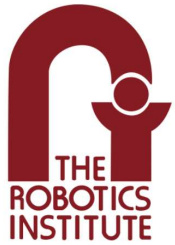
In [ ]:



Computer Vision

16720-B Fall 2022

16720 (B) Bag of Visual Words - Assignment 2

Instructor: Kris Kitani TAs: Sheng-Yu, Jinkun, Rawal, Arka, Rohan

Theory Questions

This section should include the visualizations and answers to specifically highlighted questions from P1 to P4. This section will be manually Graded

Q1.1.1 (5 Points WriteUp)

What visual properties do each of the filter functions (See Figure below) pick up? You should group the filters into categories by its purpose/functionality. Also, why do we need multiple scales of filter responses? **Answer in the writeup. Answer in your write-up.**

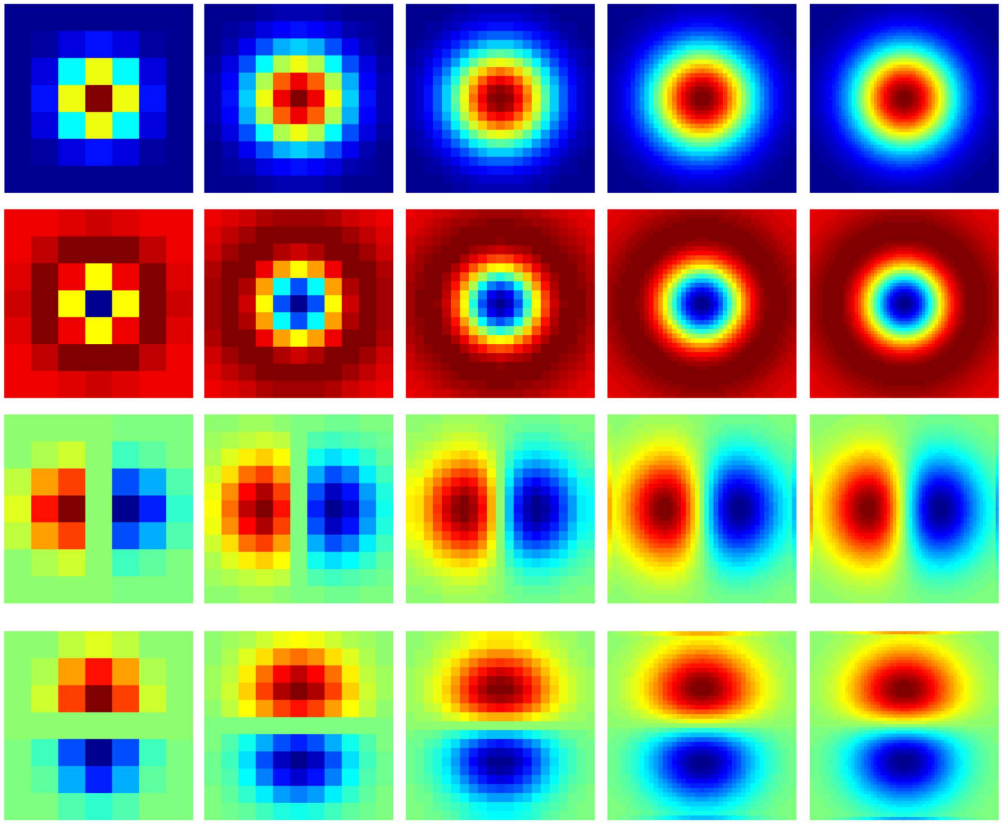


Figure1. The provided multi-scale filter bank

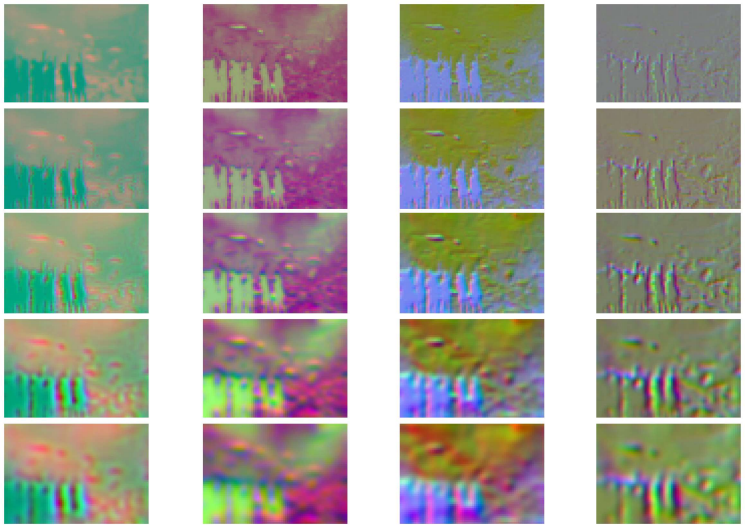
Solution

