ENPM673 – Perception for Autonomous Robots – Spring 2022

Harika Pendli

M.Eng. Robotics

University of Maryland, College Park

Email: hpendli@umd.edu

This project aims to create a stereo vision system such that given a stereo pair of images, the output should be color and grayscale images of disparity and depth map. To achieve this the task is divided into four parts:

The first part is the calibration part-

1. We first compare the pair of stereo images and find the matching pairs. This has been achieved by using ORB feature matching. The best matches are then chosen based on distance.

2. To fill the final list of feature points, points are extracted from the matches and the matches are plotted simultaneously on both images to visualize them.

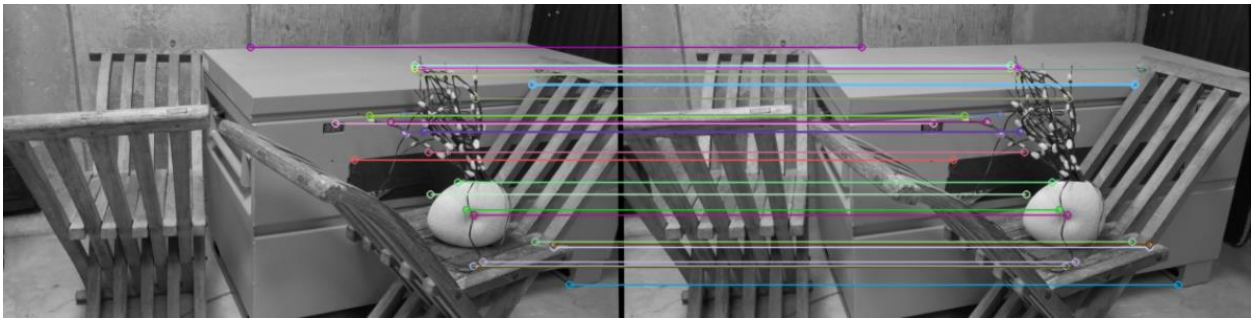

Fig: Feature matching between stereo pair

3. We then calculate the fundamental matrix using SVD method. We have 8 equations and we input exactly 8 points to form the A matrix from them (Eight point algorithm). The fundamental matrix, denoted by F, is a 3×3 (rank 2) matrix that relates the corresponding set of points in two images from different views (or stereo images).

$$\begin{bmatrix} x_1x'_1 & x_1y'_1 & x_1 & y_1x'_1 & y_1y'_1 & y_1 & x'_1 & y'_1 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_mx'_m & x_my'_m & x_m & y_mx'_m & y_my'_m & y_m & x'_m & y'_m & 1 \end{bmatrix} \begin{bmatrix} f_{11} \\ f_{21} \\ f_{31} \\ f_{12} \\ f_{22} \\ f_{32} \\ f_{13} \\ f_{23} \\ f_{33} \end{bmatrix} = 0$$

Form: Ax=0

Here, the RANSAC algorithm is applied to choose the best F matrix. 8 random points are sampled and the Fundamental matrix is calculated iteratively and we calculate the error in estimating the points. Comparing this with a predefined threshold, we can classify the inlier points and choose the Best F matrix based on the chosen criteria.

4. We now derive the essential matrix E from the Fundamental matrix F and camera intrinsic matrix K as follows $E = K^T FK$. Clearly, the essential matrix can be extracted from F and K. As in the case of F matrix computation, the singular values of E are not necessarily (1,1,0) due to the noise in K. This can be corrected by reconstructing it with (1,1,0) singular values, i.e.

$$\mathbf{E} = U \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} V^T$$

It is important to note that the F is defined in the original image space (i.e., pixel coordinates) whereas E is in the normalized image coordinates. Normalized image coordinates have the origin at the optical center of the image. Also, relative camera poses between two views can be computed using E matrix. Moreover, F has 7 degrees of freedom while E has 5 as it takes camera parameters in account.

