Computer Vision Coursework
Feature Extraction and Surface Reconstruction

CID: 01666113
Charlize Yang

November 22, 2021

Question 1

I focused on detecting corners in the frames, as there are many well-defined corners of the square-shaped buildings, swimming pool and football pitch, making them distinctive features to interpret the objects. I chose corners also because they are stable under illumination and geometrical changes and relatively easy to detect computationally. I used the SIFT (Scale Invariant Feature Transform) detector to detect the keypoints in the frames, which is invariant to illumination change, rotation and scale of images. The detected keypoints on both frames are shown in Figure 1, which lie mainly on the corners of the buildings. There are many detected keypoints in the left bottom corner of the frames, lying on some buildings' surfaces, potentially because of the texture change on those surfaces (e.g. the surfaces are made of bricks).

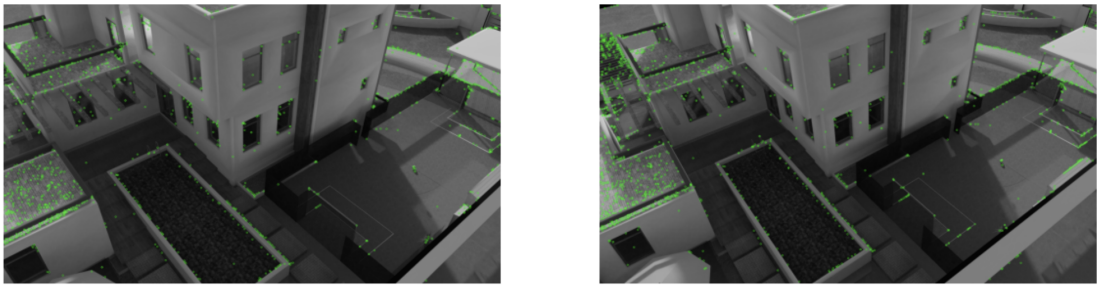

Figure 1: The detected keypoints on the pair of frames using the SIFT detector.

Question 2

I used the SIFT descriptor with the nearest neighbour distance ratio approach for feature matching, due to its robustness and efficiency. The SIFT descriptor detects the keypoints using DoG (Difference-of-Gaussian) and localises them accurately by interpolating nearby data. Then, it characterises each keypoint with a 128-dimensional vector based on orientation histograms. To match features unambiguously, I adopted the nearest neighbour distance ratio approach to only match nearby keypoints if the ratio of distances to the nearest neighbour and the second nearest neighbour is smaller than a threshold (I set it to be 0.7). The matched keypoints are shown in Figure 2.

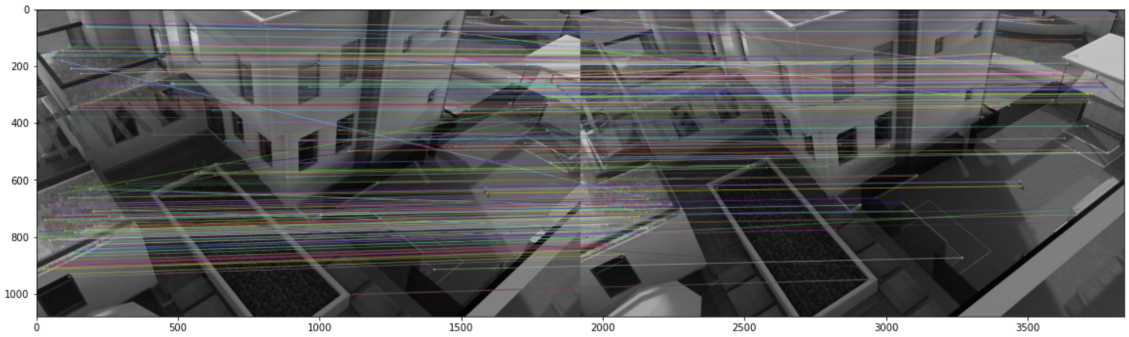

Figure 2: The matched points found using the SIFT descriptor with nearest neighbour distance ratio approach.

