Imperial College London

COURSEWORK

IMPERIAL COLLEGE LONDON

DEPARTMENT OF COMPUTING

Computer Vision

Author:
Konstantinos Mitsides (CID: 01857560)

Date: November 19, 2023

Question 1

1. The Scale-Invariant Feature Transform (SIFT) has been chosen as it identifies distinctive keypoints within an image with high localisation accuracy. The key strength of SIFT lies in its invariance to scale, rotation, and illumination changes, ensuring robust and repeatable performance. This robustness is crucial in handling variations in viewpoint, relevant to our case due to the different positions of the two cameras capturing the frames. The salient features of choice are corners, as they possess characteristics that enhance their localisability and matchability across the two video frames. In particular, corners are distinctive points, and their local nature allows for the identification of many such points across the entire image. Furthermore, their well-defined positions make them robust to scaling, translation, rotation, and illumination changes. The adaptability of corner detection to natural scenes, coupled with its good mathematical representation, enhances their suitability even further.

Question 2

2. The Nearest Neighbour Distance Ratio (NNDR) has been chosen to establish matches between salient features in the two frames. After detecting keypoints using SIFT, feature descriptors for each frame are extracted again with SIFT. For every descriptor in the first frame, its two nearest neighbours in the second frame based on the Euclidean distance between descriptors are identified. Subsequently, the ratio of the distance between the descriptor in the first frame and its closest neighbour to the distance between the descriptor in the first frame and its second closest neighbour is computed. Then a threshold is chosen, where if the NNDR ratio is larger than the threshold, the match is rejected, and if it is lower is accepted. Using NNDR prevents incorrect matches associated with the ”nearest neighbour” approach.

Question 3

(a)

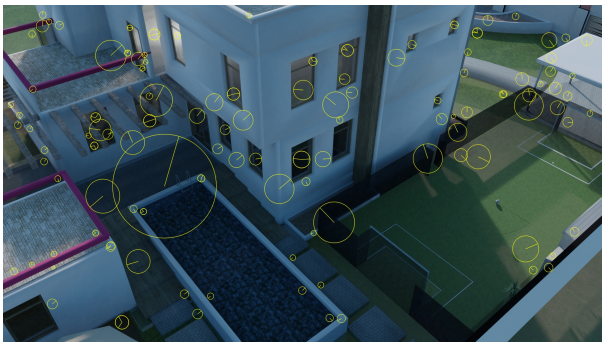

Figure 1: Detected Features on Frame 1

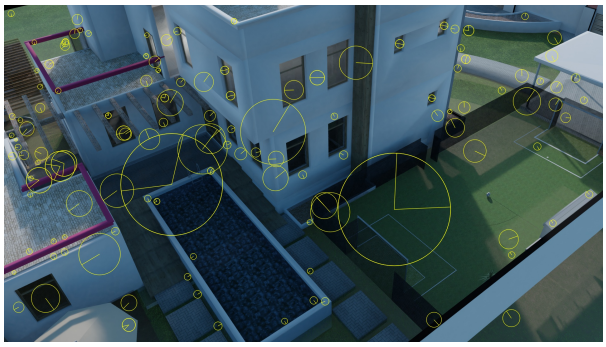

Figure 2: Detected Features on Frame 2

To extract the above features, the images were converted to grayscale, then a smoothing operation was applied, and finally OpenCV functions specifically designed for employing SIFT on the images were utilised.

(b)

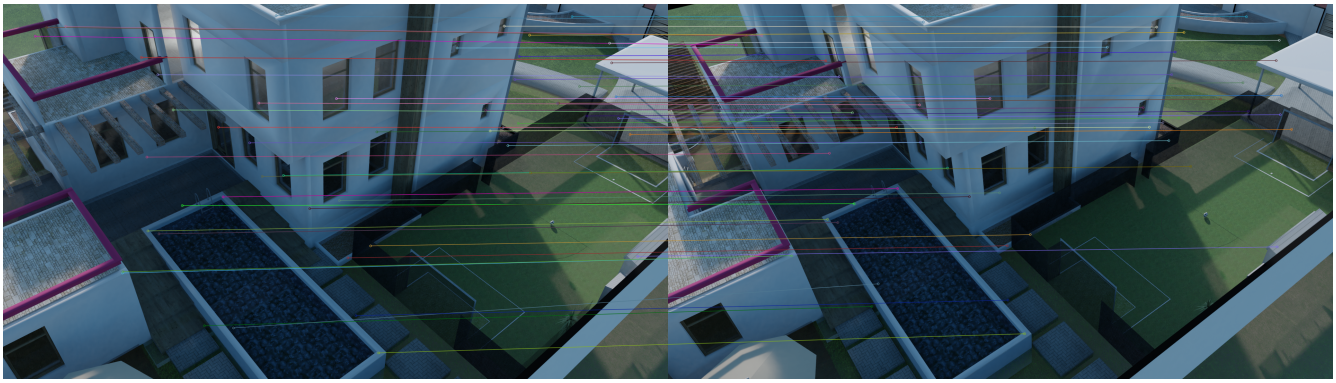

Figure 3: Matches between the two frames

To identify the matches among the previously obtained descriptors, the descriptors from the images, detected using SIFT, were fed into the 2-Nearest Neighbour Algorithm. Then, a Nearest Neighbour Distance Ratio (NNDR) threshold was set to 0.5, where matches with a ratio exceeding 0.5 were rejected to mitigate ambiguity and avoid erroneous associations.

(c)

The fundamental matrix to 3 s.f., estimated using the 8-point algorithm on the matched features (1st method) is:

$$\begin{bmatrix} -3.29\times10^{-7} & 4.39\times10^{-5} & -9.28\times10^{-3} \\ -4.20\times10^{-5} & 1.75\times10^{-6} & -1.85\times10^{-3} \\ 8.99\times10^{-3} & -3.24\times10^{-3} & 1.00 \end{bmatrix}$$

The fundamental matrix estimated using the extrinsic and intrinsic camera parameters (2nd method) is:

$$\begin{bmatrix} 7.64\times10^{-9} & 1.67\times10^{-6} & -9.48\times10^{-4} \\ -4.79\times10^{-7} & -2.60\times10^{-8} & -2.50\times10^{-2} \\ 3.59\times10^{-4} & 2.21\times10^{-2} & 1.00 \end{bmatrix}$$

The first method relies solely on the already identified matched features, while the second method incorporates the actual camera parameters. Consequently, the estimation accuracy of the first method is sensitive to noise and outliers in the correspondence data, potentially stemming from repeated patterns or deformations, making it highly reliant on the precision of the initial feature matching. The second method demonstrates greater accuracy due to the substantially lower likelihood of inaccurate calibration compared to the likelihood of encountering outliers in the matches. Its higher accuracy was validated by applying both matrices to the condition formula for the epipolar constraint,

$$|\mathbf{x'}^{\mathrm{T}}\mathbf{F}\mathbf{x}| = 0$$

resulting in 18 more pairs satisfying the condition when the second matrix is used. To improve the accuracy of the first method we can use a lower NNDR threshold or even use the RANSAC algorithm that handles outliers and improve the accuracy of feature matching.

(d)