5. After extracting the essential matrix E, we need to compute the camera pose. The camera pose consists of 6 degrees-of-freedom (DOF) Rotation (Roll, Pitch, Yaw) and Translation (X, Y, Z) of the camera with respect to the world. Since the E matrix is identified, the four-camera pose configurations: (C1,R1),(C2,R2),(C3,R3) and (C4,R4) where $C \in R^3$ is the camera center and $R \in SO(3)$ is the rotation matrix, can be computed. Thus, the camera pose can be written as: $P=KR[I_{3x3} \; -C]$ These four pose configurations can be computed from E matrix. Let $E=UDV^T$ and W be:

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

The four configurations can be written as:

$C1=U(:,3)$ and $R1=UWV^T$

$C2=-U(:,3)$ and $R2=UWV^T$

$C3=U(:,3)$ and $R3=UW^T V^T$

$C4=-U(:,3)$ and $R4=UW^T V^T$

It is important to note that the det(R)=1. If det(R)=−1, the camera pose must be corrected i.e. C=−C and R=−R. Implementing this will return list of translation and rotation solutions for each camera pose possible.

To find the correct unique camera pose, we need to remove the disambiguity. This can be accomplished by checking the cheirality condition i.e. the reconstructed points must be in front of the cameras. To check the cheirality condition, triangulate the 3D points (given two camera poses) using linear least squares to

check the sign of the depth Z in the camera coordinate system w.r.t. camera center. A 3D point X is in front of the camera iff: r3(X−C)>0 where r3 is the third row of the rotation matrix (z-axis of the camera). Not all triangulated points satisfy this condition due of the presence of correspondence noise. The best camera configuration, (C,R,X) is the one that produces the maximum number of points satisfying the cheirality condition.

The next step is the rectification part:

1. Here we warp the images to make the epipolar lines horizontal (epipoles at infinity), so that disparity calculation is restricted to one dimension only. Essence of this step is to make the camera setups parallel after homography transformation.

2. This is achieved by using in-built cv2 stereo rectify function. Here, we input the fundamental matrices, image dimensions and features points, while it returns the H1, H2 matrices. Finally we warp both the images using H matrices to get horizontal epilines.

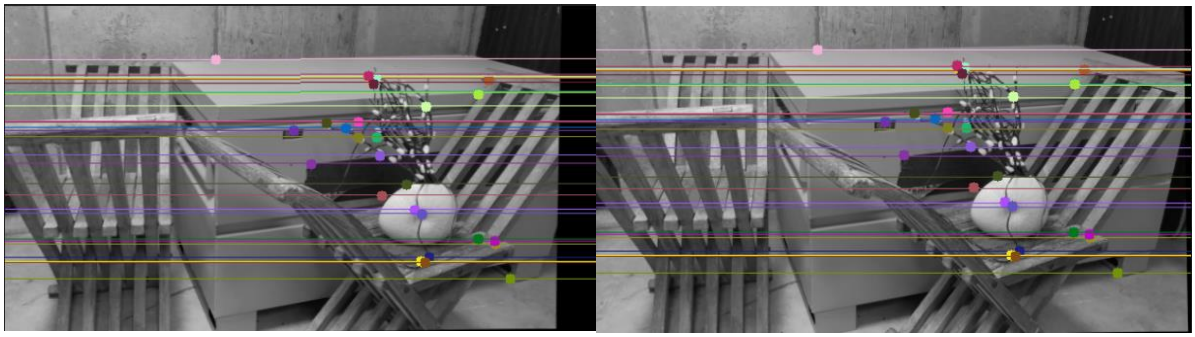

Fig: Left and right image epilines and correpondances(gray scale)

Third step is the Correspondence part:

1. For each pixel in the image we calculate disparity map using a sliding window and sum of squared differences approach and save the grayscale and heatmap images. We can also use the sum of absolute difference approach.

2.We search for the best matching pixel by sliding window over the line and calculating SSD and choosing the pixel indices with the least difference.

3. We find the disparity from the difference along the width and we scale the disparity to increase its range between 0-255 and plot the disparity map in grayscale and color using heat map conversion.

Final step is Depth image generation:

1.Using the disparity map from the previous step, we compute the depth information for each pixel using the below formula: depth = (baseline * focal length) / disparity), where baseline is the distance between the cameras.

2. This gives us the depth map in both grayscale and color using the heat map conversion.

Sometimes RANSAC, gave bad fundamental matrix which was countered by fine tuning parameters. There were some issues while plotting the epilines and finding H1 and H2. Achieved clarity by implementing cv2 function and own functions.

