16-720 B Assignment1

Sajal Maheshwari
sajalm@andrew.cmu.edu

September 23, 2019

1 Representing the World with Visual Words

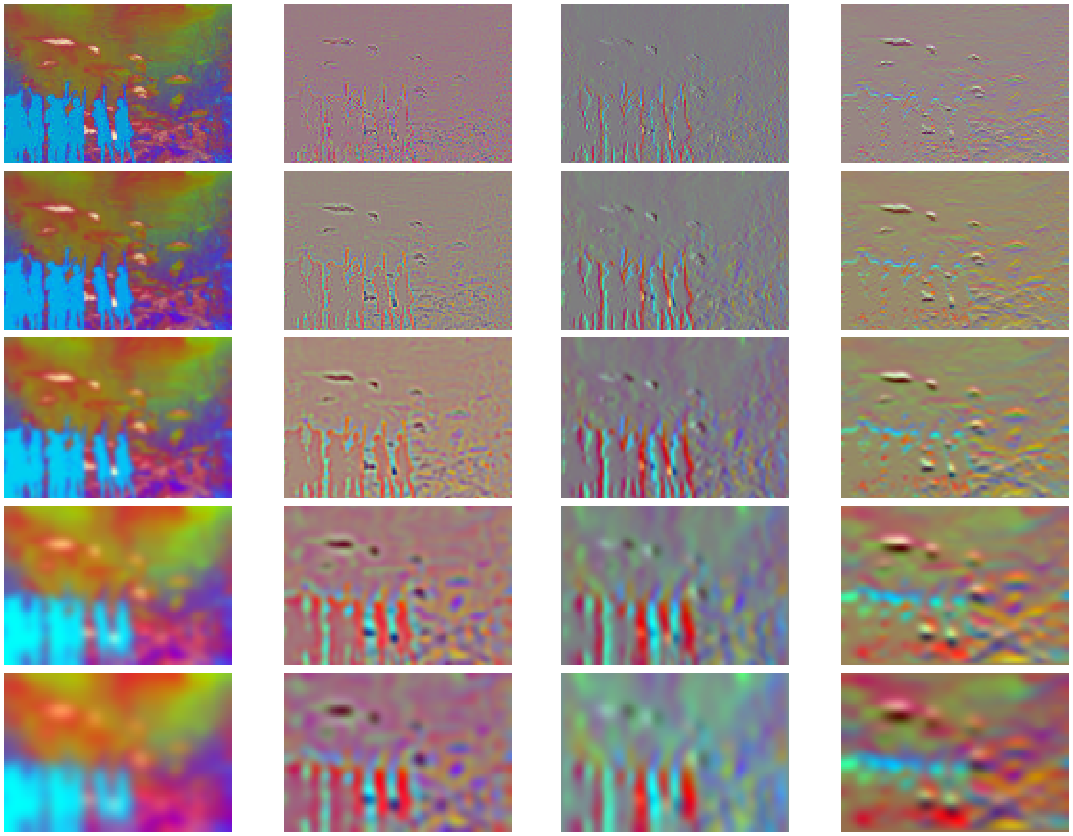
1.1 Extracting filter responses

1.1.1

- Gaussian filter - Gaussian filter acts like a blurring kernel over the image.
- Laplacian of Gaussian - Laplacian of Gaussian acts like an edge detector. Since it is a symmetric filter in 2-D, the single function is able to detect edges in both x and y direction.
- Derivative of Gaussian (x direction) - Derivative of Gaussian in x-direction also acts as an edge detector. However, this function can only extract the vertical edges, leaving the horizontal edges unchanged.
- Derivative of Gaussian (y direction) - Derivative of Gaussian in y-direction also acts as an edge detector. However, this function can only extract the horizontal edges, leaving the vertical edges unchanged.

We need multiple scales for each filter because we need to collect distinguishing features and their outputs at different scales. This helps us in generating a robust feature vector for the image which is invariant to scale changes, resulting in better performance during recognition/matching of images.

1.1.2



1.2

1.3

Examples of three images from the 'park' category with their corresponding wordmaps can be seen in Fig. 1. The wordmaps resemble segmentation maps with each segment containing the same words. This is consistent with our understanding of the visual words, as regions with a similar structure (filter response) getting assigned to the same centroid(visual word in the dictionary) during the clustering.

