1 Introduction

In this assignment, We will make an image classification and then calculate the accuracy of the model. To compare images, we will use some techniques, likes Gabor filter bank, SIFT and bag of visual words, to get some special information about images. And then we will compare the accuracy results. To compare the images, we will use the 1-NN algorithm and euclidean distance (see eq. 1).

$$distance = \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2} \tag{1}$$

The rest of the papers as follows, The method I followed and details of the solution in Section 2, Experimental results for different parameters and comments on them in Section 3, the weakness of my implementation in Section 4 and the bonus part that I used a different distance formula in Section 5.

2 Implementation Details

To start implementation details, firstly to compare the actual labels and predicted labels for query images, I separate the images class by class like in train dataset and then I load all images in grayscale form into dictionaries that is separated class by class.

2.1 Gabor Filter Bank

The first task of part I is using gabor filter bank feature vectors taht gives us information about texture of the images like orientation, frequency etc. To create gabor filter bank feature vectors, I firstly create gabor filters with different orientations to extract texture information from each image. I apply each filter to images with convolution after that I take the mean of each filtered image and store them in an array. At the end, I come up with 1xN (N is the number of filters that I used) vector for each image. Then I normalize the each vector between 0 and 1. I made filtering and normalizing for both train images and query images. Then I pass these feature vectors into k-nearest neighbor algorithm for k $= 1$ and then calculate average and class based accuracies. I repeat these steps for different number of gabor filters in different orientations. To take different orientations I increment the theta for each iteration. I determine incremented value as eq. 2

$$incrementedvalue = \frac{180}{numberofgaborfilters} \tag{2}$$

2.2 SIFT

The second task of part I is using scale invariant feature transform(SIFT) that extract key points and compute its descriptors. SIFT that is a object recognition algorithm is both scale and rotation invariant and also robust against shrinkage and expansion. I extract SIFT feature vectors for both query images and train images using openCV library and then pass the average of the descriptors to the k-nearest neighbor algorithm for k $= 1$ and then calculate average and class based accuracies.

2.3 Bag of Visual Words

In the second part of this assignment, I used bag of visual words with and without spatial tiling. The bag of visual words are used for image classification. Like in NLP, we extract some special words from images. And using these visual words, the frequency histograms are created for each of images. Firstly, I use bag of visual words without spatial tiling. To extract visual words, I firstly get the feature vectors using SIFT. Then I cluster these features using k-means clustering algorithm. (k should be a large number.) The central points, that I get from clustering algorithm, are visual words. (In another saying, codebooks) Using these words, I build frequency histogram which is a vector of occurrence counts of a vocabulary of local image features. I create these histograms for both query and train images and pass these histograms to the k-nearest neighbor algorithm for k $= 1$ and then calculate average and class based accuracies. I repeat these steps for different number of k values of k-means clustering algorithm. Secondly, I used bag of visual words with spatial tiling that means before extracting features from each image, I split the image into k-tiles. After that from each tile, I extract feature vectors using SIFT algorithm and continue like in the above(without spatial tiling).

3 Experimental Results

3.1 Gabor Filter Bank

To see the effect of number of gabor filters in different orientations, I run the code for 3 times.

	Class Based Accuracies										
Num. of filter	Avg. Acc.	Bear	Butterfly	Coffee-mug	Elk	Fire-truck	Horse	Hot Air Baloon	Iris	Owl	Teapot
40	16	20	20	0	20	0	0	20	20	20	40
60	18	20	0	40	0	0	0	20	20	20	20
90	20	20	40	20	60	0	0	0	20	0	40

Table 1: The average and class based accuracies according to different gabor filters and orientations

The table 1 shows the average and class based accuracies for different number of gabor filters in different orientations. The best number of gabor filter is 90. It

means when we use 90 different gabor filters the orientations of the filters come the best that means texture of the images are extracted best. (the degree of orientations are started from one and increased step by step - step is calculated using the eq. 2)

I take the 3 sample from 3 different class in queries, and find the 5 most similar image. And the results as follows:

Figure 1: The test images from 3 different classes in query

Figure 2: The 5 most similar image to the elk using Gabor Filter Bank

Figure 3: The 5 most similar image to the iris using Gabor Filter Bank

Figure 4: The 5 most similar image to the teapot using Gabor Filter Bank

Although the elk has the most class based accuracy, the model cannot find any class that is elk in figure 2. The model may be affected by grayscale-colors or texture of the test image is different from other elk images.

