

Computer Science 5722 - Computer Vision

Assignment 3: Visual Odometry: Motion from correspondences

Due : in class April 1

Visual Odometry We can estimate the trajectory of a moving camera based on the image motion arising from it. In this assignment you will extend your implementation of a simplified version of the work of Konolige et al., to estimate the trajectory of a robot in an outdoor environment. Although there are numerous approaches to VO, this assignment will use monocular interest point feature tracking, estimation of pose using RANSAC on the 8 point algorithm, and incremental bundle adjustment.

CenSurE Features. In the previous assignment you implemented Konolige’s CenSurE features. In this next step we use the features and their frame to frame correspondences to estimate the world motion of the camera.

Consensus correspondence. You will use RANSAC with the 8 point algorithm for estimating the Essential matrix E which describes the epipolar geometry and therefore the motion of the robot for a pair of frames (viewpoints) in the sequence.

The 8 point algorithm works as follows:

- We know that for normalized coordinates: $y_2^T E y_1 = 0$.
- We obtain normalized coordinates y by transforming pixel correspondences $p=[u,v,I]$ back to the image plane using calibrated camera intrinsics focal length f (in pixels) and image center C .

$$y = K^{-1}p = \begin{bmatrix} f & 0 & C(1) \\ 0 & f & C(2) \\ 0 & 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix}.$$

- Hartley’s paper (above) recommends further normalizing the points $y=[y_x,y_y,I]$ for each image, by translating their centroid to the origin then scaling them so their average distance to the origin is $\sqrt{2}$. This gives 2 transforms T_1 and T_2 such that $\hat{y}_1 = T_1 y_1$ and $\hat{y}_2 = T_2 y_2$, which must be undone once the E matrix is estimated by $T_2^T \hat{E} T_1$.
- Use these normalized y to construct a linear system of equations eg for each corresponding pair $y_1 = [y_{1x}, y_{1y}, y_{1z}]$ and $y_2 = [y_{2x}, y_{2y}, y_{2z}]$ we have equation:

$$y_{1x}y_{2x}e_{11} + y_{1x}y_{2y}e_{21} + y_{1x}e_{31} + y_{2x}y_{1y}e_{12} + y_{1y}y_{2y}e_{22} + y_{1y}e_{32} + y_{2x}e_{13} + y_{2y}e_{23} + e_{33} = 0$$

- We want to solve for the elements e_{ij} of E , so we construct a matrix A_E from our correspondences. To estimate E we compute the singular value decomposition (SVD) $[U, S, V] = \text{svd}(A_E^T A_E)$, the rightmost column of U (or V) is reshaped to form our estimate of E .
- To enforce that E is rank 2 and has 2 equal nonzero singular values we can compute $[\mu, \sigma, \nu] = \text{SVD}(E)$ and substitute

$$\tilde{s} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

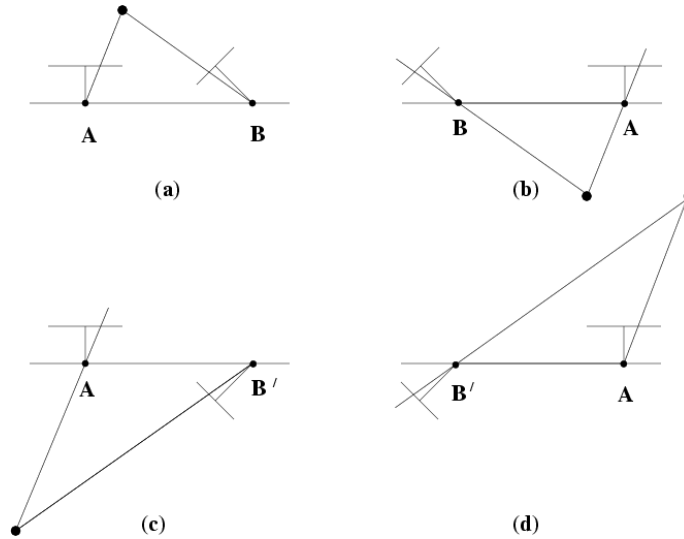
to estimate $\tilde{E} = T_2^T \mu \tilde{s} v^T T_1$.

- We know that $E = R[t_x]$, so we can extract the motion $[Rt]$ between frames.

