CS 4476: Computer Vision, Fall 2020

PS4

Name: Fei Ding

Due: Monday, Oct 23rd 2020

Problem 1

(1) When performing interest point detection with the Laplacian of Gaussian, how would the results differ if we were to (a) take any positions that are local maxima in scale-space, or (b) take any positions whose filter response exceeds a threshold? Specifically, what is the impact on repeatability or distinctiveness of the resulting interest points?

Student Response: In the scale-space we need to find interest points of varying sizes. (a) When we apply the Laplacian of Gaussian to the image and take the local maxima in as detected points, these detected points can be different ones across all runs due to different values used for the scale of the interest points. Therefore, the interest points we obtain here are more repeatable but less distinctive. (b) When we set a threshold for the filter response, the result is extremely sensitive to the threshold value: a high threshold can lead to no points detected across all runs. On the other hand, a proper threshold value is expected to capture the same interest points across most of the runs, so the result is more distinctive but less repeatable.

(2) What is an "inlier" when using RANSAC to solve for the epipolar lines for stereo with uncalibrated views, and how do we compute those inliers?

Student Response: For uncalibrated camera views, not all epipolar lines are coincident due to uncorrected F. When using RANSAC for estimating the epipolar lines, we randomly select 8 points (supposing taht 8-point algorithm is used) and solve for the missing parameters in the fundamental matrix F. Then, we use this estimated F to calculate the expected transformation, compute their distances to the original correspondences, and compare those to a certain tolerance. An "inlier" is a correspondence that falls within this tolerated error; otherwise, the correspondence is considered an "outlier". We will repeat the selection process until we have sufficient "inliers" indicating that our parameters are very close to the ground truth. Finally, we will use only these many "inliers" to compute a final estimation for F, which yields most accurate epipolar lines.

(3) Name and briefly explain two possible failure modes for dense stereo matching, where points are matched using local appearance and correlation search within a window.

Student Response: For dense stereo matching, the algorithm can fail on surfaces with little texture. In this case, it is difficult to establish correspondence from one image to the other because there are many windows with similar appearance in the other image. Similarly, a texture with repeated patterns at the same angle will also make the local patches resemble each other and create ambiguity to the algorithm. The algorithm can also fail due to occlusion in image formation. Since the camera is translated to obtain the second image, the window patch in the first image might be blocked by some object in the second image, in which case it is impossible to identify a correspondence.

(4) What exactly does the value recorded in a single dimension of a SIFT keypoint descriptor signify?

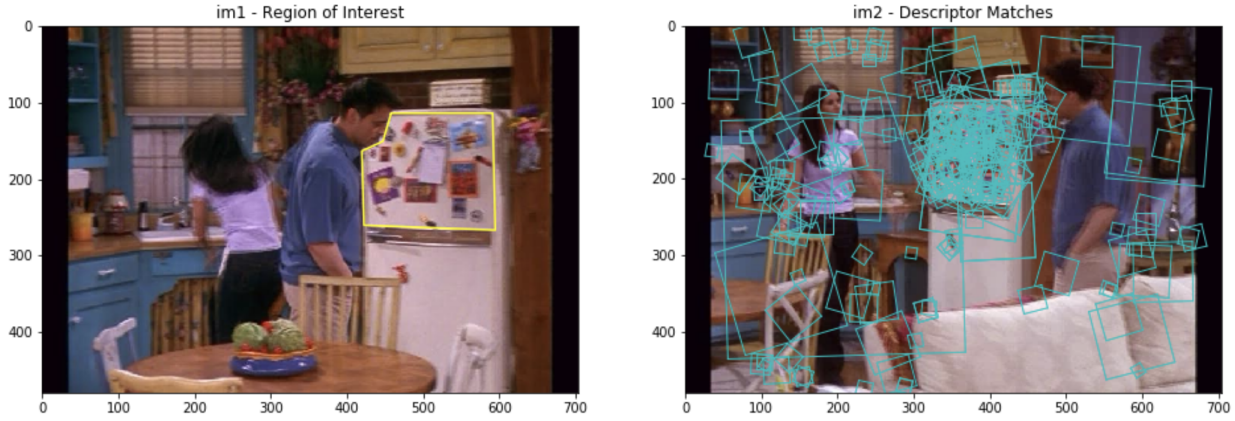
Student Response: A single dimension of a SIFT descriptor can be the response of an elementary filter. For example, as opposed to complex concepts such as visual words, it can be as simple as the derivative in the x or y direction or some combination in a particular direction. We aggregate the response values of these rudimentary filters to form a high-dimensional vector and call that a SIFT descriptor with the hope to capture more complex features.