The 4 types of filters we have used to extract image features are (in order of rows in the above image):

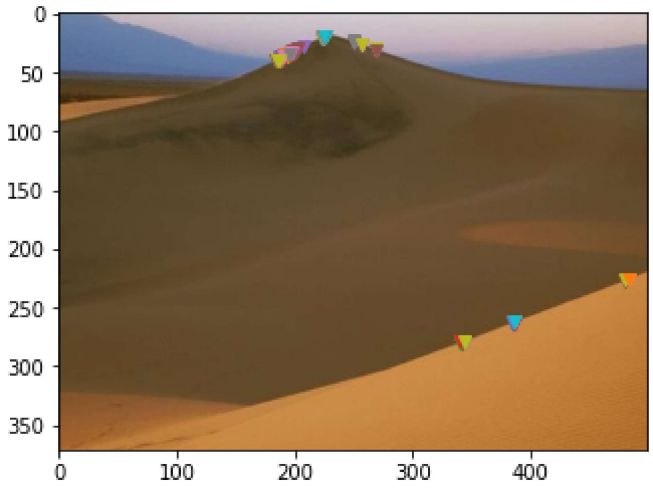
1. Gaussian: smoothes the image, hence removing noise and fine details. lower sensitivity to noise makes the image feature more representative of the class to which that image belongs.
2. Laplacian of Gaussian: this filter can be used for edge detection as well as for blob detection (taking the example of the sunflower image discussed in class, this filter could be used for detecting the center of the sunflower)
3. derivative of Gaussian in the x direction: to get vertical edge information
4. derivative of Gaussian in the y direction: to get horizontal edge information

Filters of mutliple scales are used since features might be of different scales in the images. For ex., sharp vs thick edges