– Let

$$D = \begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

- For $[u, s, v] = SVD(E)$, choose t to be the 3rd column of u , $Ra = uDv^T$ and $Rb = uD^T v^T$. Any combination of R and t will satisfy the epipolar constraint, but will not necessarily be physically valid.
- Assuming that the first camera matrix is $[I|0]$ and that t is unit length, there are 4 possible solutions for the second camera $P_1 = [Ra|t]$, $P_2 = [Ra|-t]$, $P_3 = [Rb|t]$ and $P_4 = [Rb|-t]$. These represent 4 possible solutions for reconstruction from E :



- We must determine the actual configuration. To do this reconstruct the 9 points (corresponding pairs) using $([I|0], P_1)$ as the camera pair.
- A linear method for triangulating (reconstructing) a point given a pair of camera matrices and the corresponding (normalized) image points works as follows:
 - * Given a pair of camera matrices M_1 and M_2 , and a corresponding pair of image points y_1 and y_2 arising from world point X , we know $y_1 = M_1 X$ and $y_2 = M_2 X$.
 - * We eliminate the homogeneous scale factor each image using the cross product $y \times (MX) = 0$ giving:

$$\begin{aligned} y_x(M^{3T} X) - (M^{1T} X) &= 0 \\ y_y(M^{3T} X) - (M^{2T} X) &= 0 \\ y_x(M^{2T} X) - y_y(M^{1T} X) &= 0 \end{aligned}$$

where M^{iT} is i th row of M .

- * We can use 2 equations from each image to construct the linear system $Ax = 0$ with

$$A = \begin{bmatrix} y_{1x}M_1^{3T} - M_1^{1T} \\ y_{1y}M_1^{3T} - M_1^{2T} \\ y_{2x}M_2^{3T} - M_2^{1T} \\ y_{2y}M_2^{3T} - M_2^{2T} \end{bmatrix}$$

- Solve for the homogeneous coordinates $X=[x,y,z,w]$ given $AX=0$, using SVD.
- Test cheirality¹ for each of the 9 points using: $c1 = zw$, $c2 = dw$, for $[a, b, d] = P_1X$. If $c1 < 0$ the point is behind cam1, and if $c2 < 0$ the point is behind cam2. If $c1 > 0$ and $c2 > 0$ then P_1 is the correct camera matrix. If $c1 < 0$ and $c2 < 0$ then P_2 is selected. If $c1c2 < 0$ then use

$$H = \begin{bmatrix} I_3 & 0 \\ -2v_{13} & -2v_{23} & -2v_{33} & -1 \end{bmatrix}$$

to compute $[p, q, r, n] = HX$, and test $zn > 0$, and if so select P_3 , otherwise P_4 . Each point will vote on one of the 4 configurations.

- Each reconstructed point votes for one of the 4 configurations based on these tests. If one configuration has 5 or more votes, choose the appropriate P , otherwise discard this estimate of E .

RANSAC : basically selects models by randomly sampling a small number of points and estimating parameters. The model which best fits (most inliers) the observed data is chosen. In our case:

- Repeat for N samples:
 - Randomly select 9 corresponding pairs.
 - Fit the Essential matrix using the least squares technique above.
 - Reconstruct all the correspondences and compute a *reprojection error* distance metric for each corresponding pair $[y_1, y_2]$:

$$d(y_1, \hat{y}_1)^2 + d(y_2, \hat{y}_2)^2$$

where d is the Euclidean distance in the image between the original correspondence and the projection of the reconstructed point.

- Count the number of inliers with distance less than some threshold t .
- Choose the E with the greatest number of inliers, use all the inliers to reestimate E , and the relative motion $[R \ t]$ between the frames (viewpoints).

Download the zip file available [here](#).

Create a MATLAB program (a .m file containing MATLAB commands) which does the following:

¹Constrains world points to be in front of cameras

1. For each sequential pair of frames i and $i+1$ in the image sequence compute a list of correspondences.
2. Use the RANSAC approach to find Essential matrix E which best fits the observed correspondences.
3. Record the associated motion for each pair of frames.
4. Concatenate the motions to plot the motion of the robot. For motions $T_i = [R_i \ t_i]$, if we want the i th camera position in the frame of the first camera position we concatenate:

$$P_{i,1} = T_1^{-1} T_2^{-1} \dots T_i^{-1} \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}$$

Submit your completed program, .mat file and a brief explanation of the threshold and design choices you made to Wei Xu according to his instructions, by 11am (class time) on the due date.