Ko Outlaw-Spruell

Problem Set 3

I. SHORT ANSWER PROBLEMS

1. What exactly does the value recorded in a single dimension of a SIFT keypoint descriptor signify?
   A region that the SIFT descriptor wants to describe is divided into 16 sub-regions. This is to encode some spacial information as well. Then, eight gradient directions are calculated for each of those sub-patches. So, a single dimension of a SIFT keypoint descriptor describes a single gradient direction of one of the sub-regions.

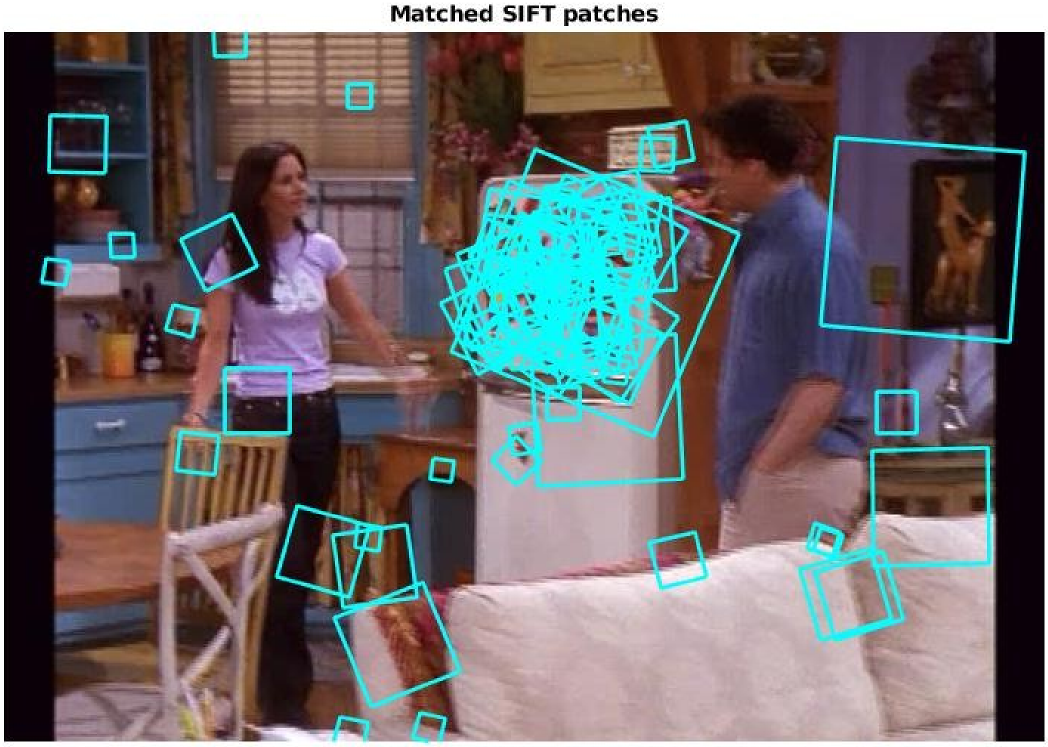
2. A deep neural network has multiple layers with non-linear activation functions (e.g., ReLU) in between each layer, which allows it to learn a complex non-linear function. Suppose instead we had a deep neural network without any non-linear activation functions. Concisely describe what effect this would have on the network. (Hint: can it still be considered a deep network?)
   The back-propagation of the error would be greatly effected, because extreme values, such as large negative values will go through the neural network. This causes large fluctuations in the calculations of the error, and the weights will be volatile. This doesn't help the network 'learn', so we need the non-linear activation functions.