Q1.1.2

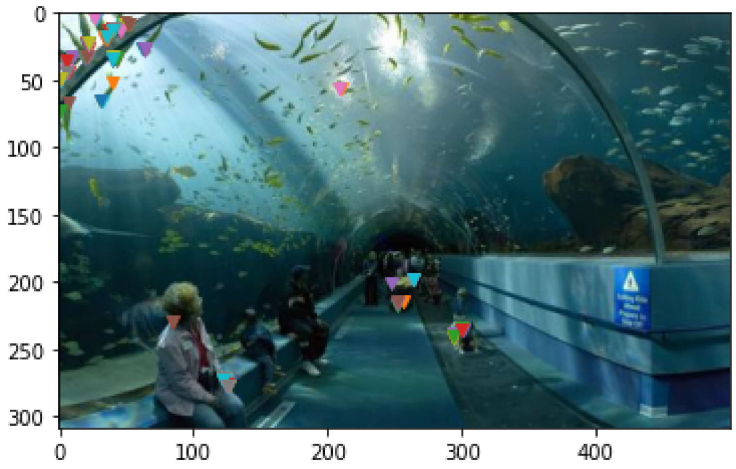


Q1.2.1

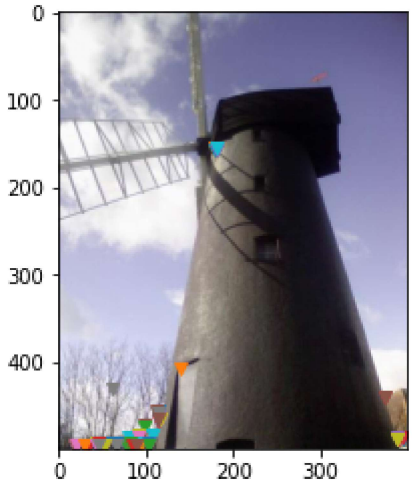


identifies the tip of the dune

identifies the gaps between the balls in the yellow game machine. also identifies corners of the rectangular shaped game machines.



identifies the sharp tips on heads and tails of the fishes. identifies only some fishes, especially the ones in the bright region, which means the detector performs better in brighter regions (more contrast)



identifes some corners in the windmill, but mostly identifies the corners in the trees.

many corners are missed, probably because of the low brightness of the image (lower contrast)

Q1.3.1 (5 Points WriteUp)

Visualize three wordmaps of images from any one of the category. **Include these in your write-up, along with the original RGB images. Include some comments on these visualizations: do the "word" boundaries make sense to you?**. We have provided helper function to save and visualize the resulting wordmap in the util.py file. They should look similar to the ones in Figure 2.

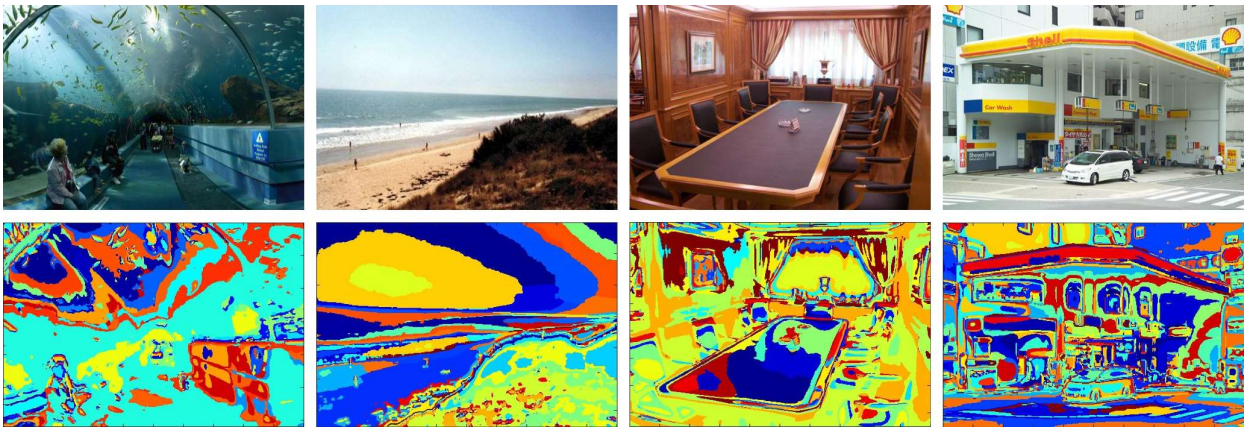
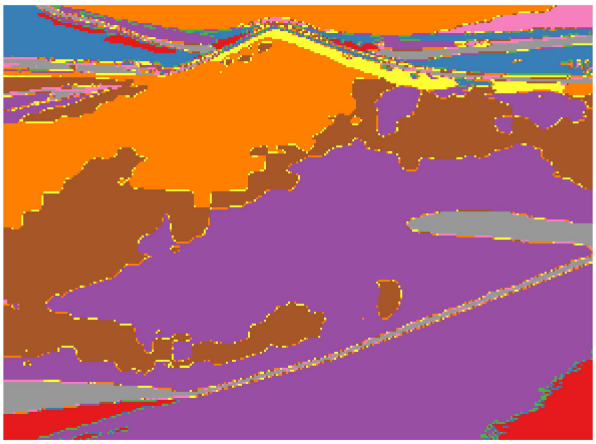


Figure 2. Visual words over images. You will use the spatially un-ordered distribution of visual words in a region (a bag of visual words) as a feature for scene classification, with some coarse information provided by spatial pyramid matching [2]

1.



