Roy Dey 914191568
Patrick Liao 913731408

PS3 Writeup

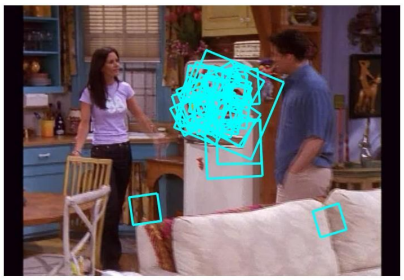
Part 1

1. A SIFT keypoint descriptor is able to identify local features by mapping high-dimensional descriptors to a quantifiable feature space. A SIFT descriptor consists of a vector with each value corresponding to a different dimension. There are a large number of dimensions involved in the calculation of the SIFT vector (e.g. 128). Thus, the significance of a value of a single dimension in a SIFT keypoint descriptor is negligible in the grand scheme of the final placement of the descriptor in the feature space.

2. A neural network without any non-linear activation functions will only be able to output a linear function of the inputs, regardless of the amount of hidden layers present. This effectively makes the neural network a linear regression calculator and not a true deep learning network. This means that such a neural network would not be able to handle complex tasks which require non-linear solutions.

Part 2

1. Note: The polygon encloses the top half of the fridge in the first image.



2.

From the two words selected, the patches corresponding to word 1 appear to depict slit-like features, possibly belonging to things like lips and eyelids. The patches corresponding to word 2 are checkered patterns, possibly belonging to woven surfaces such as carpets, table cloths, and baskets.

3. Query 1

The results of this query seem fairly accurate. Aside from the fourth image, all the other images come from the same scene and look nearly identical. The high accuracy is most likely attributed to the fact that there are a lot of distinguishable local features in the query image. As a result, the BoW for that frame is very unique and eliminates the chance of false positives with most frames in the directory.

Query 2

This output is not quite as accurate. Unlike the first query, there are very few distinguishable features in the query frame. Most of the image is dark and as a result there is little variety in colors and shapes. It looks like the algorithm found images that have a similar hair patterns as the girl in the query image, since that is the feature which stands out the most.

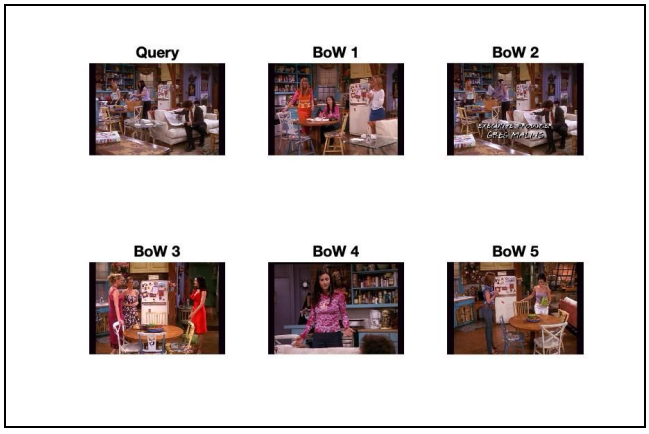
Frame no: 3664

This output is extremely accurate as well. All the matched images come from the same scene and look identical. This is because the query image has a unique feature that stands out: the neon blue lamp. The word that represents this object occurs at a high frequency, which is what most likely causes the high accuracy in image matching.

4. Query 1

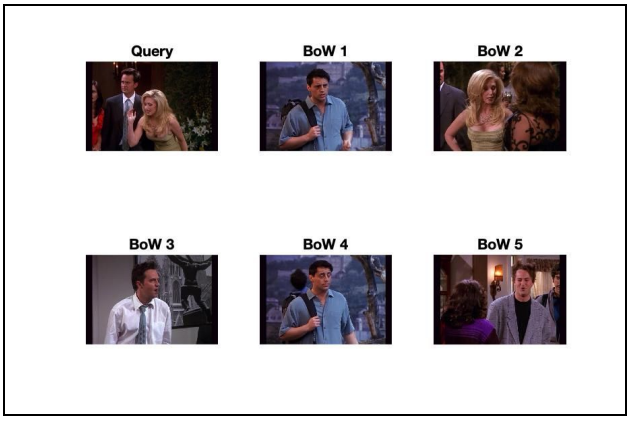
This query has success and fail cases. For BoW1 and BoW2, it most likely matched the woman's shirt to the phone. BoW3 and BoW4 are matched as well, both frames containing the same phone. BoW5 is most likely a match because of the man's shirt and its similarity in color with the phone.