3.2 SIFT

	Class Based Accuracies										
Type of the technique	Avg. Acc.	Bear	Butterfly	Coffee-mug	Elk	Fire-truck	Horse	Hot Air Baloon	Iris	Owl	Teapot
Gabor filter(k = 90)	20	20	40	20	60	0	0	0	20	0	40
SIFT	22	20	0	20	20	40	40	20	20	20	20

Table 2: The average and class based accuracies according to different gabor filters and sift feature vectors

In table 2, we see that the average accuracy of the SIFT is more than the gabor filter bank. Because SIFT is rotation and scale invariant and robust the changes but gabor filter just gives the information about texture of image.

Figure 5: The 5 most similar image to the elk using SIFT

Figure 6: The 5 most similar image to the iris using SIFT

Figure 7: The 5 most similar image to the teapot using SIFT

In figure 5, 6, 7; we see the 5 most similar images for the given test samples. Although, the class base accuracy for the elk is %20, there is not any elk image in the result images. The models may be affected by color distribution and texture of the images.

3.3 Bag of Visual Words

The first part of bag of visual words are calculating the accuracies without spatial tiling.

	Class Based Accuracies										
K	Avg. Acc.	Bear	Butterfly	Coffee-mug	Elk	Fire-truck	Horse	Hot Air Baloon	Iris	Owl	Teapot
50	16	40	0	0	0	40	40	0	0	20	20
100	20	40	0	0	0	60	60	20	0	20	0
150	24	20	20	0	20	40	60	40	0	0	40

Table 3: The average and class based accuracies according to Bag of Visual Words without spatial tiling using different number of k values of k-means clustering algorithm.

NOTE: Because of lack of my computer's resources, I cannot use big numbers for K value in k means clustering. When I try to use it takes too long time and after 4 hours, the computer crashed. So I used small numbers for k value.

In table 3, we see that when k is increased, the average accuracy is also increase. Because of k represent the number of visual words so having more visual words allow us to get more correctly distributed histogram. For example, in NLP, we can extract the topic of the document just looking at the visual words. So like in NLP, we can extract the content of an image just looking at the visual words. The more visual words we have, the more we can specify the image content correctly. Because of lack of my computer's resources, I cannot run the code for bigger k values but if I would have run the code for $k = 500$, the accuracy will be around %30 or maybe more than %30.

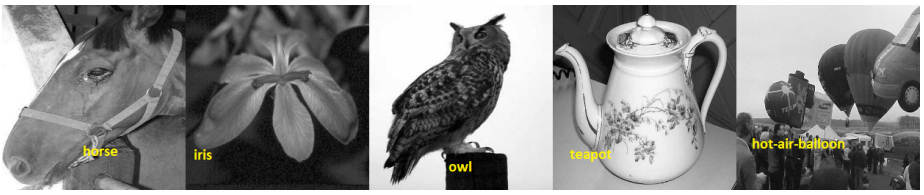

Figure 8: The 5 most similar image to the elk using bag of visual word without spatial tiling

Figure 9: The 5 most similar image to the iris using bag of visual word without spatial tiling

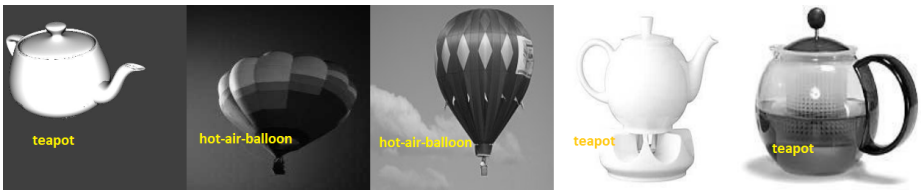

Figure 10: The 5 most similar image to the teapot using bag of visual word without spatial tiling

When we give 3 images from 3 different classes in the query dataset, the results of the 5 most similar images are shown in figure 8, 9 and 10. Although the %40 percent of the images are coming iris in figure 9, the class based accuracy of the iris in table 3 for k = 150 is 0. That means one of the other images, that are not iris, is more than close the iris in the test(see fig. 1).