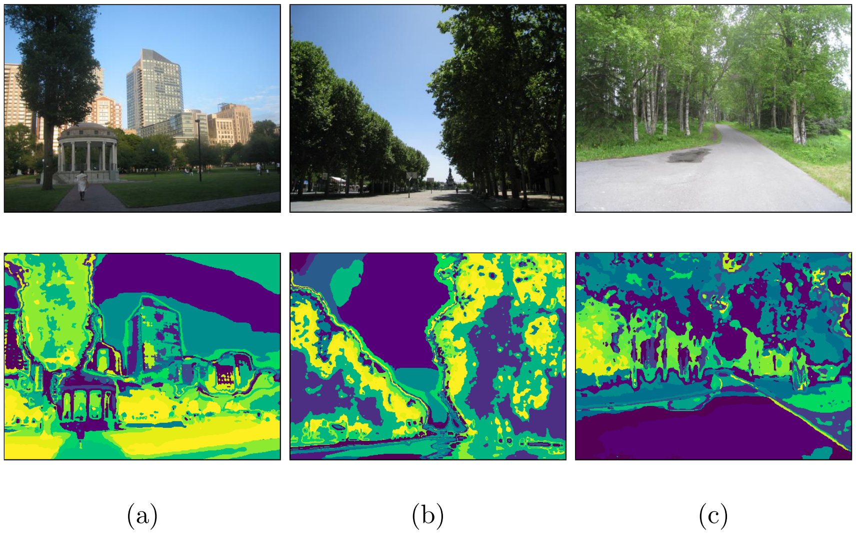


Figure 1: Comparison of input images with their visual 'wordmaps' for images of the 'park' category. As can be seen from the images, there is a clear distinction between different regions of the inputs (roads, trees and buildings have a distinct difference in (a). Similarly, (b) and (c) have a clear demarcation between trees and roads.)

