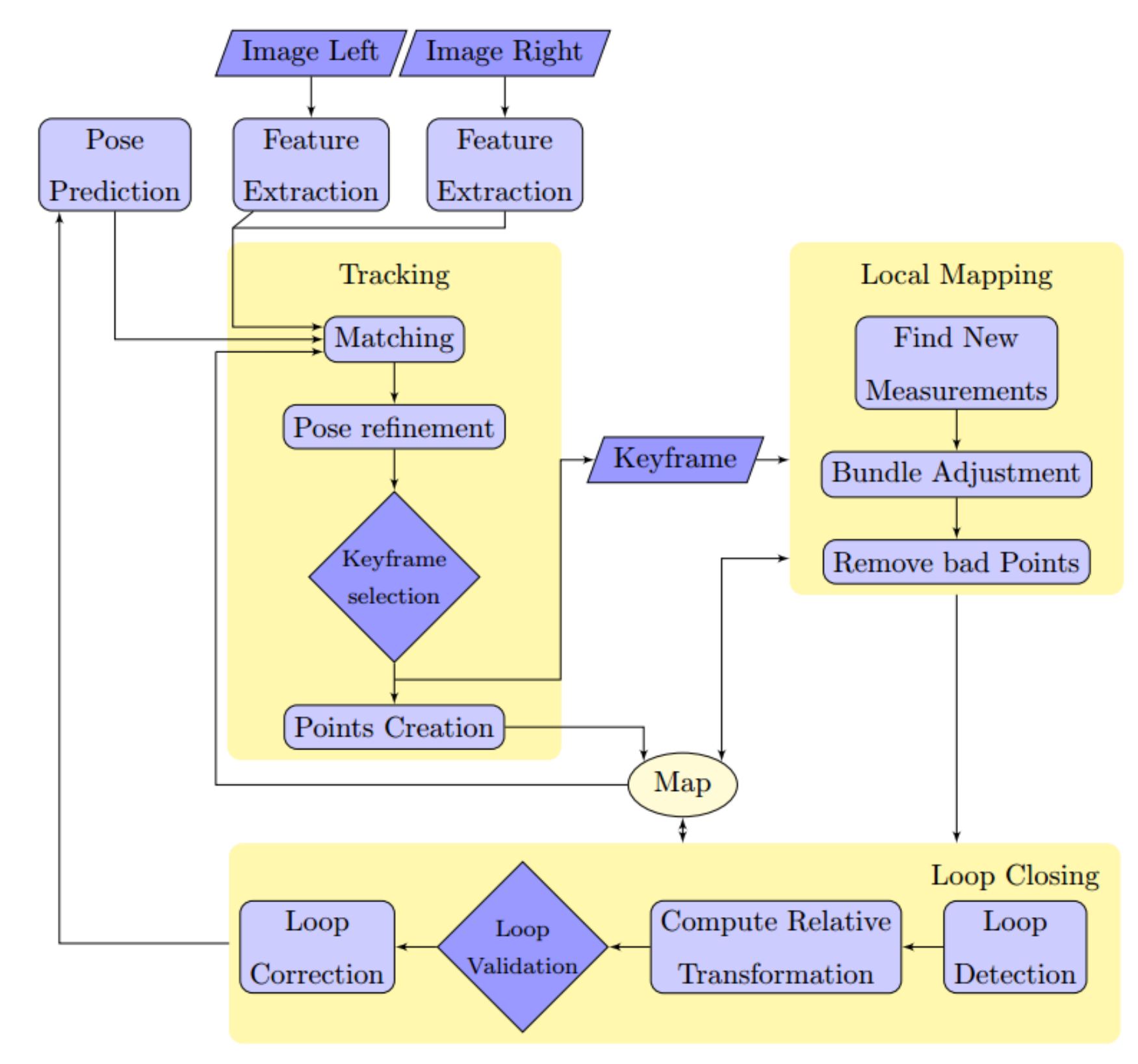
The global reference frame is the camera pose at the first frame.

An initial map is estimated by matching and triangulating salient point features in the first stereo pair.

For every frame after the tracking thread estimates the 6DOF pose for each stereo frame, by minimizing the re-projection error between the projected map points and their correspondences.

The system selects a subset of keyframes that will be used in a second thread to estimate the map at a lower rate.
The map points are triangulated from the stereo matches of each keyframe, and added to the map.

The mapping thread is constantly minimizing the local reprojection error by refining all the map points, and the stereo poses using Bundle Adjustment.
We use a pose graph to maintain the global consistency of the map.



Feature extraction and description.
- Keypoint detector: GFTT
- keypoint descriptor: BRISK.

Pose tracking:
- Matching.
  - Project each map point ins he viewing frustum of the predicted stereo pose, and search for a match in the neighborhood of the point.
  - a decaying velocity model is used for pose prediction.
  - binary desriptors are used, and the Hamming distance is calculated.

