COURSEWORK 2

IMPERIAL COLLEGE LONDON

DEPARTMENT OF COMPUTING

Computer Vision

Author:
Arnold Cheung (CID: 01184493)

Date: November 27, 2019

Question 1:

The technique to detect salient features of choice is the SIFT detector. The SIFT detector will focus on extracting Blob-like features, where the colour or intensity of the pixels in the features are relatively similar. Using SIFT, the scale space of the image is first computed, by applying Gaussian filters (blurring) of increasing scale to the image, to find the key-points in the image, the difference of Gaussian (DoG) is computed by subtracting two images with adjacent scales. The key-points are then the local extrema when a pixel is compared with its 26 neighbours from the current, above and below DoG image. The main advantage of SIFT detection over other techniques such as Harris corner detection is that SIFT is both scale and orientation invariant, which is important in a video setting where the scale of objects are constantly changing depending on the camera position.

Question 2:

A descriptor must first be computed for each of the key-points, the technique of choice is the SIFT descriptor. The key-points of each frame are first located using the SIFT detector technique. The edge orientations of each pixel in the 16x16 window around the interest points are then computed, these are then grouped into 4x4 cells, the orientations votes in each cell will be counted and a histogram of orientations will be created for each cell, the 16 orientation histograms will together form the descriptor of the key-point. With the descriptors computed, key-points across the video sequence can then be compared, one simple approach is by using the brute force matcher, where every key-point in one frame will be compared with every key point in a subsequent frame, the most similar key-points, determined with the smallest euclidean distance between the descriptors, will be matched

Question 3:

a) Using the above described SIFT detector, the key-points are determined for two consecutive frames of the video sequence, the top twenty most similar matches are drawn as shown below. The size of the circle around the key-point represents the size of the key-point, and the line represents its generation orientation:

Figure 1: Frame 1

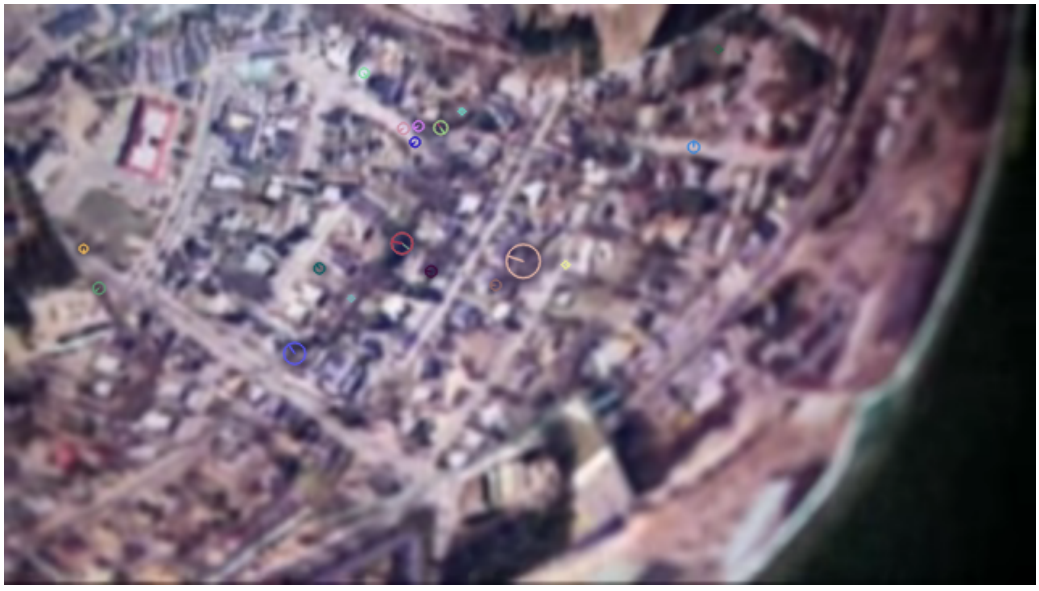

Figure 2: Frame 2

b) Applying feature matching gives the following result:

Figure 3: Lines are drawn between matched key-points in the two frames

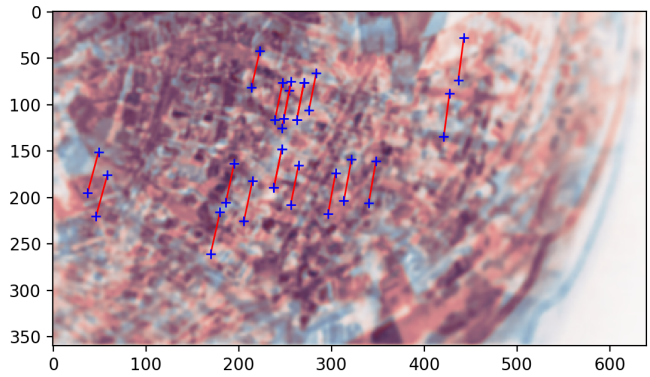

Figure 4: 2 frames are overlaid on top of each other, frame 1 recoloured in blue and frame 2 recoloured in red. The translation of the key-points are tracked and drawn

The mean translation vector calculated from the top twenty matches of key-points is $[\,8.55 \quad -42.5\,]$. This translation is applied to frame 1 and the frames are overlaid again:

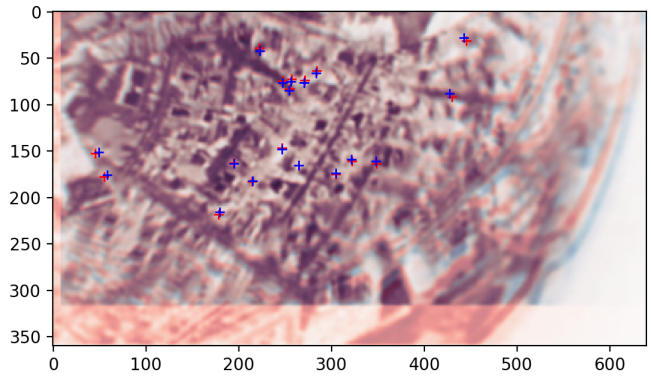

Figure 5: After frame 1 has been translated, the key-points in the two frames can be seen to be approximately aligned

c) Figure 6 shows the computed disparity map between the two frames, it can be seen that the depth of the image is evenly distributed, and can therefore be deduced that the scene is actually flat, the apparent curvature shown in the images is likely caused by a camera lens effect.

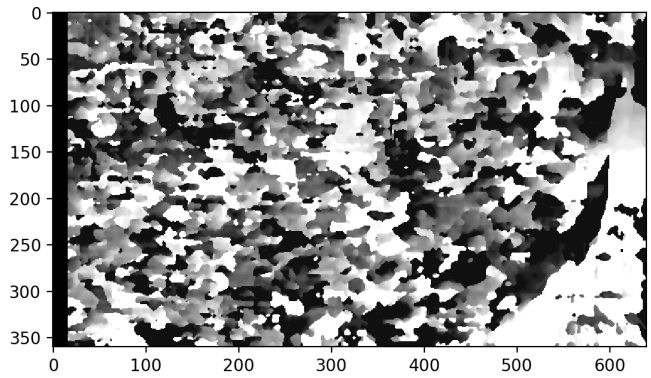

Figure 6: The disparity map computed using the two frames, white areas represent objects that are closer

d) The fundamental matrix is estimated using the matched points, points that satisfy the epipolar constraints are plot on the diagram below:

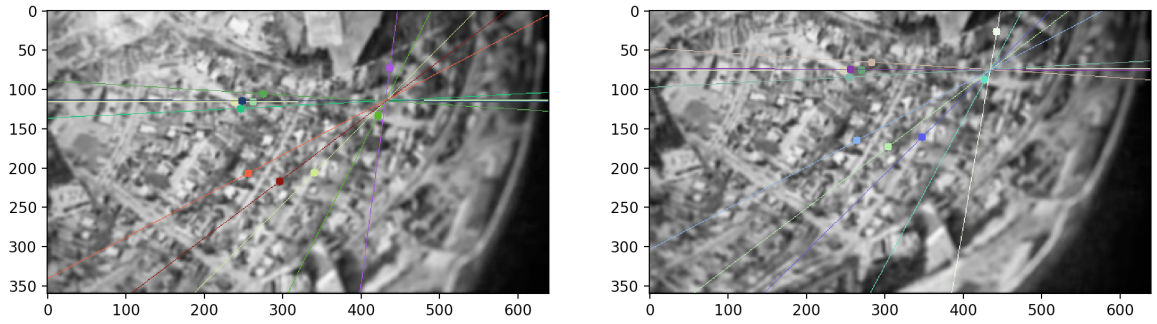

Figure 7: Epipolar-lines and the corresponding key-points that satisfy the epipolar constraints