Vision-based Navigation Exercise 3

Kerem Yildirir

May 2021

1 Part 1

Consider a 3D point X and it's correspondent 2D points x_1 and x_2 in two different images and for simplicity, pinhole camera model. When we unproject the point $x1$, we get a normalized vector $\overrightarrow{x_1}$. Due to scale ambiguity, we say that this vector needs to be scaled by factor λ_1 to be equal to original 3D point X.

$$\lambda_1 \overrightarrow{x}_1 = \mathbf{X}$$

We assume that camera rotated by $\mathbf{R}$ and translated by $\mathbf{t}$ then observed x_2. Doing the same thing as above to x_2:

$$\lambda_2 \overrightarrow{\mathbf{x}}_2 = R\mathbf{X} + \mathbf{t}$$

Inserting X from above equation,

$$\lambda_2 \overrightarrow{\mathbf{x}}_2 = R\lambda_1 \overrightarrow{x_1} + \mathbf{t}$$

Cross product with t from left leads to:

$$\lambda_2 t \times \overrightarrow{x_2} = t \times R\lambda_1 \overrightarrow{x_1}$$

Since $t \times \overrightarrow{x_2}$ results in a vector perpendicular to $\overrightarrow{x_2}$, applying dot product from left makes left hand side 0.

$$x_2^T t \times \overrightarrow{x_2} = 0 = \overrightarrow{x_2}^T t \times R\lambda_1 \overrightarrow{x_1}$$

We can omit λ_1 and represent $t\times$ with the skew-symmetric matrix $\hat{T}$built from t:

$$0 = \overrightarrow{x_2}^T \hat{T} R \overrightarrow{x_1}$$

This equation is the epipolar constraint, and $E = \hat{T}R$ is the essential matrix.

2 Part 2

When we do brute force matching over all frames, we consider all pairwise combinations minus stereo pairs. This leads to:

$$\frac{N * (N - 1)}{2} - N$$

For 1000 images we need to evaluate 498500 pairs. When BoW is used, this number is limited by number of returned query images retrieved from the database. For Q returned images, this becomes at most

$$\frac{Q * (Q - 1)}{2} - Q$$