Assignment 1

Problem 1

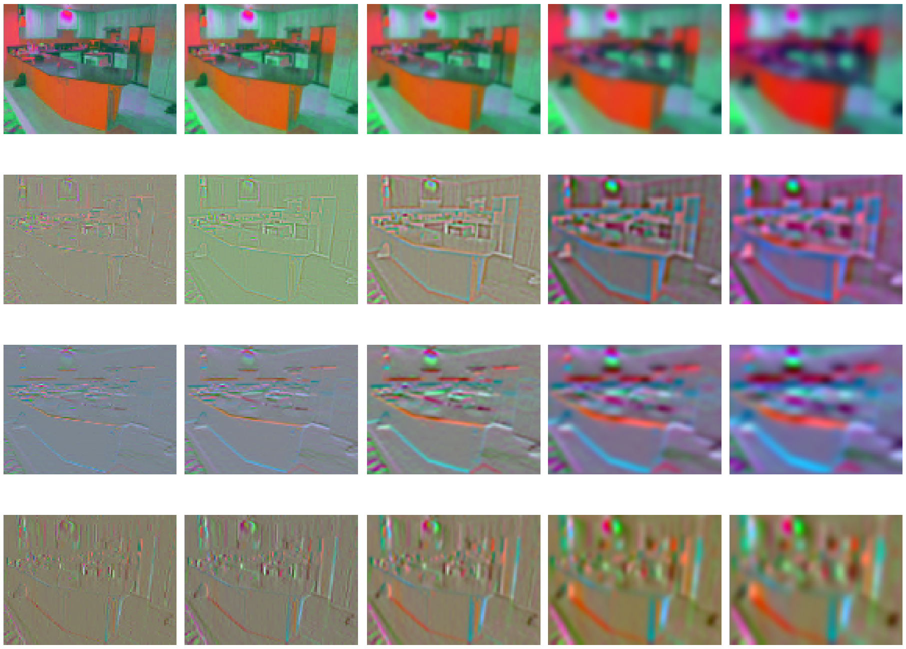
Q1.1.1

A Gaussian filter is a low pass filter and removed the high frequencies in an image, resulting in a smoother version of the image. This can help in reducing unnecessary details and noise from the image. In contrast, the derivatives of Gaussian in x or y direction captures the change in intensity across pixels, thus acting as edge detectors in x and y dimensions. Finally, the LoG uses Laplacian which is the sum of the partial second derivatives in each direction and can thus identify zero crossings in the gradients and pick up edges that form closed contours.

We need multiple scales in the filters because the features or edges which need to be detected occur in different sizes in the image. Large scale values allow capturing of bigger objects and smaller scale values allow detection of finer features.

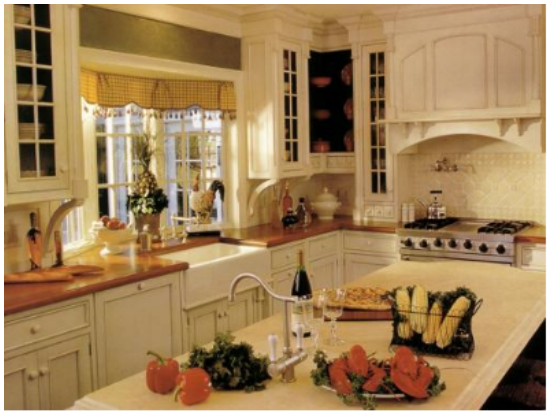
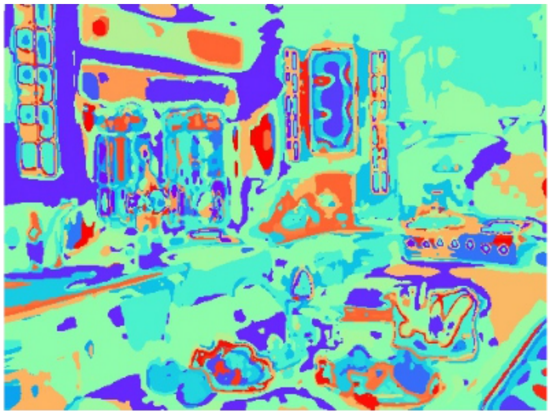
Q1.1.2