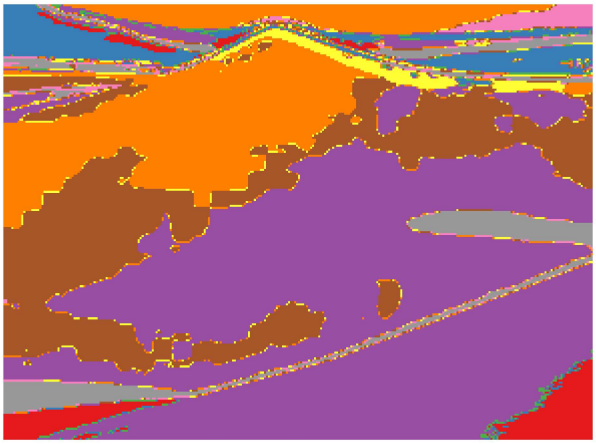
2.

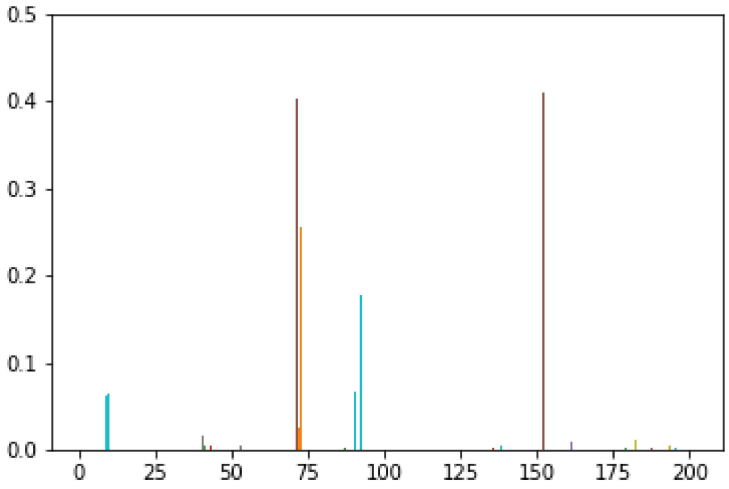
3.

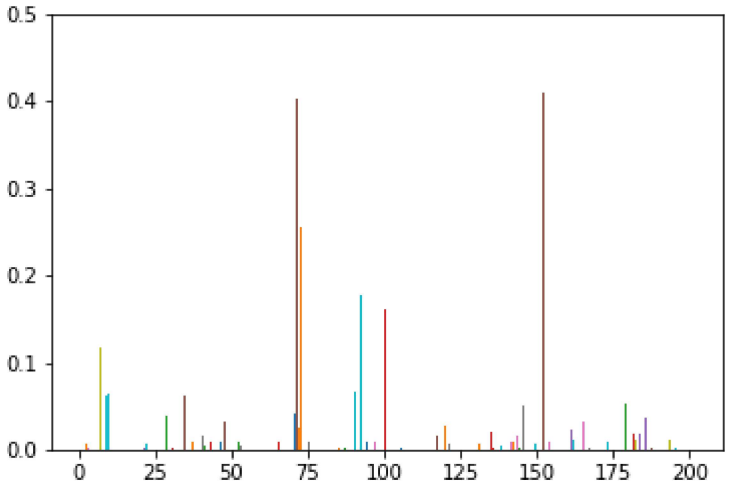
The word boundaries do make sense to an extent. It is clear that the word boundaries are able to partition objects. For example, in the first image, a clear distinction can be seen between the mountain and the background. In the second image, the words boundaries successfully demarcate the circular shapes which are washing machine doors; this feature is common in a laundry and hence will help in identifying similar images.