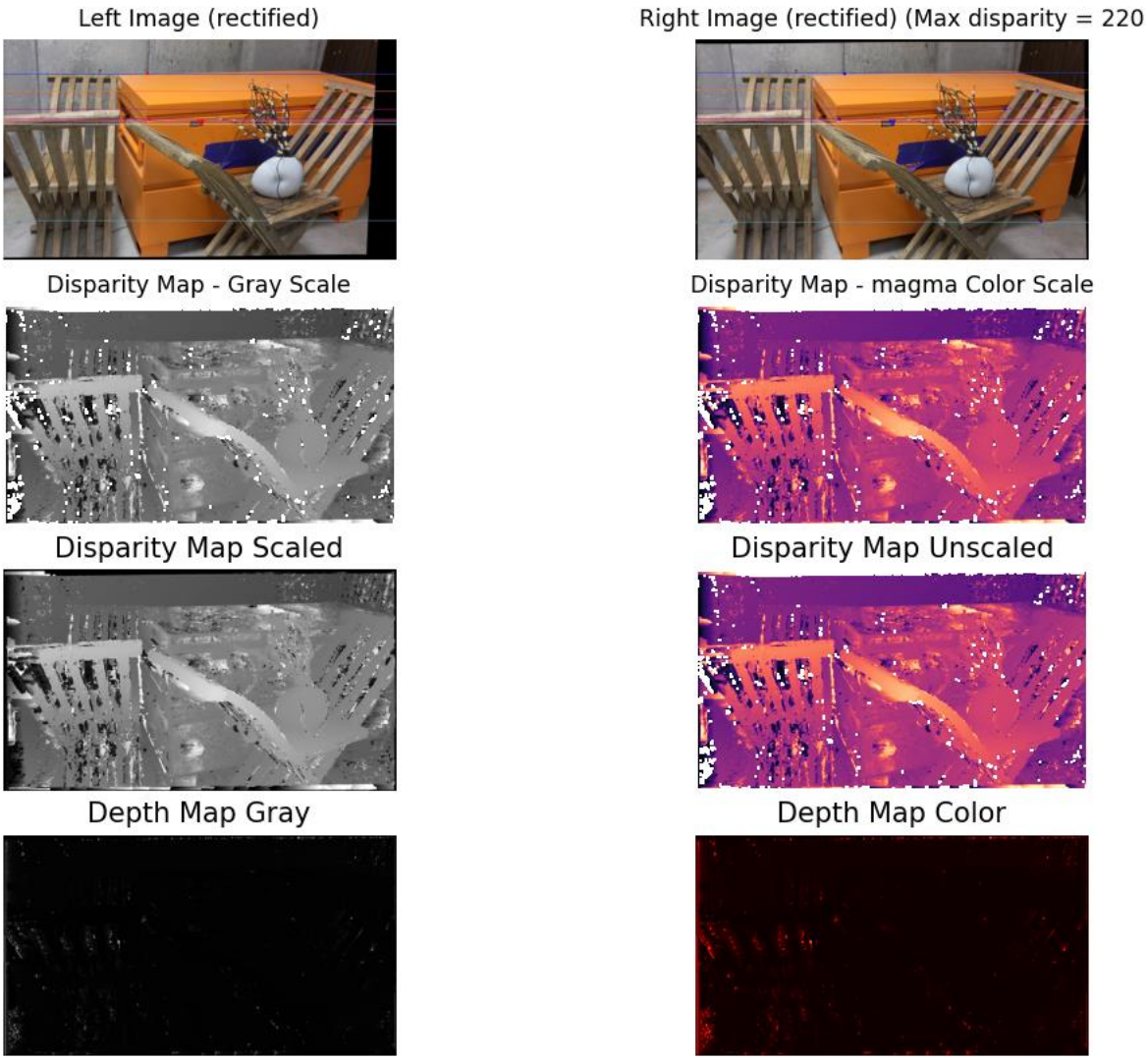

Fig: Required depth and disparity maps

(Depth might look unclear. Refer the results folder for clearer pictures)