Pose Tracking: Structure from Motion and SLAM

CS 6334 Virtual Reality
Professor Yu Xiang
The University of Texas at Dallas
Tracking in VR

• Tracking the user’s sense organs
  • E.g., Head and eye
  • Render stimulus accordingly

• Tracking user’s other body parts
  • E.g., human body and hands
  • Locomotion and manipulation

• Tracking the rest of the environment
  • Augmented reality
  • Obstacle avoidance in the real world
Feature-based Tracking

The PnP problem
- Known: 3D locations, 2D locations, camera intrinsics
- Unknown: 6D pose of the camera

What if we do not have the 3D locations of these feature points?
Feature-based Tracking

- Idea: using images from different views and feature matching

Geometry-aware Feature Matching for Structure from Motion Applications. Shah et al, WACV’15
Feature-based Tracking

• Idea: using images from different views and feature matching

• Triangulation from pixel correspondences to compute 3D location

Intersection of two backprojected lines

\[ X = 1 \times 1' \]

Unknow
Structure from Motion

• Input
  • A set of images from different views

• Output
  • 3D Locations of all feature points in a world frame
  • Camera poses of the images
Structure from motion

\[ \Pi_1 X_1 \sim p_{11} \]

minimize

\[ g(R, T, X) \]

non-linear least squares

Camera 1

\[ R_1, t_1 \]

Camera 2

\[ R_2, t_2 \]

Camera 3

\[ R_3, t_3 \]
Structure from Motion

• Minimize sum of squared reprojection errors

\[ g(X, R, T) = \sum_{i=1}^{m} \sum_{j=1}^{n} w_{ij} \cdot \| P(x_i, R_j, t_j) - [u_{i,j}, v_{i,j}] \|^2 \]

m points, n images

indicator variable: is point i visible in image j?

A non-linear least squares problem

• E.g. Levenberg-Marquardt
The Levenberg-Marquardt Algorithm

• Nonlinear least squares \( \hat{\beta} \in \arg\min_{\beta} S(\beta) \equiv \arg\min_{\beta} \sum_{i=1}^{m} [y_i - f(x_i, \beta)]^2 \)

• An iterative algorithm
  • Start with an initial guess \( \beta_0 \)
  • For each iteration \( \beta \leftarrow \beta + \delta \)

• How to get \( \delta \)?
  • Linear approximation \( f(x_i, \beta + \delta) \approx f(x_i, \beta) + J_i \delta \)
  • Find to \( \delta \) minimize the objective \( S(\beta + \delta) \approx \sum_{i=1}^{m} [y_i - f(x_i, \beta) - J_i \delta]^2 \)
The Levenberg-Marquardt Algorithm

- Vector notation for \( S(\beta + \delta) \approx \sum_{i=1}^{m} [y_i - f(x_i, \beta) - J_i \delta]^2 \)

\[
S(\beta + \delta) \approx \|y - f(\beta) - J\delta\|^2 \\
= [y - f(\beta) - J\delta]^T [y - f(\beta) - J\delta] \\
= [y - f(\beta)]^T [y - f(\beta)] - [y - f(\beta)]^T J\delta - (J\delta)^T [y - f(\beta)] + \delta^T J^T J \delta \\
= [y - f(\beta)]^T [y - f(\beta)] - 2[y - f(\beta)]^T J\delta + \delta^T J^T J \delta.
\]

Take derivation with respect to \( \delta \) and set to zero \( (J^T J) \delta = J^T [y - f(\beta)] \)

Levenberg's contribution \( (J^T J + \lambda I) \delta = J^T [y - f(\beta)] \) damped version

\[
\beta \leftarrow \beta + \delta
\]
Structure from Motion

\[
g(X, R, T) = \sum_{i=1}^{m} \sum_{j=1}^{n} w_{ij} \cdot \left\| P(x_i, R_j, t_j) - \left[ \begin{array}{c} u_{i,j} \\ v_{i,j} \end{array} \right] \right\|^2
\]

indicator variable: is point \( i \) visible in image \( j \) ?

\[
\beta = (X, R, T)
\]

How to get the initial estimation \( \beta_0 \) ?

Random guess is not a good idea.
Matching Two Views

• Fundamental matrix

\[ x' \text{ is on the epipolar line } \quad 1' = Fx \]

\[ x'TFx = 0 \]

\[
\begin{bmatrix}
  x_i' & y_i' & 1
\end{bmatrix}
\begin{bmatrix}
  f_{11} & f_{12} & f_{13} \\
  f_{21} & f_{22} & f_{23} \\
  f_{31} & f_{32} & f_{33}
\end{bmatrix}
\begin{bmatrix}
  x_i \\
  y_i \\
  1
\end{bmatrix} = 0
\]

\[ x_ix_i'f_{11} + x_iy_i'f_{21} + x_if_{31} + y_ix_i'f_{12} + y_iy_i'f_{22} + y_if_{32} + x_i'f_{13} + y_i'f_{23} + f_{33} = 0 \]

We need 8 points to solve this system.
Matching Two Views

• Essential matrix $E$

\[
x'^T F x = 0
\]

\[
(K'^{-1} x')^T E (K^{-1} x) = 0
\]

\[
F = K'^{-T} E K^{-1}
\]

Credit: Thomas Opsahl
Matching Two Views

• In 1981 H. C Longuet-Higgins proved that one could recover the relative pose $R$ and $t$ from the essential matrix $E$ up to the scale of $t$