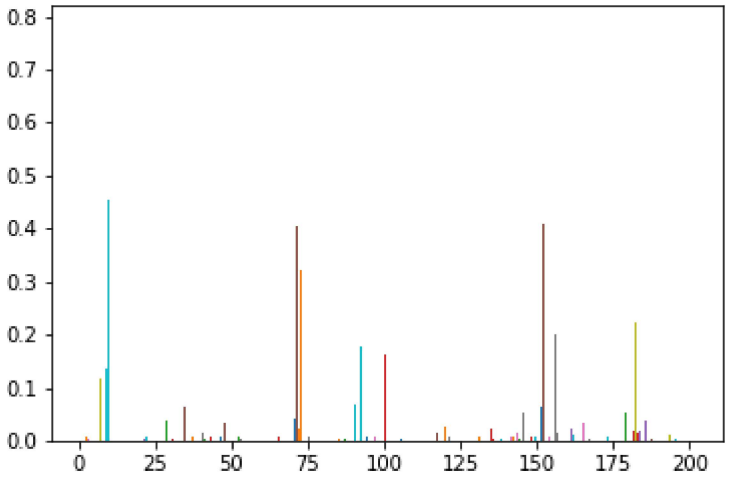
Q2.1

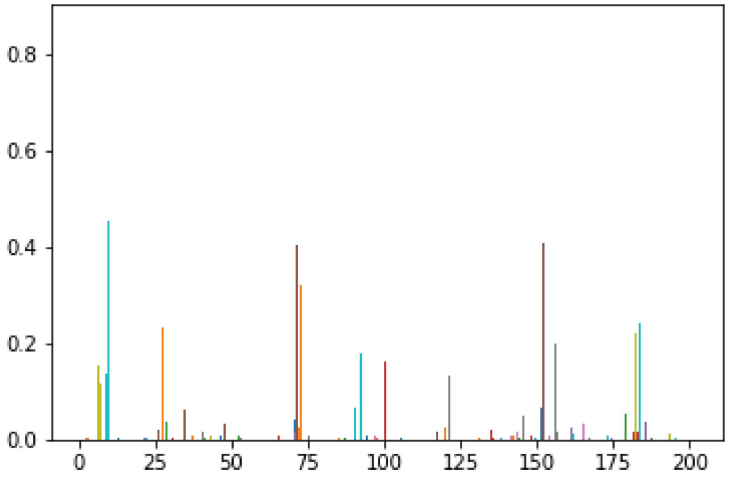
For 5 Images, include their visual word maps and histograms