Query 2



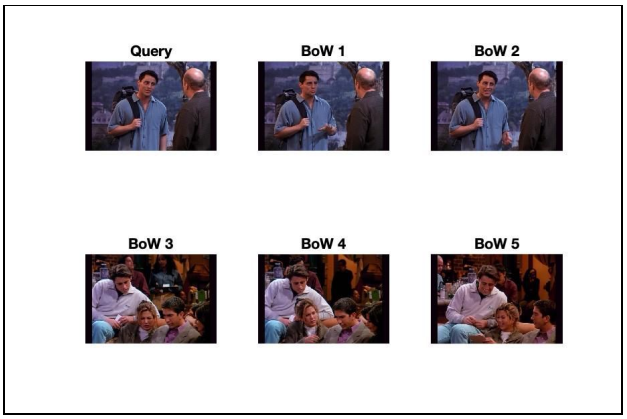
This query is a success in terms of accuracy of capturing the fridge. Since the top half of the fridge contains many unique shapes and colors, there is a lower rate of getting false positives among all the frames in the directory. BoW4 is an exception, most likely because of the similar variety in pattern at the bookshelf area.

Query 3



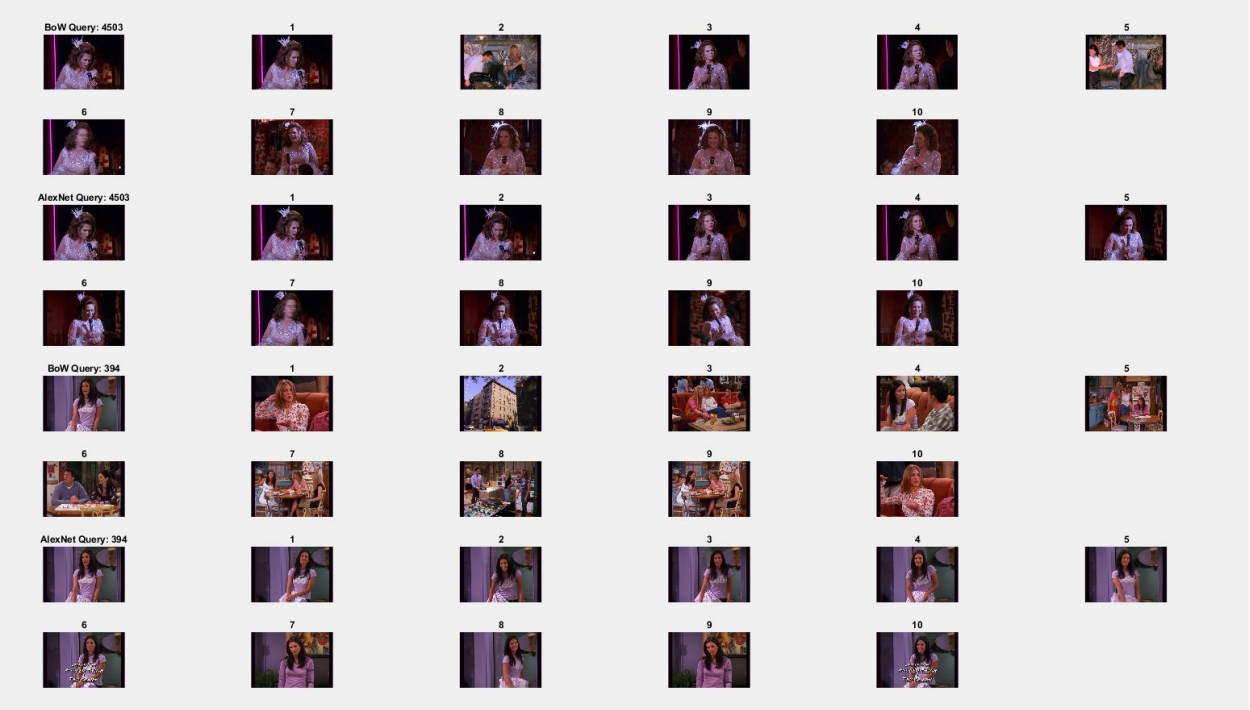
This query seems to be more of a fail case, as there were a lot of false positives that were classified as matching images. The selected region had a unique shape, but there probably was not enough defining colors to separate its BoW out from other images with common words.

Query 4



The results of this query could be better. There are a few other scenes with this man's bald head in it, but the algorithm failed to capture those scenes. Instead, all the similar frames are from the same scenes. The bottom half seem to not have any features of the bald head, and it is interesting that the algorithm grabbed frames from the same scene too. In the algorithm's defence, there are very few distinct textures in the selected region. Therefore the bag of words for the head may contain noise that would produce lots of false positive matches.

5.



When querying frame 4503, the bag of words query retrieved 8/10 relevant images while alexnet retrieved 10/10 relevant images. When querying frame 394, the bag of words query retrieved 1/10 relevant images while alexnet again retrieved 10/10 relevant images. Therefore, it can be seen that alexnet is significantly more accurate when retrieving similar frames. This could be for a multitude of reasons. For instance, alexnet has already been pre-trained on datasets much larger than our hardware can handle with k-means. Furthermore, k-means and neural networks are fundamentally different approaches to image recognition. K-means only finds the average feature dimensions that constitute specific features, whereas neural networks can offer more nuanced and complex solutions using their interconnected layers of neurons.