Triangulation

Intersection of two backprojected lines

$$X = l \times l'$$

How to get the initial estimation $\beta_0$?

$$\beta = (X, R, T)$$

Estimated from essential matrix $E$
Structure from Motion

• Bundle adjustment
  • Iteratively refinement of structure (3D points) and motion (camera poses)

• Levenberg-Marquardt algorithm

\[ g(X, R, T) = \sum_{i=1}^{m} \sum_{j=1}^{n} w_{ij} \cdot \left\| P(x_i, R_j, t_j) - \left[ \begin{array}{c} u_{i,j} \\ v_{i,j} \end{array} \right] \right\|^2 \]

Examples: [http://vision.soic.indiana.edu/projects/disco/](http://vision.soic.indiana.edu/projects/disco/)
Basics

• Image feature matching
Harris Corner Detector

• Corners are regions with large variation in intensity in all directions

“flat” region: no change in all directions

“edge”: no change along the edge direction

“corner”: significant change in all directions
Harris Corner Detector

\[ f(\Delta x, \Delta y) = \sum_{(x_k, y_k) \in W} (I(x_k, y_k) - I(x_k + \Delta x, y_k + \Delta y))^2 \]

Taylor expansion

\[ I(x + \Delta x, y + \Delta y) \approx I(x, y) + I_x(x, y)\Delta x + I_y(x, y)\Delta y \]

\[ f(\Delta x, \Delta y) \approx \sum_{(x, y) \in W} (I_x(x, y)\Delta x + I_y(x, y)\Delta y)^2 \]

\[ f(\Delta x, \Delta y) \approx \begin{pmatrix} \Delta x & \Delta y \end{pmatrix} \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix} \]

\[ M = \sum_{(x,y) \in W} \begin{bmatrix} I_x^2 & I_x I_y \\ I_x I_y & I_y^2 \end{bmatrix} = \begin{bmatrix} \sum_{(x,y) \in W} I_x^2 & \sum_{(x,y) \in W} I_x I_y \\ \sum_{(x,y) \in W} I_x I_y & \sum_{(x,y) \in W} I_y^2 \end{bmatrix} \]

\[ R = \det(M) - k(\text{trace}(M))^2 \]

- \( \det(M) = \lambda_1 \lambda_2 \)
- \( \text{trace}(M) = \lambda_1 + \lambda_2 \)
- \( \lambda_1 \) and \( \lambda_2 \) are the eigenvalues of \( M \)
Harris Corner Detector

https://docs.opencv.org/master/dc/d0d/tutorial_py_features_harris.html
Invariance

• Can the same feature point be detected after some transformation?
  • Translation invariance

• 2D rotation invariance

• Scale invariance

Are Harris corners scale invariance?

No
Scale Invariance Feature Transform (SIFT)

• Keypoint detection

• Compute descriptors

• Matching descriptors
SIFT: Scale-space Extrema Detection

• How to detect keypoints?
  • E.g., applying a second derivative of Gaussian kernel to an image (Laplacian of Gaussian)

\[ G(x, y, \sigma) = \frac{1}{2\pi \sigma^2} e^{-(x^2+y^2)/2\sigma^2} \]

Scale \( \sigma \)
In pixels, radius of the kernel
SIFT: Scale-space Extrema Detection

Maxima and minima of DOG images

$L(x, y, \sigma) = G(x, y, \sigma) \ast I(x, y)$

$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{-\frac{(x^2+y^2)}{2\sigma^2}}$

$D(x, y, \sigma) = (G(x, y, k\sigma) - G(x, y, \sigma)) \ast I(x, y)$

$= L(x, y, k\sigma) - L(x, y, \sigma)$.
SIFT Descriptor

- Divide the 16x16 window into a 4x4 grid of cells (2x2 case shown below)
- Compute an orientation histogram for each cell
- 16 cells * 8 orientations = 128 dimensional descriptor
SIFT: Rotation Invariance

• Rotate all orientations by the dominant orientation
SIFT Matching Example
Simultaneous Localization and Mapping (SLAM)

- Localization: camera pose tracking
- Mapping: building a 2D or 3D representation of the environment
- The goal here is the same as structure from motion, usually with video input

ORB-SLAM2
- Point cloud and camera poses
ORB-SLAM

- Oriented FAST and Rotated BRIEF (ORB)
- Tracking camera poses
  - Motion only Bundle Adjustment (BA)
- Mapping
  - Local BA around camera pose
- Loop closing
  - Loop detection

https://webdiis.unizar.es/~raulmur/orbslam/
3D Scanning

- Using laser to create “point clouds”

Figure 9.26: (a) The Afinia ES360 scanner, which produces a 3D model of an object while it spins on a turntable. (b) The Focus3D X 330 Laser Scanner, from FARO Technologies, is an outward-facing scanner for building accurate 3D models of large environments; it includes a GPS receiver to help fuse individual scans into a coherent map.
3D Scanning

https://matterport.com/
Further Reading

• Section 9.5, Virtual Reality, Steven LaValle

• SIFT: Distinctive Image Features from Scale-Invariant Keypoints, David Lowe, IJCV’04

• ORB-SLAM: ORB-SLAM: a Versatile and Accurate Monocular SLAM System, Mur-Artal et al., T-RO’15