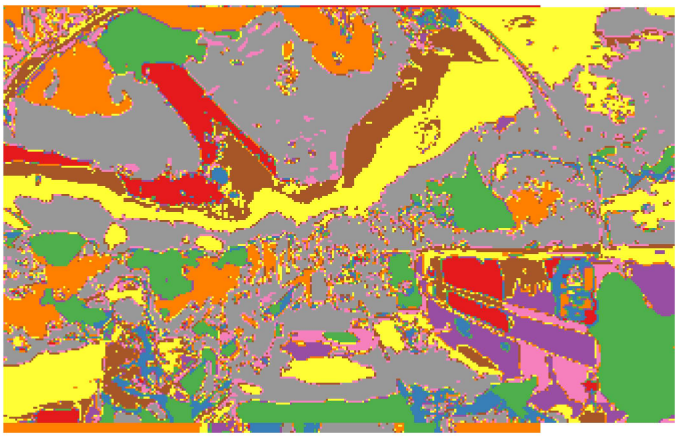


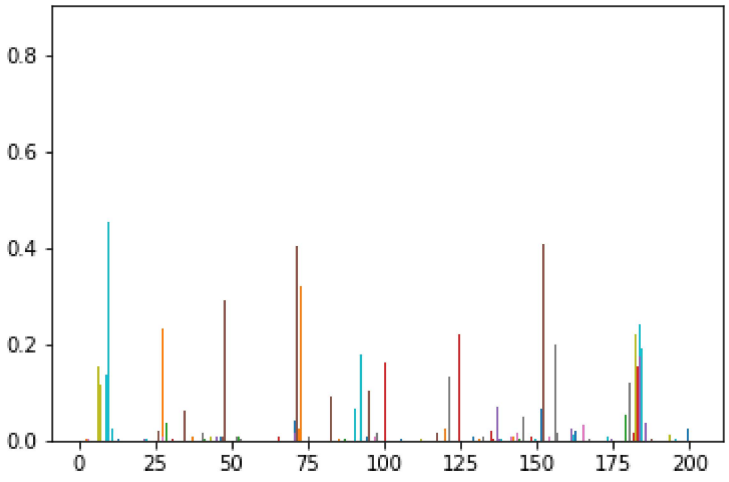



0.5
0.4
0.3
0.2
0.1
0.0
0
25
50
75
100
125
150
175
200


0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
25
50
75
100
125
150
175
200