(5) If using SIFT with the Generalized Hough Transform to perform recognition of an object instance, what is the dimensionality of the Hough parameter space? Explain your answer.

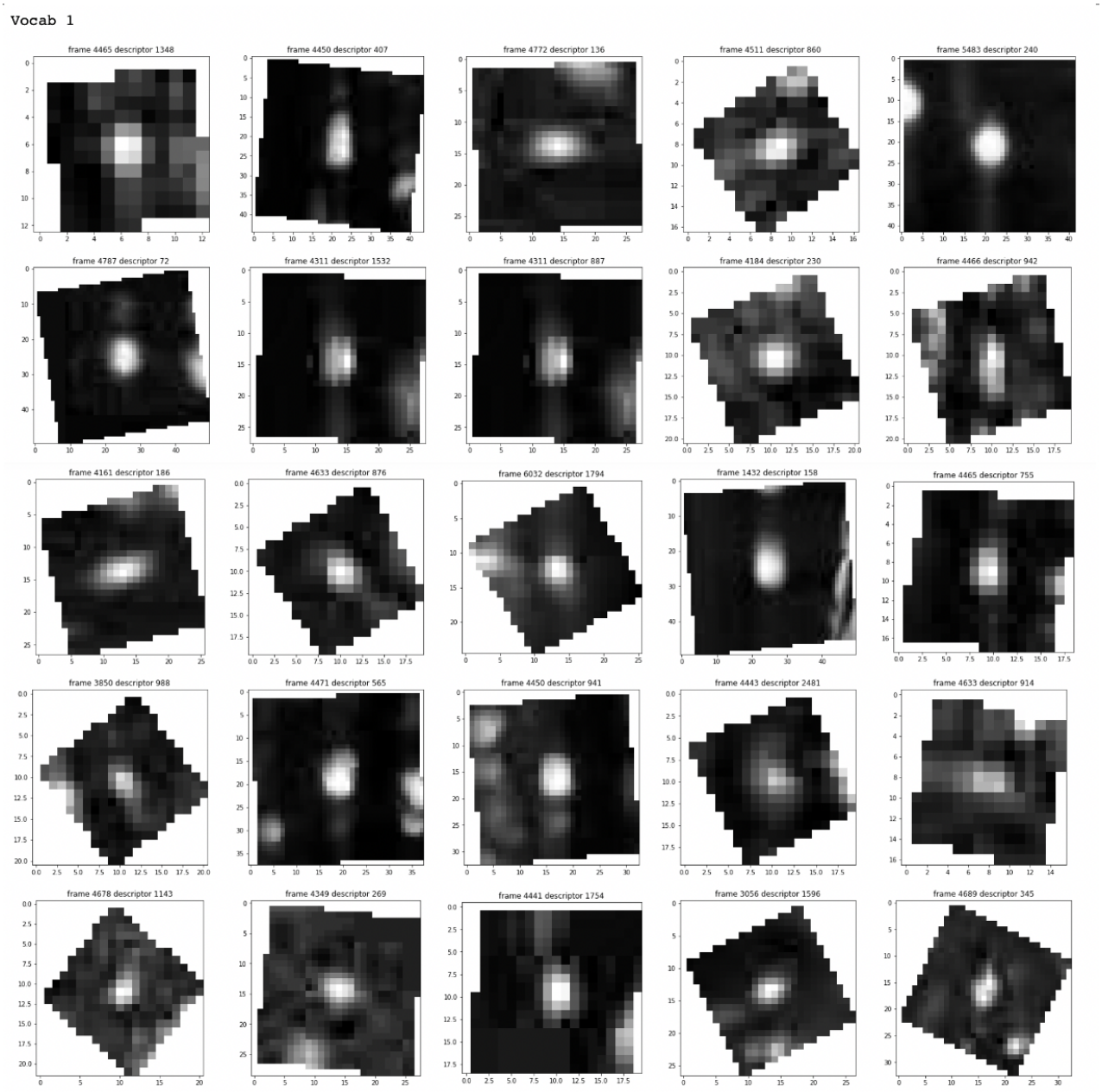
Student Response: The Hough parameter space for voting will be 4-dimensional. The four dimensions will be scale, x-translation, y-translation, and rotation/orientation. Each SIFT descriptor match will vote for a hypothesis in the Hough space, and at the end we verify the parameters with enough votes.