Question 3

3(c)

Using the eight-point algorithm, the estimated fundamental matrix based on the matched features is:

$$F = \begin{bmatrix} 1.457\,808\,72 \times 10^{-7} & 1.514\,898\,31 \times 10^{-5} & -7.426\,475\,85 \times 10^{-3} \\ -1.496\,327\,60 \times 10^{-5} & 1.852\,478\,27 \times 10^{-6} & 3.836\,791\,60 \times 10^{-4} \\ 6.776\,903\,95 \times 10^{-3} & -3.237\,378\,39 \times 10^{-3} & 1.000\,000\,00 \end{bmatrix}$$

Using the formula $F = [K't]_\times K'RK^{-1}$, the estimated fundamental matrix based on the extrinsic and intrinsic camera parameters is:

$$F' = \begin{bmatrix} -4.012\,103\,72 \times 10^{-10} & -8.780\,193\,27 \times 10^{-8} & 4.976\,996\,67 \times 10^{-5} \\ 2.511\,864\,47 \times 10^{-8} & 1.365\,831\,64 \times 10^{-9} & 1.314\,461\,24 \times 10^{-3} \\ -1.883\,367\,35 \times 10^{-5} & -1.158\,447\,31 \times 10^{-3} & -5.249\,279\,26 \times 10^{-2} \end{bmatrix}$$

F and F' have distinctive entries with different signs. In general, F' has entries on a smaller scale than F. Such difference implies that not all selected matching points lie on the epipolar lines identified by the camera parameters, i.e., false matchings may be identified. To evaluate the accuracy of methods, I calculated the averaged values of $x'^T Fx$ and $x'^T F'x$ across all matching pairs (x, x') and compared how close they are to zero. The results are shown below ($n = 327$ in our case):

$$\frac{1}{n}\sum_{i=1}^{n} x_i'^T F x_i = -0.013266731$$

$$\frac{1}{n}\sum_{i=1}^{n} x_i'^T F' x_i = 0.004365877$$

The averaged value of $x'^T F'x$ is slightly more closer to zero, suggesting the estimation using camera parameters is more accurate. To improve the accuracy of F, I could reduce the threshold used in the nearest neighbour distance ratio approach or try other descriptors such as SURF to find more accurate matched pairs.

3(d)

The correctly matched pairs of points (x, x') that meet the epipolar constraint $x'^T F'x = 0$ are illustrated in Figure 3, with the corresponding epipolar lines identified for each feature point in the matching pairs in the other image. Take the left frame as an example, for each feature point x in the matching pairs, we calculate the corresponding epipolar line $I' = Fx$ on the right image and check if the matched point x' lies on the epipolar line. If yes, this pair of matching points meet the epipolar constraint.

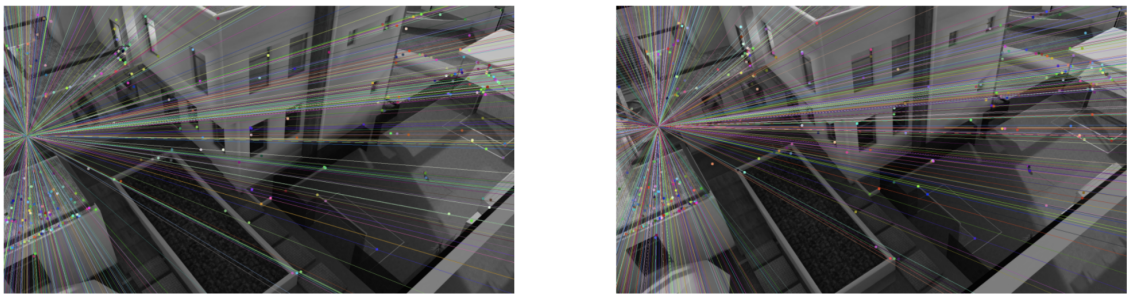

Figure 3: The corresponding epipolar lines to feature points in the matching pairs in the other frame.

3(e)

The estimated area of the swimming pool is $2772m^2$ ($33m \times 84m$), the estimated length of the football match is $96m$. This is estimated based on the disparity map between the frames, as shown in Figure 4. The depth d at each point is calculated using the formula $d = tf/disp$, where t is the distance between two cameras, f is the focal length of the (left) camera, and $disp$ is the disparity at the point. To retrieve a 3D point M = (X, Y, Z) from the depth value, I used the formula $M = dK^{-1}(u, v, 1)^T$, where (u, v) are the image coordinates of the point. By establishing the 3D coordinates of suitable points in the image, we can easily estimate the real-world distances such as the area and the length we need.

Figure 4: The disparity map between two frames.