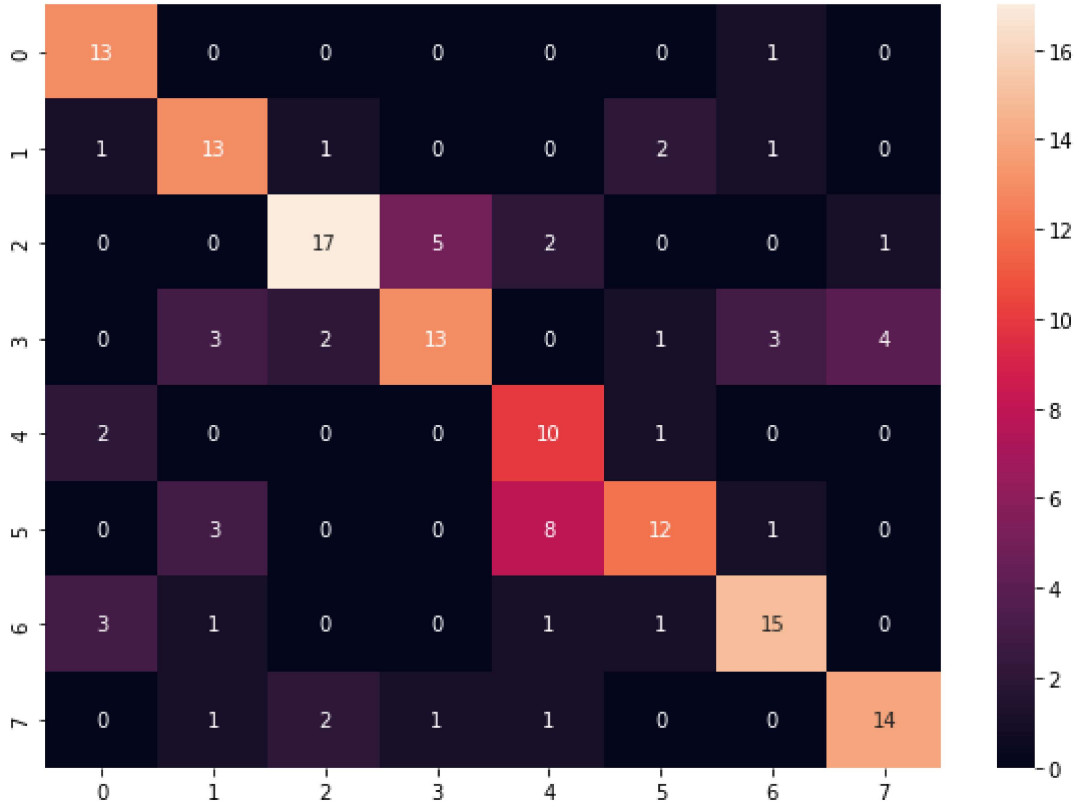

0.8
0.6
0.4
0.2
0.0
0
25
50
75
100
125
150
175
200





Q3.1.1

Submit the visualization of Confusion Matrix and the Accuracy value



Accuracy: 66.875%

Q3.1.2 (5 points WriteUp):

As there are some classes/samples that are more difficult to classify than the rest using the bags-of-words approach, they are more easily classified incorrectly into other categories. **List some of these classes/samples and discuss why they are more difficult.**

Answer

Class 5 (Laundromat) and Class 3 (highway) seem to have performed the worst. This can be attributed to the dense nature of a these classes, for ex: a kitchen has many small objects and doesn't have many defining features (in comparison to the best well performing classes like 2 (desert), which has very few unique objects, mostly the dunes and the sky). Also, the highway might have many vehicles and other objects on the side of the road, and these other objects could vary between different images in the same class. Since we are selecting features from only a few points (Harris corner points), the bag-of-words approach misses a lot of information, espcially in classes where many objects are present in one image.

Q3.1.3 Extra Credit (10 points) Manually Graded:

Now that you have seen how well your recognition system can perform on a set of real images, you can experiment with different ways of improving this baseline system.

Include the changes, modification you made and the impact it had on accuracy.

Tune the system you build to reach around 65\% accuracy on the provided test set (`data/test_data.npz`). **In your writeup, document what you did to achieve such performance: (1) what you did, (2) what you expected would happen, and (3) what actually happened.** Also, include a file called `custom.py/ipynb` for running your code.

YOUR ANSWER HERE

Q3.1.4 [Extra Credit] (5 points write up):

GIST feature descriptor: As introduced during the lecture, GIST feature descriptor is a feature extractor based on Gabor Filters. When we apply it to images, we have to implement the 2D Gabor Filters as described below

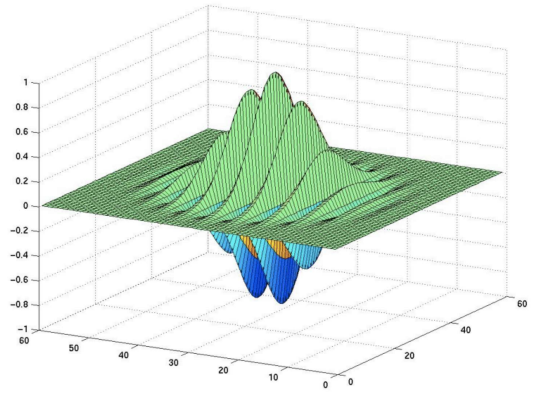
2D Gabor Filters

'Envelope' signal 'Carrier' signal

$$\frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma}} \cos(2\pi(k_x x + k_y y))$$