	Class Based Accuracies										
K	Avg. Acc.	Bear	Butterfly	Coffee-mug	Elk	Fire-truck	Horse	Hot Air Baloon	Iris	Owl	Teapot
50	16	20	20	0	40	0	40	0	40	0	0
100	18	0	20	20	20	40	40	0	20	0	20
150	20	100	40	20	0	40	0	0	0	0	0

Table 4: The average and class based accuracies according to Bag of Visual Words with spatial tiling for 500 tiles using different number of k values of k-means clustering algorithm.

Bovw	Avg. Acc.
Without spatial tiling	24
With spatial tiling	20

Table 5: Comparison of the bag of visual words' average accuracy. K-means(k = 150) and for spatial tiling, number of tile = 500

In table 4, we see that when number of k is increased, the accuracy is also increased like in 3. Because if number of visual words is increased, we can specify the image more clearly. On the other hand, in table 5 there is a comparison between two distinct bag of visual words. Both of them have the same clustering number but one of them without a spatial tiling and the other is with spatial tiling. (with 500 tile.) We see that with spatial tiling, the accuracy decrease. When we are doing tiling, we may split the image over an important feature and this time we lost this feature. So it may cause the decrease the accuracy.

Figure 11: The 5 most similar image to the elk using bag of visual word with spatial tiling

Figure 12: The 5 most similar image to the iris using bag of visual word with spatial tiling

Figure 13: The 5 most similar image to the teapot using bag of visual word with spatial tiling

Although class based accuracy of the elk is %0 , in figure 11, there are two different elk images. It shows that one of the other classes are most similar to elk in the test images.(see figure 1).

If I change the test images, I may get more proper results from 5 most similar images.

Num. of Tile	Avg. Acc.
500	20
200	16
100	12

Table 6: The difference in the average accuracy when we used different number of tiles for ba of visual words

In table 6, we see that when the number of tile goes down, average accuracy is also goes down. Beacuse while the more we have small pieces of image, the more we can extract feature from each piece. So the accuracy of the visual words are more convenient.

4 Conclusion

Technique	Avg. Acc.
Gabor Filter Bank	20
SIFT	22
Bag of Visual Words Without Spatial Tiling	24
Bag of Visual Words With Spatial Tiling(tile = 500)	20

Table 7: The last accuracies

In table 7, we see the average accuracies from all operations. The best accuracy comes with bag of visual words without spatial tiling. But I think, we can come up with more accuracy if we could have run 500 as cluster number. And also to get

more accuracy, we can arrange the dataset because there are several things that make the prediction wrong.

(a) Butterfly (b) Iris

Figure 14: 2 images from 2 different categories

For example in figure 14, we see 2 images from 2 different categories. But in the first image whose category is butterfly, the butterfly is on an iris and the size of the butterfly is so small. The models can predict the butterfly's category as iris. As a future work, we can do object recognition to improve the accuracies. And maybe we can use some image processing techniques that are compatible with our problem.

5 Bonus

To see the affect of distance function, I used L1 distance as another distance metric. (see eq. 3)

$$distance = \sum_{i=1}^{k} \|x_i - y_i\| \tag{3}$$

I made this just for Gabor Filter Bank and SIFT. And the results as follows:

The method	Euclidean	Manhattan(L1 Distance)
Gabor Filter Bank	20	20
SIFT	22	20

Table 8: The average accuracies using different distance equations

In table 8, we see that the accuracy of gabor filter bank is the same for both distance matrix. And their class based accuracies are also same but on the other hand, the accuracy of the SIFT decrease when we use Manhattan distance. Although the calculation of the Manhattan distance is easier, the euclidean distance is better. Because the euclidean distance is rotation invariant but the Manhattan is not.