II. PROGRAMMING

1. Raw descriptor matching

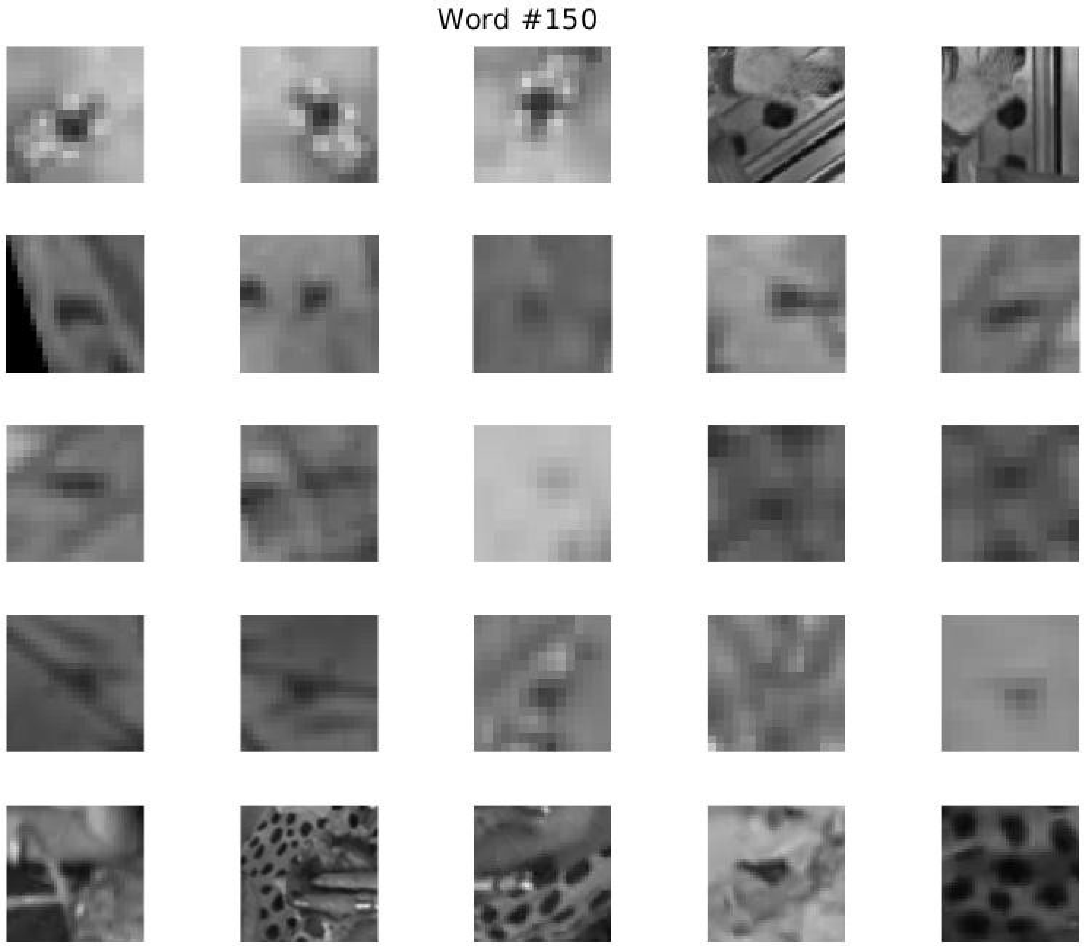
Selected Region

Matched SIFT patches



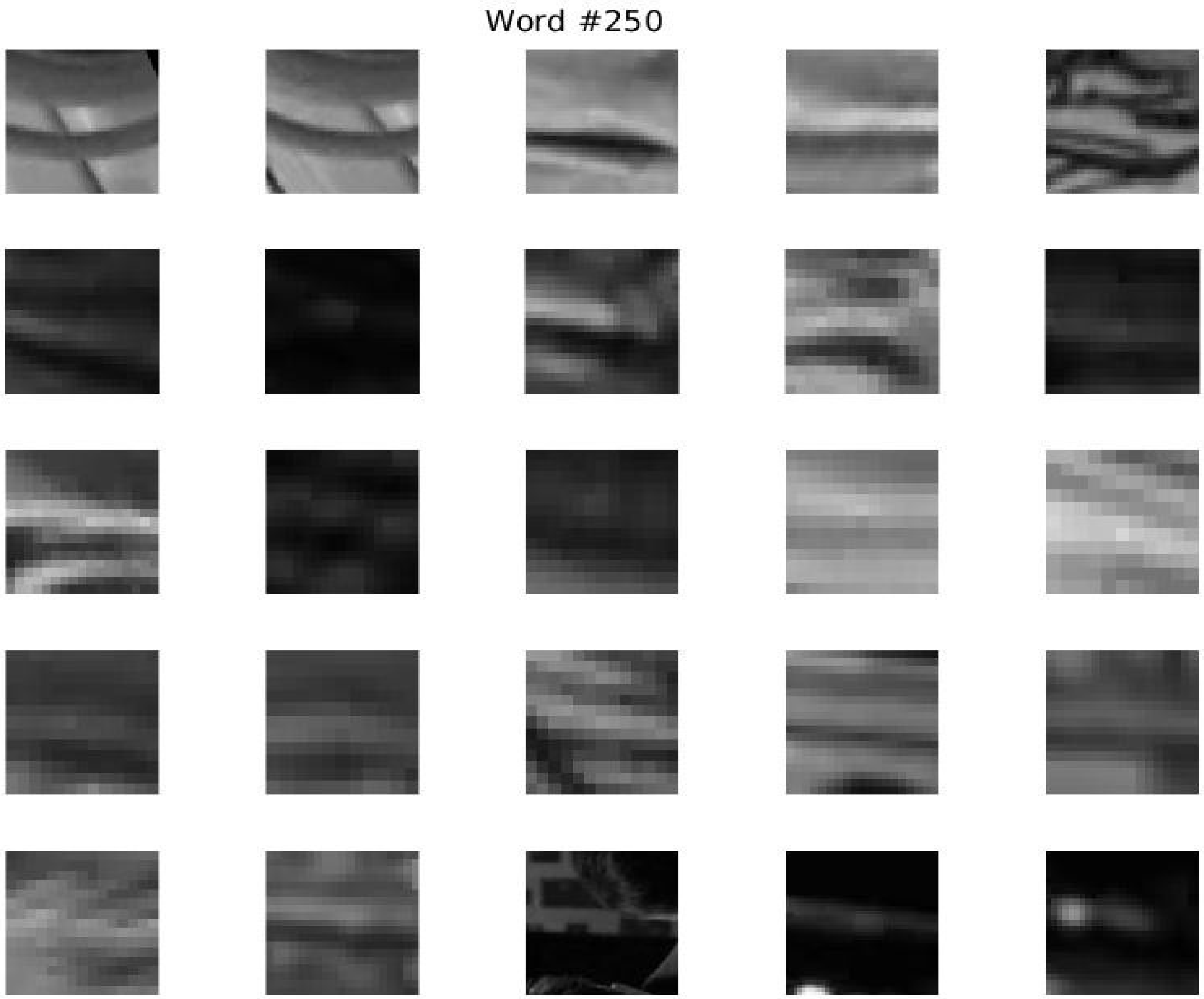
2. Visualizing the vocabulary

Sample 1:



Observation: This vocab describes a shape where the middle part is low in intensity compared to the other parts. It represents a patch with a small black hole in the middle

Sample 2:



Observation: This vocab describes a patch that has a dark line going through the middle. The intensity goes down in the middle horizontal region, and goes back up in other places.

3. Full frame queries (similar frames are ranked left-to-right & top-to-bottom)

Sample 1:

Original Frame: ./frames//friends$_0$000000459.jpeg

Similar Frames

./frames//friends$_0$000000460.jpeg

./frames//friends$_0$000000458.jpeg

./frames//friends$_0$000000461.jpeg

./frames//friends$_0$000003567.jpeg

./frames//friends$_0$000000457.jpeg

Sample 2:

Original Frame: ./frames//friends$_0$000000759.jpeg

Similar Frames

./frames//friends$_0$000000775.jpeg

./frames//friends$_0$000000757.jpeg

./frames//friends$_0$000000758.jpeg

./frames//friends$_0$000000771.jpeg

./frames//friends$_0$000000756.jpeg

Sample 3:

Original Frame: ./frames//friends$_0$000001759.jpeg

Similar Frames

./frames//friends$_0$000001760.jpeg

./frames//friends$_0$000001762.jpeg

./frames//friends$_0$000001761.jpeg

./frames//friends$_0$000001767.jpeg

./frames//friends$_0$000001763.jpeg

Observation:
The algorithm was successful in recognizing similar frames. This was probably pretty easy, since there were many frames that had very similar features. Because of the temporal locality in videos, the similar frames were found very close to the input frame. This made sense because there are not many changes between frames in a video.

Note: I am aware that the number of words and the frames used to create the image vocabulary is smaller than the prompted values. However, my computer cannot take a larger number, and this is the best combination of k and numframes that I could come up with.

4. Region Queries

Sample 1:

Sample 2:

Original Image

Similar Frames

./frames//friends$_0$000001416.jpeg

./frames//friends$_0$000001437.jpeg

./frames//friends$_0$000001414.jpeg

./frames//friends$_0$000001215.jpeg

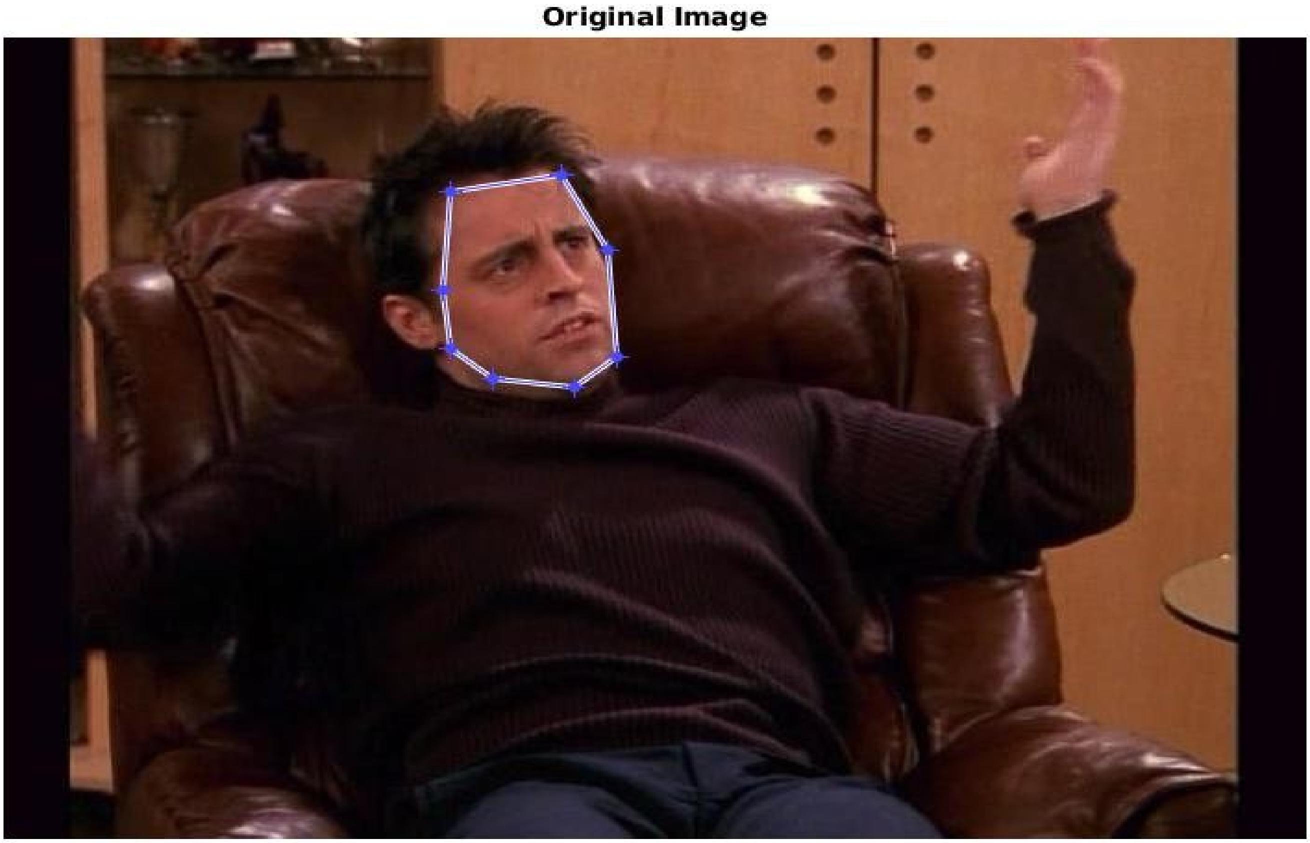
./frames//friends$_0$000001442.jpeg

Sample 3:

Original Image

Similar Frames

Sample 4 (failure case):



Observation:
The regional queries were able to detect frames that contain similar objects in the input region. The objects were in different orientation, or the matches frames had different objects in them, but the algorithm was able to still detect patches that is similar to the the input regions.
Some regions failed to be recognized correctly. When a face was picked as the region, the algorithm simply detected any face, as opposed to a specific one. This has shown that this algorithm can detect features and objects, but it does it in a coarse way.

5. Full frame queries, part 2

Sample 1:
Input Frame:

Original Frame: ./frames//friends$_0$000004503.jpeg

Similar frames found by BoW:

Similar frames found by AlexNet:

Sample 2:
Input Frame:

Original Frame: ./frames//friends$_0$000000394.jpeg

Similar Frames found by BoW:

Similar Frames found by AlexNet:

Observation:

Using the features extracted from AlexNet was much better at finding similar frames. Using CNN allowed the features to be more complex, and was able to detect more high-level features in the frames. Also, the sheer number of features were very different between BoW and AlexNet, so that might have contributed in the different results as well. Overall, using CNN allows us to extract more high-level features, compared to using SIFT.