Gaussian function Modulated by sinusoid

Even filter



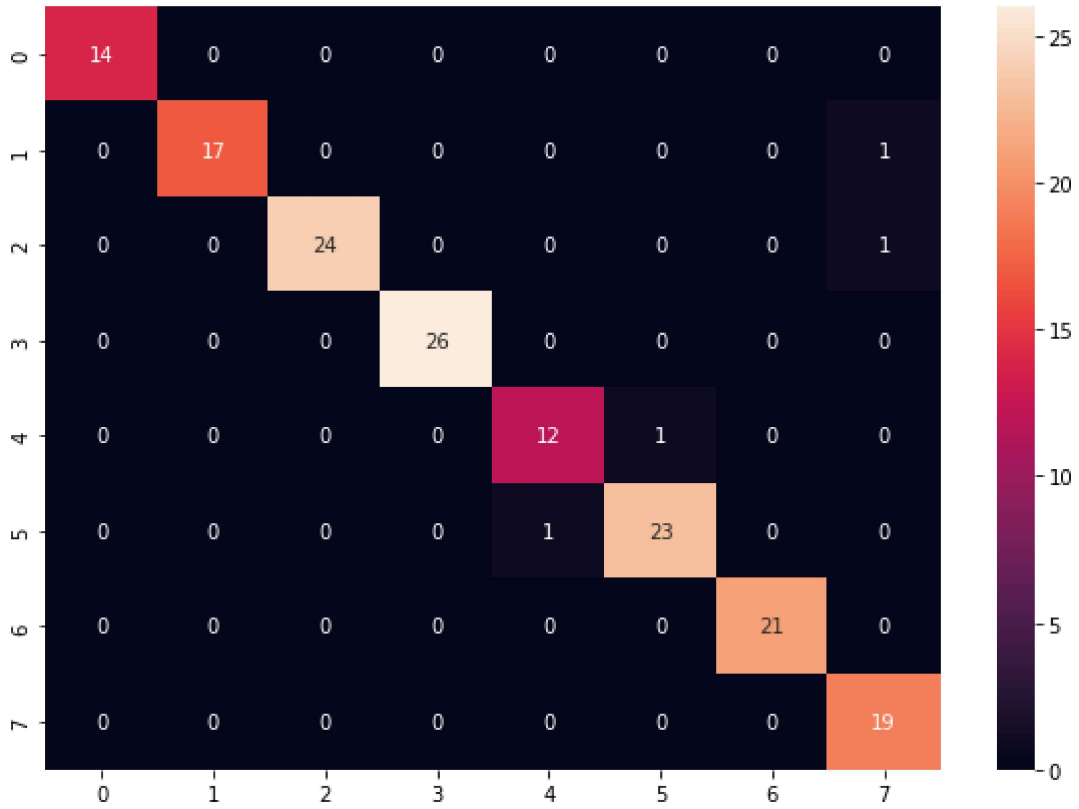
Assuming symmetric Gaussian: $\sigma_x = \sigma_y = \sigma$

In your writeup: How does GIST descriptor affect the performance? Better or worse? Explain your reasoning?

YOUR ANSWER HERE

Q4.2.1 (2 points write up)

Report the confusion matrix and accuracy for your results in your write-up. Can you comment in your writeup on whether the results are better or worse than classical BoW - why do you think that is?



Accuracy: 97.5%

The results are far better than classical BoW approach.

The VGG network is a deep neural network with many layers. As we go from the first hidden layer to deeper layers in this network, the layer learns more and more sophisticated features. The VGG network being a Convolutional Neural Network-based model, it is translation invariant. Also, since the network has mulitple layers where the image is downsampled, the network is also scale invariant. The BoW approach is not sophisticated enough as compared to a Deep Neural Network such as VGG. Also, we only take features from a few points in the image during SPM, hence lot of information is lost, unlike VGG which extracts features from every pixel in the image. Hence the classical BoW approach has lower accuracy.

Q4.3.2 [Extra Credit] (2 points write up)

Report the confusion matrix and accuracy for your ViT results in your write-up. Can you comment in your writeup on whether the results are better or worse than VGG - why do you think that is? A short answer is okay.

YOUR ANSWER HERE

References

[1] James Hays and Alexei A Efros. Scene completion using millions of photographs.ACM Transactions onGraphics (SIGGRAPH 2007), 26(3), 2007.

[2] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recogniz-ing natural scene categories. InComputer Vision and Pattern Recognition (CVPR), 2006 IEEE Conferenceon, volume 2, pages 2169–2178, 2006.

[3] Jian xiong Xiao, J. Hays, K. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recogni-tion from abbey to zoo.2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition,pages 3485–3492, 2010.14

In [ ]: