1. **Propose a technique to detect salient features of your choice on the video frames above. Explain the type of features on which you will focus and justify your choice. (Word limit: 150 words)**

We can use SIFT to detect the salient features. SIFT find extrema by applying Gaussian filtering across multiple scales that are divided into two octaves. We convolve the same Gaussians in both octaves but image size is halved in upper octave. The Difference of Gaussians (DoG) is calculated and compared across spatial and scale domain to identify local extrema. I chose SIFT as DoG is efficient to compute (comparing to LoG).

The type of features I will focus on are the corner points. Examples include the corners of the building, swimming pool, football field, yard lines, football doors as well as the football itself. I chose these features primarily because they are distinctive and not repeated in the image. Also, these points have good locality as they are all small areas. They are invariant to scale, orientation changes and deformation. They can also be identified on both Frame1 and Frame2.

2. **Propose a technique to match the detected salient features between the video frames. Explain how you would approach this task and the steps you would follow. (Word limit: 150 words)**

First, localise the salient features by eliminating those doesn't have enough contrast and those along edges. Then extract the 1*128 SIFT vector descriptor for these refined interest points. We can then find match between corresponding features on the two images. For each descriptor on Frame1, compute its Euclidean distances with all descriptors in Frame2 and apply the Nearest Neighbour Distance Ratio (NNDR). In particular, for each key points on Frame1, find the nearest neighbour distance (d1) and the second-nearest neighbour distance (d2). If d1/d2 is close to 1, the match is ambiguous and should be rejected. If d1/d2 is close to 0, it means the descriptor is very close to one neighbour and relatively far away from the rest of the neighbours, this indicates a distinctive match and we would keep this feature match. This technique improves accuracy by reducing false positives while retaining strong matches.

3. Use a programming environment of your choice to:

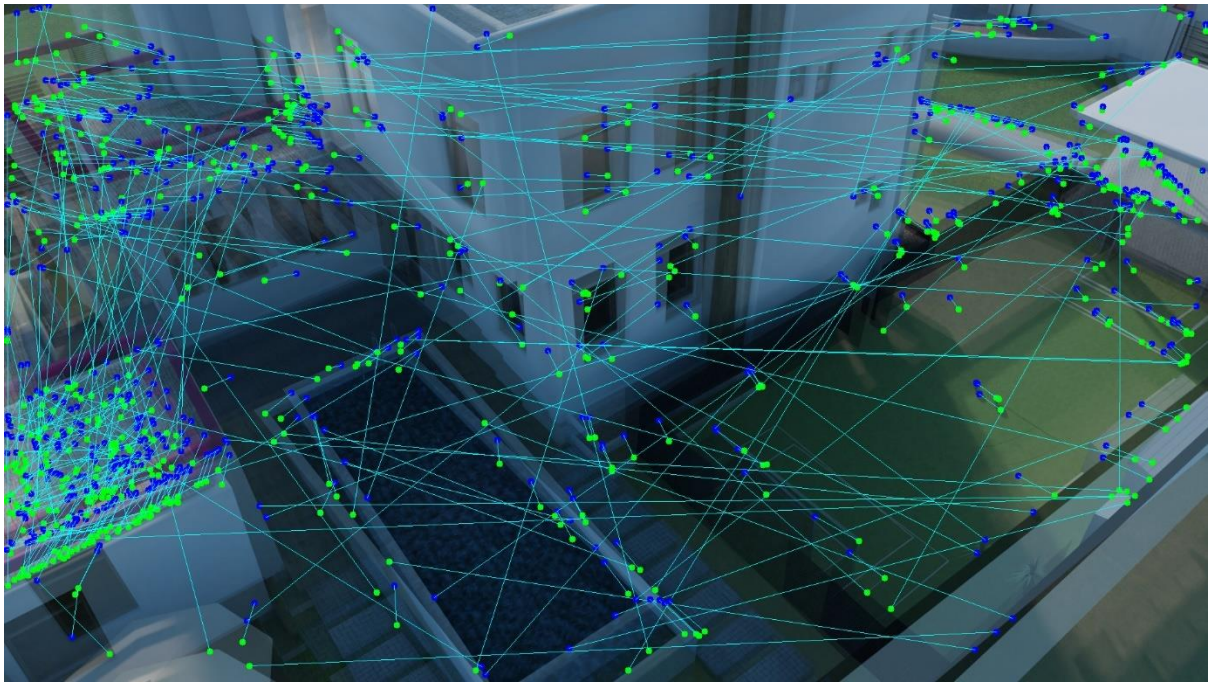
a. Implement your proposed salient feature detector and plot the detected features on the provided pair of frames.

(answers in next page)

Frame 1

Frame 2

b. Find corresponding features between the two frames and illustrate those matches. To illustrate the matches you can, for example, create a composite image (e.g centered overlay image) from the two frames.



c. Use the matched features to estimate the fundamental matrix between the two images. Now estimate the fundamental matrix using the extrinsic and intrinsic camera parameters. Compare the estimated fundamental matrices and explain any possible disagreement between the two methods. Which method is more accurate? Justify your answer and suggest how you could improve the least accurate method. (Word limit: 150 words)

Fundamental Matrix from Matched Points:

[[-1.84052511e-07 -2.94164190e-06 5.20242546e-03]

[5.12894635e-06 -4.64694042e-07 -4.67723602e-02]

[-5.59085270e-03 4.24675270e-02 1.00000000e+00]]

Fundamental Matrix from Camera Parameters:

[[-4.01210372e-10 -8.78019327e-08 4.97699667e-05]

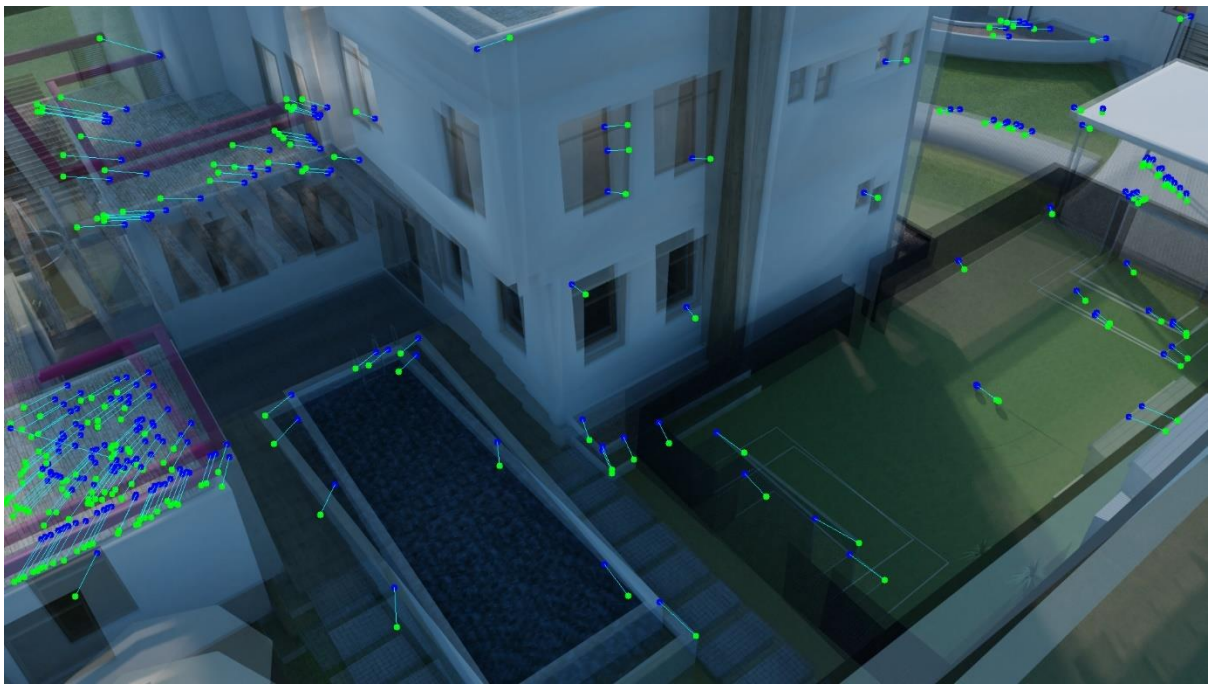
[2.51186447e-08 1.36583164e-09 1.31446124e-03]

[-1.88336735e-05 -1.15844731e-03 -5.24927926e-02]]

The two estimates differ significantly. The Fundamental matrix from matched points could be prone to errors due to incorrect matches identified. There could be absence of reliable features or structures with a lot of repeated structure points could cause difficulty in robust matching. For this, we need to identify matches that satisfies uniqueness, smoothness and ordering constraints. Calculating fundamental matrix from camera parameters is more accurate, because the error in camera calibration is likely to be very small. To improve the estimate from matched points, we can first select 8 matches, calculate the initial F. Then re-project the feature points using the estimated F,

remove the points that are re-projected with large errors (outliers) and look for additional matches. We repeat the process until the number of outliers are below a pre-set threshold %. This ensures the F estimate is based on all correct matches.

d. Find the correctly matched points that meet the epipolar constraint and illustrate these matches. Briefly explain how these matches have been identified. (Word limit: 150 words)



The correctly matched points are identified by enforcing the epipolar constraint: find the matched pairs x, x' where $x'^T Fx = 0$ is satisfied. The logic for identifying these matches is as follows:

From previous questions, the total number of matches detected are 608. For these pairs, I checked if they satisfy the equation above. In this, I've relaxed the threshold to 1e-3 instead of 0 to get more meaningful matches. I've also used the Fundamental Matrix from camera parameter as it is more accurate. As a result, I got, among the 608 matches at the beginning, 238 matches satisfy epipolar constraints, which are drawn on the picture above.

e. Estimate the area of the swimming pool and the length (touchline) of the football field. (hint: you can establish the disparity map between these frames or you can apply 3D surface reconstruction)

I would first apply rectification so that both cameras are facing the same direction. I will then calculate the disparity by establishing the disparity map. Then, calculate the depth for the swimming pool corners and football field corners. These are the Z variable value for these points. I will then calculate X and Y values by formulas X = (x-x0) t/disp and Y = (y-y0) t/disp. The X&Y values will be calculated for the corner points of the swimming pool and football. Then, the width and height of the swimming pool can be calculated as the Euclidean distance between the coordinates, so is the area. The length of the football field can also be established.