2 Building a Recognition System

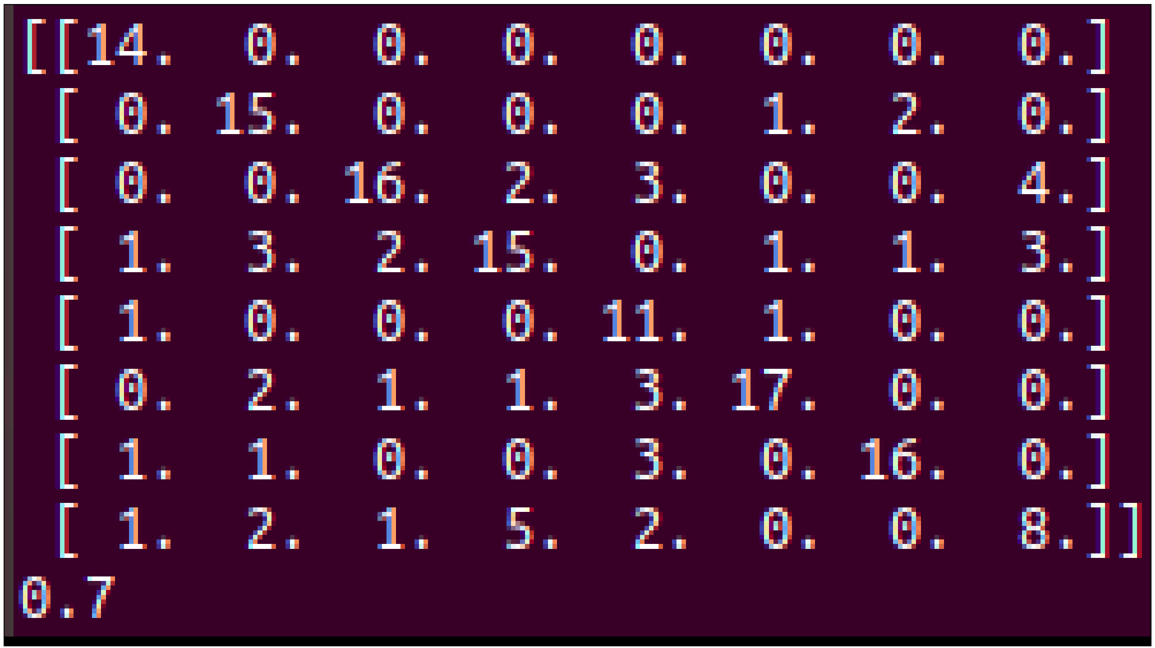
2.1

2.2

2.3

2.4

2.5

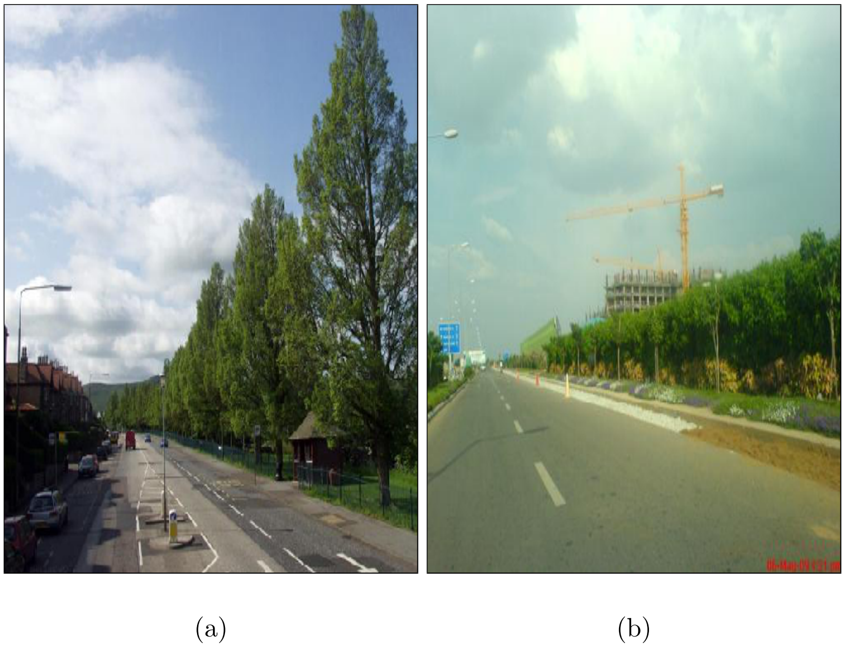


```
[[14.  0.  0.  0.  0.  0.  0.  0.]
 [ 0. 15.  0.  0.  0.  1.  2.  0.]
 [ 0.  0. 16.  2.  3.  0.  0.  4.]
 [ 1.  3.  2. 15.  0.  1.  1.  3.]
 [ 1.  0.  0.  0. 11.  1.  0.  0.]
 [ 0.  2.  1.  1.  3. 17.  0.  0.]
 [ 1.  1.  0.  0.  3.  0. 16.  0.]
 [ 1.  2.  1.  5.  2.  0.  0.  8.]]
0.7
```

$Accuracy = 70\%$

2.6

In this section, we show some hard examples, seen in Fig. 2. The accuracy of the recognition system generated are most susceptible to making an error by misclassifying a ‘highway’ image to ‘windmill’ image. The possible reasons for this misclassification can be the additional lamp post in the (a) example and the construction equipment at the lane ends. Since both the classes primarily consist of outdoor scenes, such anomalies are more likely to happen, as a large part of the images from both categories can be similar.



(a)

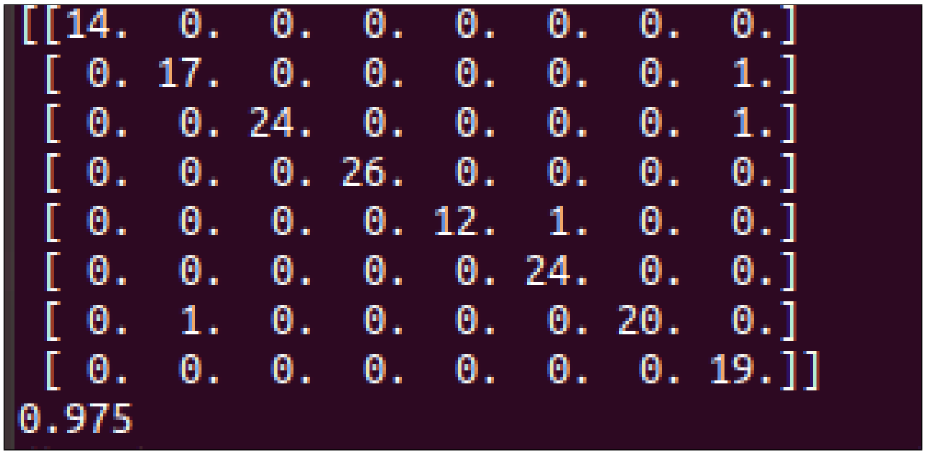
(b)

Figure 2: Examples of failure cases when recognition done using the handcrafted features. These images belong to the ‘highway’ category but are mis-classified as being in the ‘windmill’ category by the prediction system.)

3 Deep Learning Features - An Alternative to "Bag of Words"

3.1

3.2



```
[[14.  0.  0.  0.  0.  0.  0.  0.]
 [ 0. 17.  0.  0.  0.  0.  0.  1.]
 [ 0.  0. 24.  0.  0.  0.  0.  1.]
 [ 0.  0.  0. 26.  0.  0.  0.  0.]
 [ 0.  0.  0.  0. 12.  1.  0.  0.]
 [ 0.  0.  0.  0.  0. 24.  0.  0.]
 [ 0.  1.  0.  0.  0.  0. 20.  0.]
 [ 0.  0.  0.  0.  0.  0.  0. 19.]]
0.975
```

$$Accuracy = 97.5\%$$

We get a very high accuracy as compared to the BoW approach. This is because the deep features extracted at multiple scales correlate extremely well within classes. The pretrained model, trained on ImageNet captures the discriminative features extracted from the image through its series of convolution and Non-linearities. These discriminative features efficiently ensure that all images similar to each lie close, thus producing better results.

In this section we also verify the correctness of the implementation of network layers in solve 3.1. The difference between the outputs after the second dense layer (fc7) match in both the implementations. The error obtained is of the order 10^{-12}).