Short Answer 1

If we were to take any positions whose filter response exceeds a threshold, the repeatability of the resulting interest points would suffer since the different scales of different images would mean that some interest points might not be recognized. However, it would help create greater feature intensity, making it easier to determine the distinctiveness of the interest points. Taking positions that are local maxima would have the opposite effect, since there would be a lot more noise, making it difficult to distinguish between some interest points though it would make it easier to identify matching interest points.

Short Answer 2

The inliers are the interest points that lie close to the epipolar lines. These can be found by randomly selecting epipolar planes until a certain number of inliers are found.

Short Answer 3

One possible failure mode is a textureless surface since it's hard to find corresponding points on surfaces that look the same throughout. Another possible failure mode is a non-Lambertian surface which due to images being refracted/reflected differently based on perspective, it can be difficult to find corresponding points.

Short Answer 4

SIFT provides a set of features of an object that aren't affected by the usual complications of scaling or rotation

Short Answer 5

The Hough parameter space is 4D because each matched feature casts a vote on location (which is 2D), scale, and orientation.

2.1

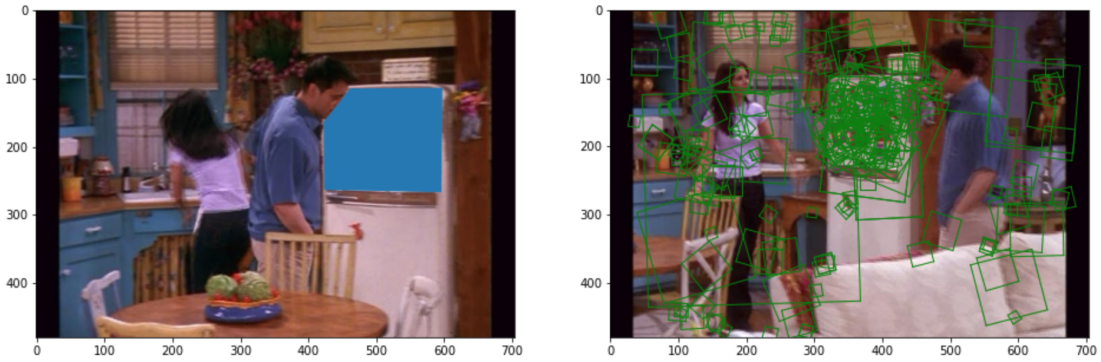

Figure 1: Selected Region and Matched Features

There is a cluster of matched features around the freezer door in the second image, which was the selected region in the first image.

2.2

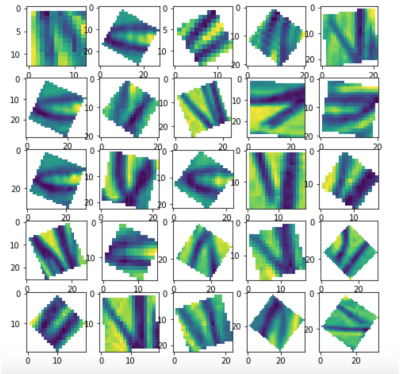

Figure 2: Examples of first visual word

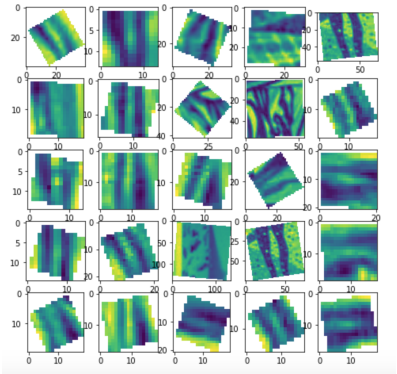

Figure 3: Examples of second visual word

Based on the example patches being displayed, the first visual word appears to be a sort of dark v shape while the second visual word looks more like two dark parallel lines. It seems that visual words are relatively simple visually, which makes it easier for computers to compare much more complicated visual "sentences" or "paragraphs."

2.3

Figure 4: 3 full frame queries and their 5 most similar frames

The first image has the most matching frames probably due to the frame or very similar frames being used to build the vocabulary. The second image was able to get one closely matching frame perhaps due to it not being as present in the frames used to build the vocabulary. The final frame has no closely matching frames maybe due to being even less present in terms of number of similar frames within the set of images used to build the vocabulary.

2.4

Figure 5: 4 region queries and their 5 most similar frames

I noticed that the biggest failure case out of the four images was when I marked a polygon surrounding only Joey's face in the third image. The closely matching frames seemed to show a bunch of characters who seemed to have similar expressions as Joey but none of the images included Joey, oddly enough. Perhaps the vocabulary wasn't detailed enough to include faces. Out of the biggest success cases (the second and the fourth), I selected much bigger regions, usually including the two people being shown in the frame.