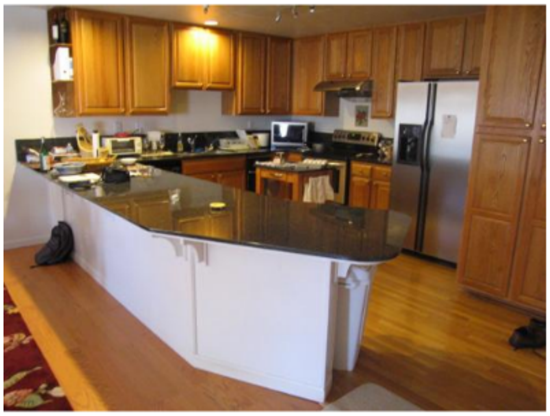
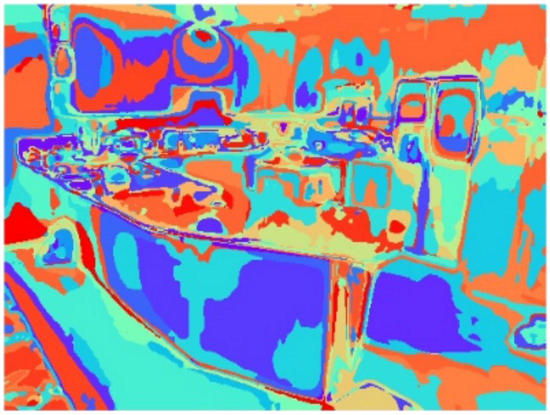
The rows in top-to-bottom order are filters: Gaussian, LoG, Derivative of Gaussian in x and Derivative of Gaussian in y. The columns from left to right are scale size of filters: 1, 2, 4, 8, $8\sqrt{2}$.

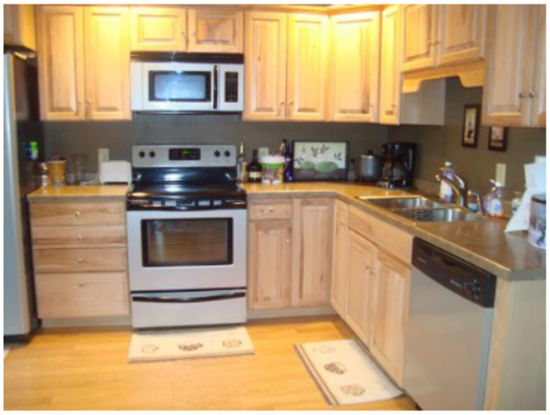
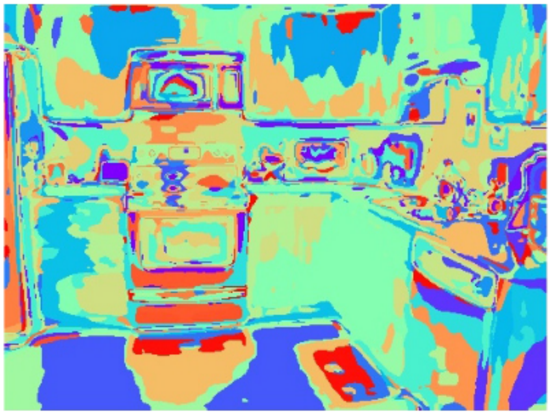


Q1.3

The visual wordmaps from 3 random images from kitchen scenes are visualized below. They are obtained with a $\alpha = 300$ and $K = 150$. The wordmap clearly shows that there is a local continuity in features and neighbouring areas in the image separated by edges belong to the same cluster. Additionally, there are some semantic similarities which can be seen in the illumination and texture of the cluster colors of similar regions across images. For example the lighter blue colors often refer to the smooth wooden surfaces in a kitchen.

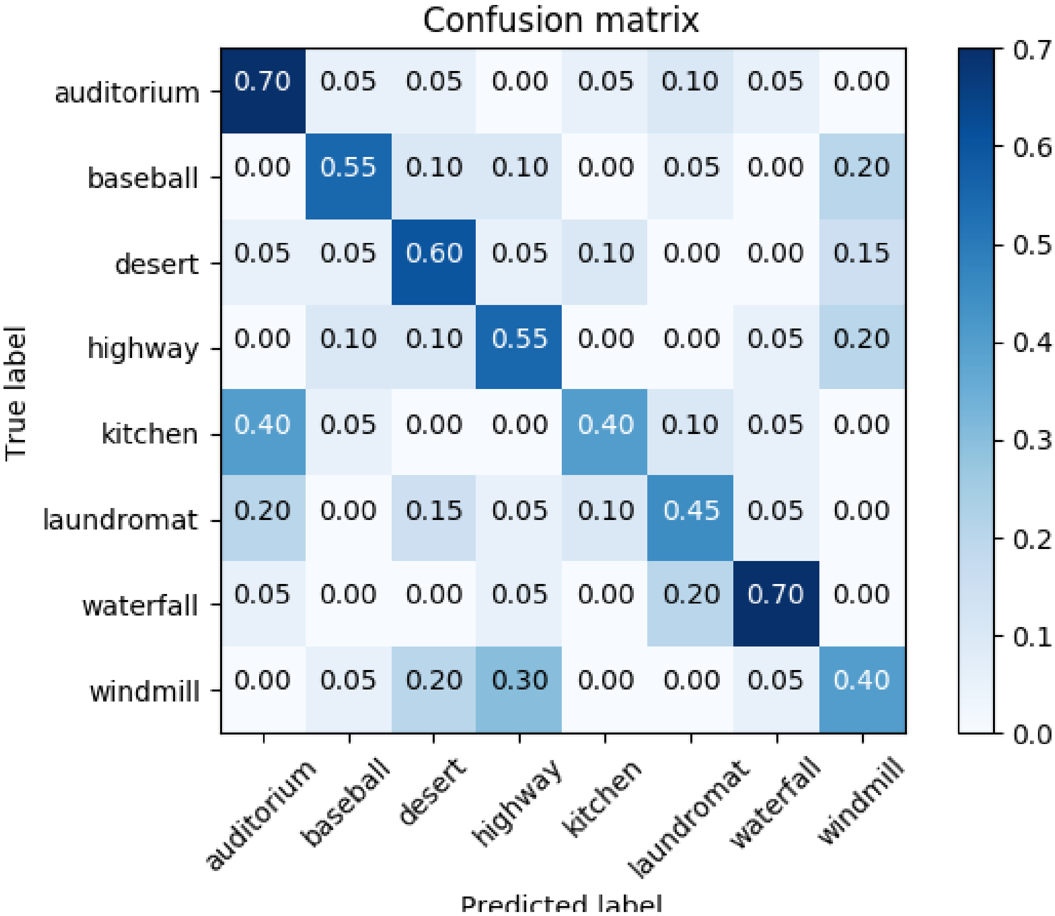



Problem 2

Q2.5

The obtained accuracy of the spatial pyramid of features is 54.375%.
The confusion matrix from the evaluation is as follows:



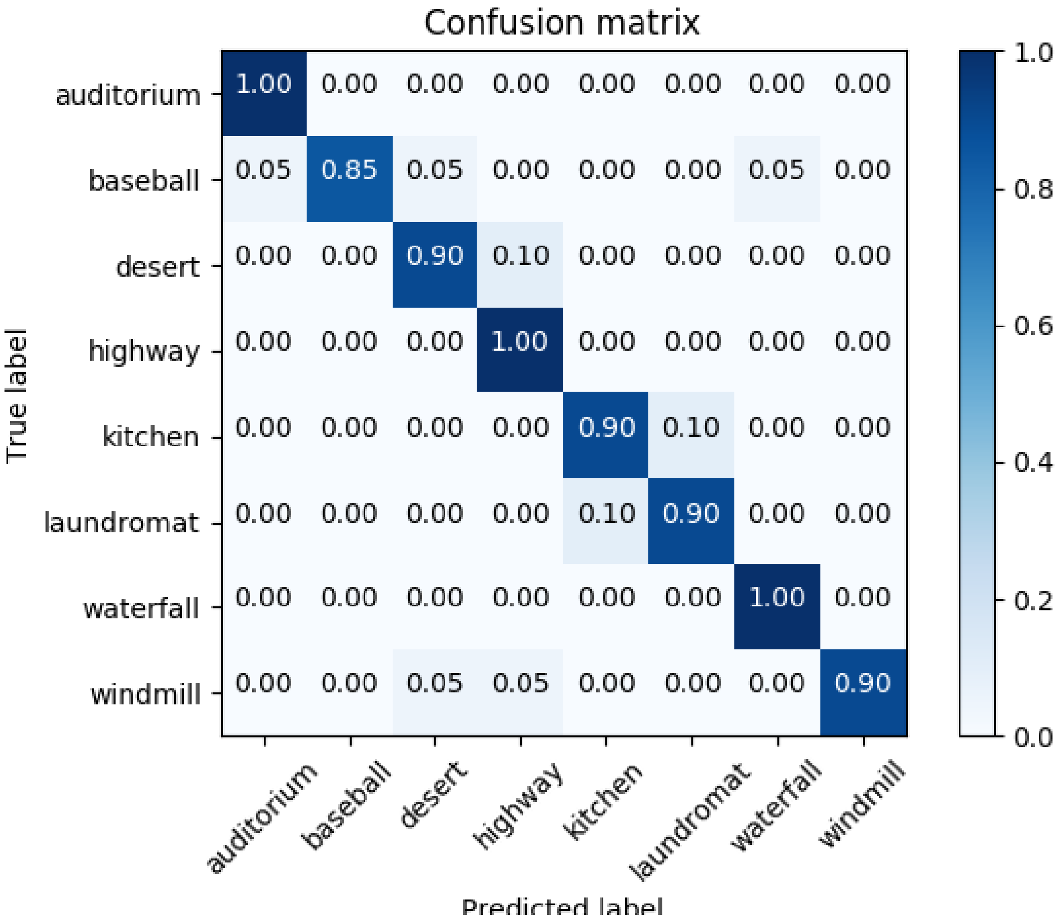
Q2.6

The failure cases of the SPM based recognition model can be observed from the confusion matrix above. Three classes get an accuracy of lower than 50%. The first is the kitchen class which is predicted correctly only 40% times. A plausible reason for low performance here is that it contains many small and cluttered objects which are not modelled well by the histogram based bag of features. It is also interesting to see that 40% of the times kitchen is mistaken for an auditorium scene which visuall often has the same illumination and global texture. The second class on which the model also performs badly is the windmill class with 40% accuracy. The model confuses it with the highway class which makes sense since windmills are usually found in open, grassy areas. The third class is the laundromat class which gets an accuracy of 45%. 20% of the mis-classifications are predicted as an auditorium. The common theme in the results is that the word of visual bags model with nearest neighbour classification bases its predictions on global scene information which are often common to multiple classes. Also, smaller objects which might not contribute much to the scene feature in an histogram but are good indicators of a scene class are not utilized. Finally, the loss of spatial order might also have an effect on the performance.

Problem 3

Q3.2

The obtained accuracy of the pre-trained VGG16 model is 93.125%.
The confusion matrix from the evaluation is as follows:



	auditorium	baseball	desert	highway	kitchen	laundromat	waterfall	windmill
auditorium	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
baseball	0.05	0.85	0.05	0.00	0.00	0.00	0.05	0.00
desert	0.00	0.00	0.90	0.10	0.00	0.00	0.00	0.00
highway	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
kitchen	0.00	0.00	0.00	0.00	0.90	0.10	0.00	0.00
laundromat	0.00	0.00	0.00	0.00	0.10	0.90	0.00	0.00
waterfall	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
windmill	0.00	0.00	0.05	0.05	0.00	0.00	0.00	0.90

The results from the deep features are much better than the bag of words model. Even though the VGG16 model has been trained on ImageNet, it learns general visual features which are useful across tasks pertaining to natural images. This can be attributed to multiple reasons. One is the fact that deep features are not hand-engineered but are learned through back-propagation. Through this learning, the initial layers across tasks end up learning basic features like edges, colors and textures which are as useful for scene classification as they were for object recognition in ImageNet. Another reason for improvement in performance is that the spatial information in CNNs is preserved to an extent greater than BoW models. On the same line of reasoning, the set of salient objects which can significantly determine the scene class are better recognized by a VGG16 model than a bag of words approach.