Depth Estimation Using Stereo Vision - Project 4

Vinay Krishna Bukka (118176680)

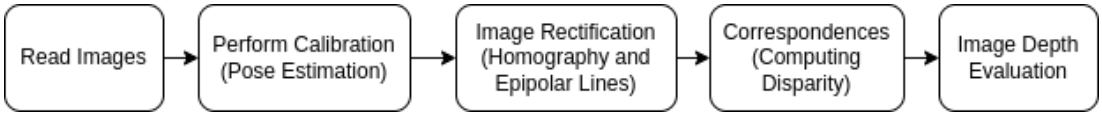
April 2023

Contents

1 Pipeline

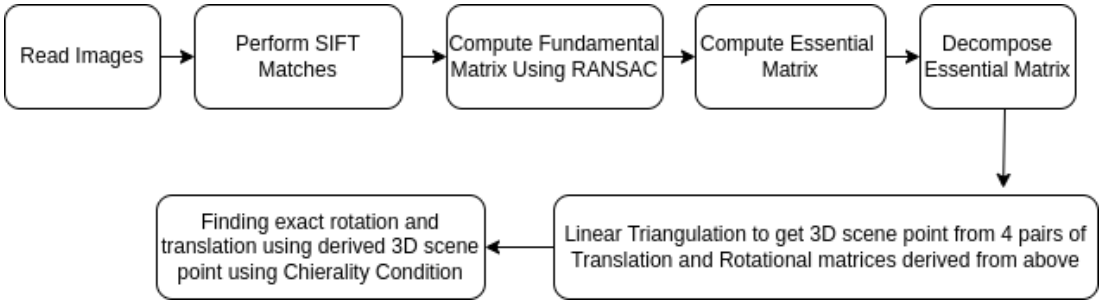
Stereo Vision is the ability of deriving how far the objects are placed from camera. This project aims to compute the depth of the scene using two images taken from different camera angles using Stereo Vision. The same process is repeated for the given 3 different scenes along with given Intrinsic matrix data.

The pipeline followed for this project is below :



2 Calibration

The pipeline followed for calibration is below followed by explanation of detailed approach.



2.1 Approach

- **Approach** The below approach will be similar for all the three scenes given.

  1. The Images are converted first to grayscale images. The corresponding matching points in the two images are found using the SIFT matching detector.
  2. To Compute the Fundamental matrix we follow below approach:
     - The Fundamental Matrix is a mathematical matrix that represents the relationship between two calibrated cameras. It is a 3x3 matrix that relates matching points in two camera images through the epipolar geometry. The fundamental matrix is used to calculate the epipolar lines in one image corresponding to points in the other image.
     - The matching points in both images represent the equation of epipolar constraint along with fundamental matrix as below [ref [2]]:

$$\begin{bmatrix} x_i^1 & y_i^1 & 1 \end{bmatrix} * \begin{bmatrix} f11 & f12 & f13 \\ f21 & f22 & f23 \\ f31 & f32 & f33 \end{bmatrix} * \begin{bmatrix} x_i \\ y_i \\ 1 \end{bmatrix}$$

       where x_i^1 represent point in right image and x_i represent a point in left image.
     - Solving this equation to get the F matrix using homogeneous system of equations solving gives below:

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

- So to solve the above equation of the form $Ax = 0$, we need atleast 8 corresponding matching points from both images as the last element in fundamental matrix is zero.
- To obtain the best 8 points, RANSAC method is implemented. In RANSAC, 8 random points are selected and a fundamental matrix is obtained by above method where SVD is performed on the matrix A. The last column of V from above SVD gives fundamental matrix.
- The fundamental matrix derived above contains noise. So, to eliminate the noise the last Singular value from SVD of above fundamental matrix are made 0. This also ensures the matrix is not full rank and remultiplied to get the Final matrix.
- So, the RANSAC algorithm is run for 500-1000 iterations with a threshold of 0.05 - 0.5. In each iteration, with the derived fundamental matrix from above, error is calculated by the epipolar constraint calculation which is $x_2^T * F * x_1 = 0$ (where x_2^T is the point in image 2 plane and x_1 is point in image 1 plane). Due to noise and rounding errors, make sure this constraint is less than the threshold defined.
- If the error is less than threshold, inlier count is increased showing this is the correct matching point. At the end of iterations, the point set which has maximum number of inliers and corresponding fundamental matrix is determined as best fundamental matrix.

3. Once the Fundamental Matrix is known, essential matrix is found by $(K^T * F * K)$ where K is the given intrinsic camera matrix.
4. The Essential matrix is decomposed into translation and rotational matrices using SVD(Singular Value Decomposition). The U, V from SVD and W matrix leads to formation of 4 pairs of translation and rotation matrices since two camera planes are present. The correct pair of rotation and translation matrices is known by :
   - By using Linear triangulation method, the estimated 3D points of the scene with respect to each image plane are figured out. This is achieved by calculating projection matrices for each image plane with respect to each set of rotation and translation matrices. The matching points in both the images and projection matrices are put together to form a homogeneous system of equations of form $Ax = 0$ (Ref [3]).

$$\begin{pmatrix} v_x p_3^T - p_2^T \\ p_1^T - u_x p_3^T \\ v_y q_3^T - q_2^T \\ q_1^T - u_y q_3^T \end{pmatrix} X = 0$$

- Performing SVD on above matrix A and taking out the vector corresponding to least eigen value gives the estimated 3D pose for the given matching points.
- Once these points are derived, using chierality condition $r_3 * (X - C) > 0$ check is performed for each set of rotation and translation matrices, where r_3 is the last row of rotation matrix, X is the 3d point pose, C is the translation vector.
- The set satisfying above condition is the translation and rotation matrix

- **Problems & Solutions**

1. The major problem is to find the best fundamental matrix using RANSAC. By increasing number of iterations and testing with different threshold values gave the best fundamental matrix.
2. Understanding and performing linear triangulation method is one of the major problem.

- **Results** The results pertaining to different images are below

```
The Fundamental Matrix for artroom dataset is :
[[-7.30504459e-09  1.60038249e-06 -1.05136351e-03]
 [-4.44111035e-07  2.41340747e-06  2.41473491e-02]
 [ 4.11268618e-04 -2.74042145e-02  9.99332097e-01]]
The Essential Matrix for Artroom Dataset is:
[[-2.32553026e-04  1.11322438e-01 -1.69383337e-02]
 [-3.00034650e-02  8.27369957e-02  9.96089873e-01]
 [ 4.11437881e-03 -9.90312340e-01  8.15122091e-02]]
Rotation matrix for artroom dataset is:
[[ 0.99658646 -0.01037975 -0.08190052]
 [ 0.00358577  0.99657062 -0.08266892]
 [ 0.08247773  0.08209305  0.993206  ]]
Translation matrix for artroom dataset is:
[-0.99363995 -0.00770364 -0.11234011]
The Homography Matrices H1 and H2 are:
[[ 2.57206988e-02  3.52940595e-03 -2.64250189e+00]
 [-4.30148654e-04  2.75199914e-02  1.49879110e-02]
 [-4.72525397e-07  2.55459295e-06  2.57045790e-02]]
[[ 9.36677295e-01  7.45903886e-03  5.67619154e+01]
 [-4.35642308e-02  9.99684792e-01  4.19918741e+01]
 [-6.59281240e-05 -5.25005188e-07  1.06357450e+00]]
```