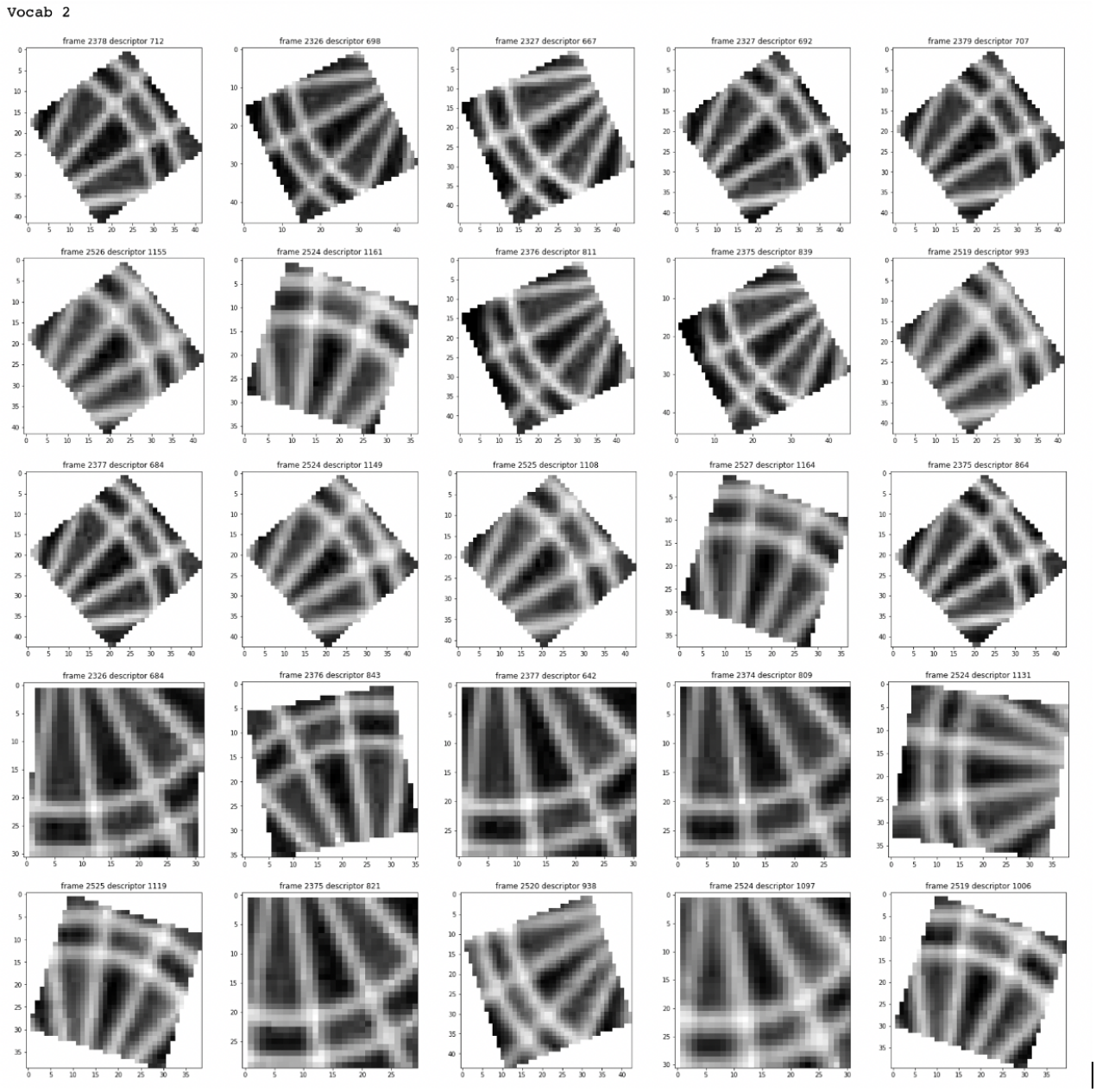
Problem 2

(1) Add the figure from step 3 to your answer sheet.



(2) Display patches corresponding with two visual words in your answer sheet.





Discuss the results in your answer sheet.

Student Response: The two visual words I picked above (vocab #12 and #10) illustrate the difference in the features they capture. The first visual word tends to recognize image patches with a dark background and a bright spot in the middle, whereas the second visual word captures a particular circular lattice pattern, which might be from a specific chair back in the episode.