- Pose refinement
  - E^{cw}: current camera pose in global frame w.
  - Previous camera pose E^{cw}_{prev} with the relative motion M^c in the local camera frame.

$$E^{cw} = M^c E^{cw}_{prev}$$

  - To find the M^c,

$$J\mu = \Delta z(\mu_{prev}) \quad (6)$$

$\mu = (t_x, t_y, t_z, \theta_{roll}, \theta_{pitch}, \theta_{yaw})^T$.
Δz: re-projection error that only depends on the camera motion μ, as we consider the map fixed,
J: Jacobian of reprojection error wrt camera motion.

$$J_{ij} = \frac{\partial \Delta z_i(\mu)}{\partial \mu_j} = \frac{\partial \left(\begin{bmatrix} u \\ v \end{bmatrix}_i - P(\exp(\mu) E^{cw}_{prev} x_i^w) \right)}{\partial \mu_j}$$

$$= - \frac{\partial P(x_i^c)}{\partial x_i^c} \frac{\partial x_i^c}{\partial \mu_j}$$

where

$$\frac{\partial P(x_i^c)}{\partial x_i^c} = \begin{bmatrix} \frac{f_u}{z^c} & 0 & -\frac{f_u x^c}{z^{c2}} \\ 0 & \frac{f_v}{z^c} & -\frac{f_v y^c}{z^{c2}} \end{bmatrix}$$

$$\frac{\partial x_i^c}{\partial \mu_j} = G_j E^{cw}_{prev} x_i^w$$

μ is found by solving (6). In order to do this, given a set

$$S = \{z_1, \dots z_n\}$$

of matched measurements, the new value for μ is obtained by minimizing an objective function:

$$\mu' = \arg\min_{\mu} \sum_{i \in S} \rho(J_i \mu - \Delta z_i(\mu_{prev}))$$

$\rho(\cdot)$ is the Huber function, the optimization is done via LM.

- keyframes selection and map points creation.
  - A frame is selected to be a keyframe if the number of tracked points is less than 90% of the points tracked in last keyframe.
  - The remaining unmatched features from the stereo pair are triangulated to create new map points.
  - The keyframe is quened into the map refinement thread

- local mapping.
  - The refinement of the keyframe poses, and the 3D map points, is done via BA, that minimizes the reprojection error of every point in every image.
  - The problem can be stated as follows:

Given an initial set of N keyframe poses $\{E_1, \dots E_N\}$, an initial set of M 3D points $x^w = \{x_1^w, \dots, x_M^w\}$, and measurements $\{S_1, \dots, S_N\}$, each set S_j contains the measurements z_{ij} of the i-th point in the j-th keyframe, the simultaneous estimation of the multiple cameras and the point cloud is achieved by solving

$$J \begin{bmatrix} \mu \\ x \end{bmatrix} = \Delta z(\mu_{prev}, x^w_{prev})$$

$$\Delta z(\mu, x^w) = z - P(\exp(\mu) E^{cw}_{prev} x^w)$$

We must minimize

μ_1 is fixed in BA, as it defines the world frame.

$$\{\mu'_{j=2\dots N}, x'^w_{i=1\dots M}\} = \arg\min_{\{\{\mu\}, \{x^w\}\}} \sum_{j=1}^{N} \sum_{i \in S_j} \rho_{6T}(\psi_{ji})$$

$$\psi_{ji} = J_{ji} \begin{bmatrix} \mu_j \\ x_i^w \end{bmatrix} - \Delta z_i(\mu_{prev,j}, x^w_{prev,i})$$

Given that the vector of parameters is divided in two groups, (cameras and points), the Jacobian can be decomposed as:

$$J = \left[\frac{\partial \Delta z(\mu, x^w)}{\partial \mu} \middle| \frac{\partial \Delta z(\mu, x^w)}{\partial x^w} \right].$$

$$J_{ji} = \frac{\partial \Delta z(\mu_j, x^w)}{\partial x_i^w}$$

$$= \frac{\partial \left(\begin{bmatrix} u \\ v \end{bmatrix}_i - P(\exp(\mu_j) E^{c_j w}_{prev} x_i^w) \right)}{\partial x_i^w}$$

$$= - \frac{\partial P(x_i^{c_j})}{\partial x_i^{c_j}} \frac{\partial x_i^{c_j}}{\partial x_i^w}$$

$$\frac{\partial x_i^{c_j}}{\partial x_i^w} = \frac{\partial (M^{c_j} E^{c_j w}_{prev} x_i^w)}{\partial x_i^w} = R.$$

We have dealt with the BA in the left image. Adding the stereo constraint, the relative motion between the left and the right cameras is fixed, so we can obtain the pose of the right camera from the left camera:

$$E^{Rw} = E^{RL} M^L E^{Lw}_{prev}$$

Now we can use the right camera measurements to add stereo constraints to BA:

$$z^R = P^R(E^{RL} M^L E^{Lw}_{prev} x^w)$$

The Jacobian rows related to the right camera measurements have the form:

$$J^R_{ji} = \frac{\partial \Delta z^R(\mu_j, x_i^w)}{\partial x_i^w} \quad (21)$$

$$= \frac{\partial \left(\begin{bmatrix} u^R \\ v^R \end{bmatrix}_i - P^R(E^{RL} \exp(\mu_j) E^{L_j w}_{prev} x_i^w) \right)}{\partial x_i^w}$$

$$= \begin{bmatrix} \frac{f_u}{z^R} & 0 & -\frac{f_u x^R}{z^{R2}} \\ 0 & \frac{f_v}{z^R} & -\frac{f_v y^R}{z^{R2}} \end{bmatrix} R^{RL} R$$

If the stereo camera is rectified, then the transformation between cameras is a pure translation in the x-axis, and the intrinsics are the same, ∴

$$y^L = y^R, \quad z^L = z^R$$

and (21) can be rewritten as

$$J^R_{ji} = \begin{bmatrix} \frac{f_u}{z^L} & 0 & -\frac{f_u x^R}{z^{L2}} \\ 0 & \frac{f_y}{z^L} & -\frac{f_v y^L}{z^{L2}} \end{bmatrix} R$$

$$J_{ji} = \begin{bmatrix} \frac{f_u}{z^L} & 0 & -\frac{f_u x^L}{z^{L2}} \\ 0 & \frac{f_v}{z^L} & -\frac{f_v y^L}{z^{L2}} \\ \frac{f_u}{z^L} & 0 & -\frac{f_u x^R}{z^{L2}} \end{bmatrix} R.$$

PTAM.