Figure 1: Calibration Results of Artroom Scene

Figure 2: SIFT Matching Points of Artroom

```
The Fundamental Matrix for chess dataset is :
[[ 7.34629874e-09  1.53684228e-06 -3.99655935e-04]
 [-5.66171941e-07  9.90247848e-07  4.31260522e-02]
 [-7.50051564e-05 -4.55608174e-02  9.98030155e-01]]
The Essential Matrix for Chess Dataset is:
[[ 3.18305324e-04  6.14377342e-02  9.31471153e-03]
 [-2.32072685e-02  1.98535177e-02  9.99477921e-01]
 [-9.18283269e-03 -9.97875303e-01  2.02544272e-02]]
Rotation matrix for chess dataset is:
[[ 9.99274474e-01  5.96377216e-04 -3.80810988e-02]
 [-1.35472778e-03  9.99801229e-01 -1.98913958e-02]
 [ 3.80616666e-02  1.99285536e-02  9.99076655e-01]]
Translation matrix for chess dataset is:
[ 0.9980674  -0.01054259  0.06123981]
The Homography Matrices H1 and H2 are:
[[ 4.30189802e-02  1.20158269e-03 -4.88330555e+00]
 [ 7.02965611e-05  4.55771275e-02 -6.38537070e-01]
 [-5.85109995e-07  1.03964989e-06  4.45702683e-02]]
[[ 9.66296658e-01 -9.82075945e-03  3.76584186e+01]
 [-8.76630954e-03  1.00014074e+00  8.33965779e+00]
 [-3.50538541e-05  3.56262713e-07  1.03345932e+00]]
```

Figure 3: Calibration Results of Chess Scene

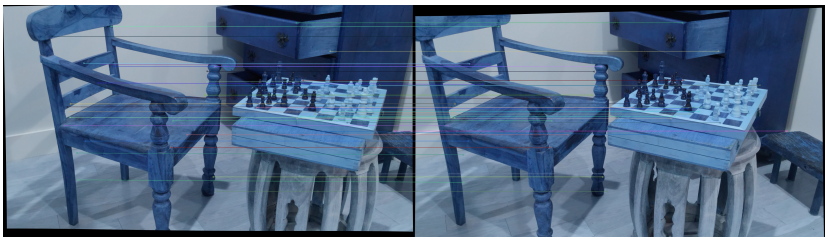


Figure 4: SIFT Matching Points of Chess

```
The Fundamental Matrix for Ladder dataset is :
[[-8.68735718e-08 -5.18538830e-06 -4.89184374e-03]
 [ 6.41323430e-06  1.78664801e-07  1.62548466e-01]
 [ 3.38991010e-03 -1.63579497e-01  9.73028429e-01]]
The Essential Matrix for Ladder Dataset is:
[[-9.12342135e-04 -5.42493202e-02 -5.97085032e-02]
 [ 6.72205542e-02  1.30192415e-03  9.95944012e-01]
 [ 5.74179179e-02 -9.96879642e-01  6.79571054e-04]]
Rotation matrix for Ladder dataset is:
[[ 0.99991131 -0.00239675 -0.01310097]
 [ 0.00237773  0.9999961  -0.0014677 ]
 [ 0.01310443  0.00143642  0.9999131 ]]
Translation matrix for Ladder dataset is:
[ 0.99674022  0.0597932  -0.05416364]
The Homography Matrices H1 and H2 are:
[[-1.67939064e-01 -7.94885435e-03  1.73942237e+01]
 [ 3.28831832e-03 -1.63657660e-01 -1.64099324e+00]
 [-6.29100726e-06 -4.78356524e-07 -1.59607833e-01]]
[[ 1.01497760e+00  6.05573795e-02 -6.62229878e+00]
 [-2.97751932e-02  1.00000180e+00  1.60768754e+01]
 [ 3.10235986e-05  1.85098453e-06  9.81470312e-01]]
```

Figure 5: Calibration Results of Ladder Scene

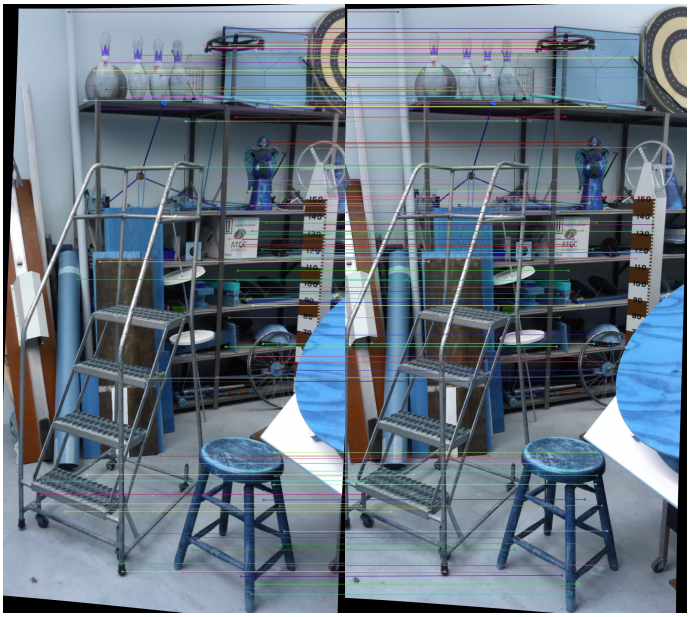


Figure 6: SIFT Matching Points of Ladder

3 Rectification

Rectification is the problem of determining two homographies that map corresponding epipolar lines onto parallel horizontal lines sharing the same ycoordinate.

- **Approach**

  1. The epipolar lines are constructed on the original images for each scene among the three given. The Epilines corresponding to each image of the scene are found by cv2.computeCorrespondEpil Using these lines and matching points, the epilines are drawn in the images. So, the image points corresponding to image 1 lie on the epiline in image 2.
  2. Perspective transformation is applied to get the rectified points in both images to make epilines horizontal which helps in finding out the correspondence.
  3. By taking out homographies from cv2.stereoRectifyUncalibrated and applying warp transformations between homography matrices and left or right image of the scene gives the rectified image. Epilines are drawn on this rectified image.

- **Results** The results showing the epilines before and after rectification are below:

Figure 7: Artroom Epilines Before Rectification

Figure 8: Artroom Epilines After Rectification

Figure 9: Chess Scene Epilines Before Rectification

Figure 10: Chess Scene Epilines After Rectification

Figure 11: Ladder Scene Epilines Before Rectification

Figure 12: Ladder Scene Epilines After Rectification

4 Dense Correspondences

- **Approach**

  1. Using the rectified image points, disparity between the two images is calculated. Disparity refers to the difference in the location of matching points in both left and right. Disparity map represents the displacement between the corresponding pixels of two images.
  2. A window of size 10 is considered surrounding each pixel point of each image. This window is slided across the image 2 means a window of same size is taken for image 2 and compared for a certain horizontal pixels range in both images.
  3. If a match is found, the corresponding pixel values in both images are taken a difference. The sum of squared differences(SSD) is calculated. The area in image 1 corresponding to window for which the SSD is least is used to calculate the disparity for that pixel.
  4. The Disparity map is thus formed by calculating the disparity value corresponding to difference of pixel values in each image. The disparity heat map is obtained by using cv2.applyColorMap on the disparity map obtained.