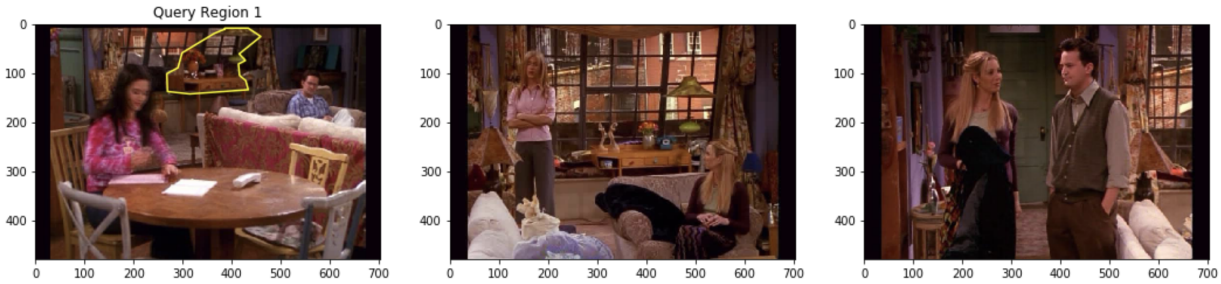
(3) Display 3 different query frames and the M=5 most similar frames from the video dataset (don't include the query in the results) in your answer sheet.

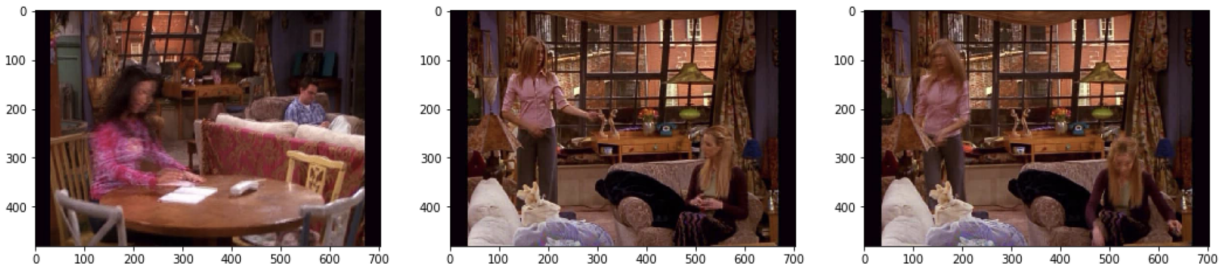
Explain the results in your answer sheet.

Student Response: The three query frames I used was frame 60 (the restaurant), 83 (the couch), and 423 (the living room). For each frame, I searched over all available frames (6000+) in the database and find the top 5 frames that have the greatest scalar products of histograms with that query frame. The histogram is calculated by classifying the SIFT descriptors in each frame into a bag of words.

The five other frames matching the query frame are distinct from the query frame by an extra condition in the program. We can see the identified frames are all associated with the original frame but look a bit different from each other for all three query frames.

(4) Display the selected query region and the M=5 most similar frames for 4 different queries in your answer sheet.






Query Region 3
Query Region 4

Explain the results in your answer sheet.
Student Response: To implement this section, I used the `select_roi.py` script to circle out the region of interest enclosed a polygon, and only use the SIFT descriptors falling within the construct the bag-of-words histogram. Afterward, I searched over all frames and calculated the dot product for each frame to find the top 5 matches.

Query Region 1 illustrates one success scenario. The interest region is a stylistic vase and lamp. The most similar 5 frames found mostly include theses objects, and all of them are from different angles.

Query Region 2 illustrates one success scenario. The interest region is a coffee maker and a juice maker. The most similar 5 frames found mostly include theses objects. In particular, the fourth frame found is drastically different from the others: the target objects lie under the shelf in this frame.

Query Region 3 illustrates one success scenario. The interest region is a blue lamp on a dinning table. The query image is shot in a closer distance (the table is not clearly visible), but the five other identified scenes are shot from a farther distance (the table is clearly included).

Query Region 4 illustrate a **failure** scenario. The interest region is a flat vase decoration on the dinning table, but none of the five other scenes include it in its content. One possible explanation is that the vocabulary list trained over randomly selected 500 images over all 6000, and we may accidentally not included this particular frame or any frames similar, so none of the typical SIFT descriptors (words) can accurately capture the features of this object. In this case, the algorithm does no better than randomly guessing in face of tremendous noise in the data images. I further inspected the list of the actual 500 images used for training to confirm my conjecture.