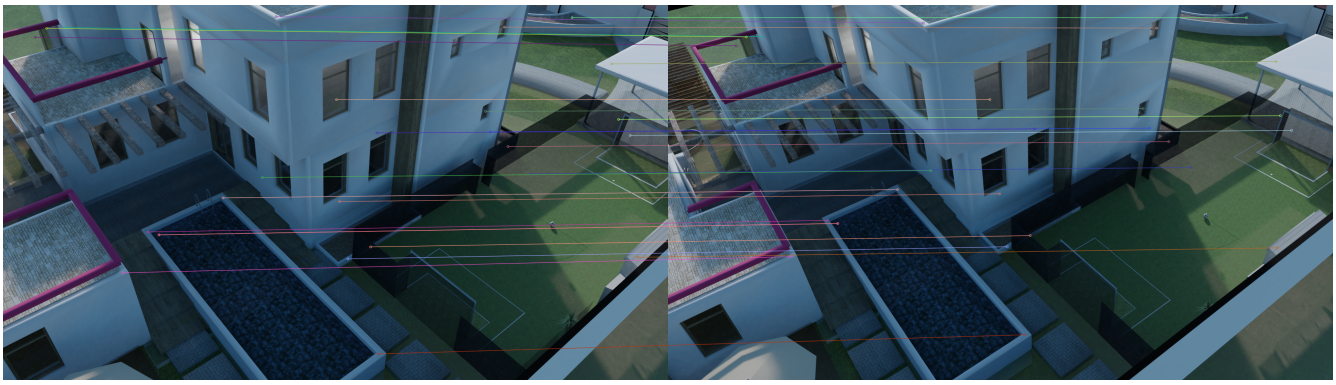

Figure 4: Corrected matches between the two frames

If the following condition is satisfied,

$$\mathbf{x'}^{\mathrm{T}}\mathbf{F}\mathbf{x} = 0 \qquad (*)$$

where **F** is the fundamental matrix
and **x**, **x'** are the matched points in frame 1 and 2 respectively

then the matched points **x** and **x'** meet the epipolar constraint, and thus they are correctly matched.

To identify correctly matched points, all previously matched points were applied to the left-hand side (LHS) of the equation $(*)$ using the fundamental matrix **F** which was derived using the camera parameters. If the absolute result exceeded 0.06, the pair of points were regarded as incorrectly matched and then were removed from consideration.

(e)

Swimming pool area: 33.42 m^2 (2dp)

Length of the field: 15.02 m (2dp)

To determine the area of the pool, rectification has been applied to the images, extracting essential parameters like the focal length, baseline, and the coordinates x_0 and y_0 (consistent with the notation in the slides). Pixel points corresponding to the top-left, top-right, and bottom-right corners of the pool in both images were then manually identified, and disparity values were calculated for each pair of these points. Using the obtained disparity values, focal length and baseline, the depth for each identified point was determined, i.e the Z-coordinate. Then, using the relevant values and equations presented in the slides, the pixel points from the left image were converted into 3D coordinates. The area of the pool was then calculated using the cross product equation with the associated vectors. A similar procedure was employed to calculate the length of the field, where instead two pixel points that define the length of the field, located near the building, were manually identified.

Question 4

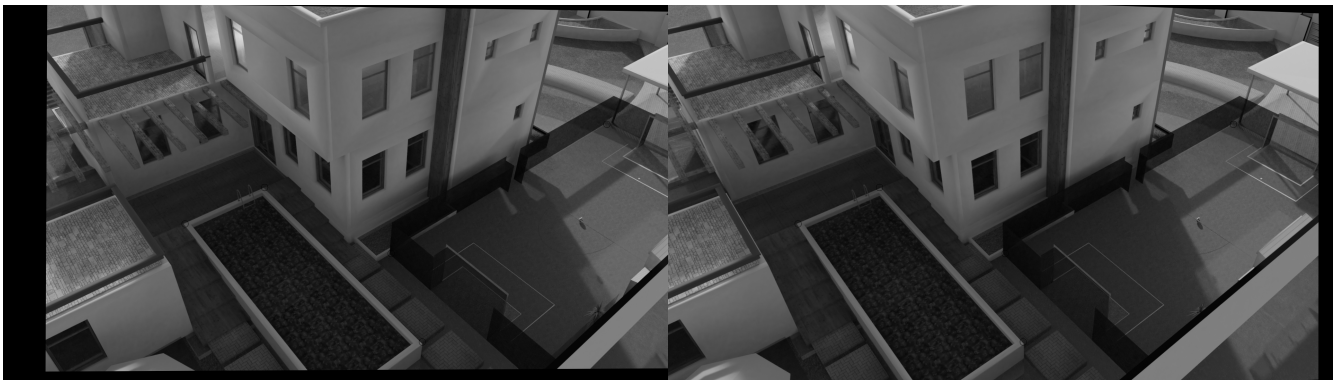

Figure 5: Rectification result of the video frames