- **Results** The Disparity Maps and heat maps for the images are shown below

Figure 13: Artroom Dispartity Map

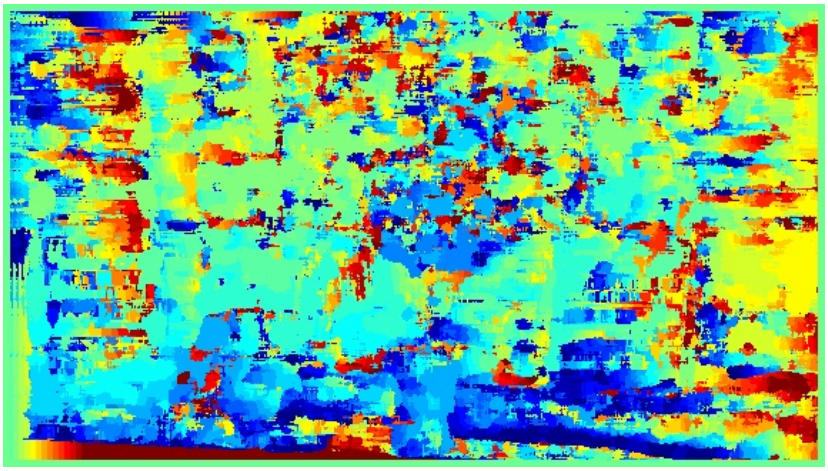


Figure 14: Artroom Disparity Heat Map

5 References

1. Lecture Notes-Perception for Autonomous Robots, Dr. Samer Charifa.
2. `https://cmsc733.github.io/2022/proj/p3/#fundmatrix`
3. `https://3d.bk.tudelft.nl/courses/geo1016/handouts/05_MVS.pdf`

Figure 15: Chess Scene Dispartity Map

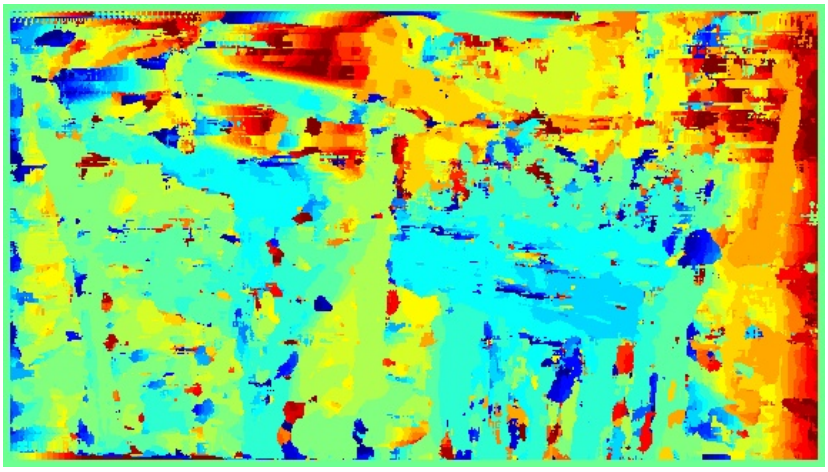


Figure 16: Chess Scene Disparity Heat Map

Figure 17: Ladder Scene Dispartity Map

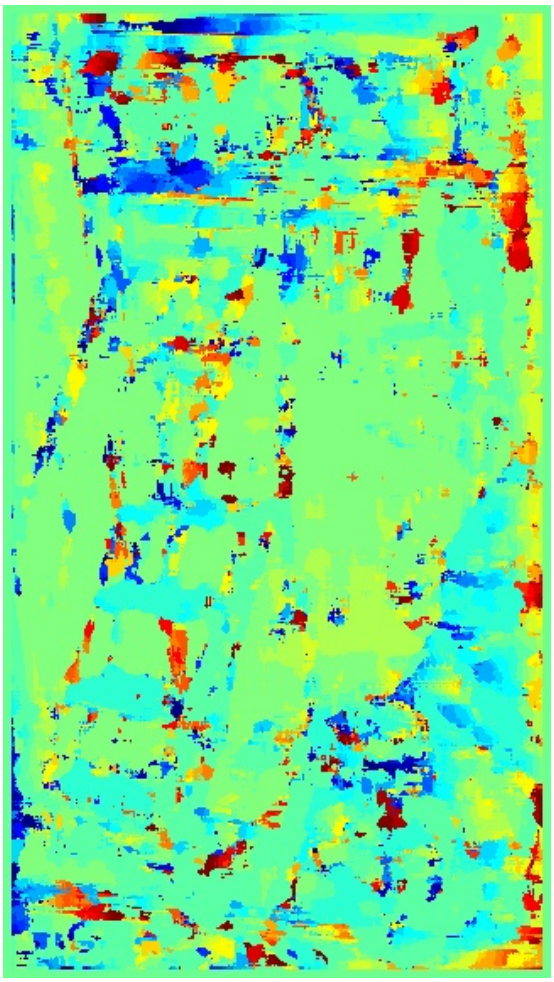


Figure 18: Ladder Scene Disparity Heat Map