1 Introduction

In this project, we explore image classification using Bag-of-Words (BoW). We build a BoW model and evaluate its performance with the mean average precision and mean accuracy metrics. Furthermore, we do experiments on different settings such as different number of vocabularies/clusters, different color spaces and different descriptors and discuss the results of different settings.

2 Image Classification using Bag-of-Words

The Bag-of-Words model is an effective method for image classification. Each image is represented by frequencies of visual words which corresponds to feature descriptors. This frequency representation can then be used with classification algorithms in order to make prediction.

To train our model we use the STL-10 dataset [2]. There are a total of ten classes in this dataset, but we will only focus on the following five classes: *airplanes, birds, ships, horses, cars*. The dataset includes 500 training and 800 testing images for each class. Besides that, each image is of size $96 \times 96 \times 3$. We train our model in several different settings, which are different color spaces, different vocabulary sizes and different descriptors.

2.1 Feature extraction and description

Extracting features is a vital process in image classification. In this project we use three different descriptors: scale-invariant feature transform (SIFT), speed up robust features (SURF) and histogram of oriented gradients (HOG). SIFT is an algorithm to detect and describe local features in images. It finds extreme points in the spatial space and extracts translation, scale, and rotation invariants [1]. For this descriptor, we use the built-in function `cv2.xfeatures2d.SIFT_create()` from OpenCV. This function computes descriptors at points of interests in images, such as edges, corners, blobs and so on. In this way the image can be described by characteristic points.

SURF is similar to SIFT with the main difference being that SIFT uses Gaussian smoothing, while SURF uses square-shaped filters which makes the process faster. For SURF, we also use the built-in function `cv2.xfeatures2d.SURF_create()` from OpenCV.

HOG divides the image into small connected areas, which are called cell units. It collects the gradient or edge direction histogram of each pixel in the cell unit. Finally, these histograms can be combined to form a feature descriptor. In this case we use the HOG-function from *skimage*.

In training dataset each class has 500 images. In order to build up visual vocabulary we use 200 images in each class. Furthermore, in order to examine BoW model we also use different color space: gray scale and RGB. To compare with these two color space we use SIFT descriptors.

Figure 1 shows plots of key points for five classes. The colorful point in each plots represent key points detected by SIFT and the circles represent orientation of key points.

(a) Plot of key points in car image based on SIFT. (b) Plot of key points in airplane image based on SIFT . (c) Plot of key points in bird image based on SIFT . (d) Plot of key points in ship image based on SIFT . (e) Plot of key points in horse image based on SIFT .

Figure 1: Plots of key points for 5 classes.(a-d) are images of car,airplane, bird, ship and horse class

2.2 Building Visual Vocabulary

After extracting features from subset we build up visual vocabulary in this part. We perform **kmeans** on extracted descriptors. In order to run **kmeans** we need to decide the number of clusters to divide extracted descriptors. We use three different clusters number: 400, 1000 and 4000. Each cluster is considered as a 'word' in visual vocabulary.

2.3 Encoding Features Using Visual Vocabulary

Since we have visual vocabulary the next step is representing the rest of images into collection of visual vocabulary. Firstly, we extract key points and descriptors from 1500 images. For each image we assign each descriptor to the closest 'word' in given visual vocabulary.

2.4 Representing images by frequencies of visual words

In this step we quantify the process of encoding features. We calculate frequency of each 'word' in each image. In this way, each image can be represented by histogram of its visual words. Besides, we need to normalize the histogram because each image has different number of features. And these histograms are used to train classifiers.

Since different images have different number of features we normalize the histogram by number of total features in corresponding class. Then the Y axis in each histogram can be represented as the probability of features that are included by these clusters in each class. We use 300 images for each class and add them up to show the histogram.

Figure 2 shows histogram plots for 5 classes and we normalize histogram because different image has different number of features. The size of histogram is 400. It is clear to see that histogram of bird is significantly different from other classes. Histogram of bird is sparser and higher than other classes. The possible explanation is that when building histogram there are a fewer amount features extracted from bird class than other classes, which will makes prediction of bird classifier worse than other classes' classifiers. Furthermore, we see that histogram of car is quite similar to airplane. Because when we look carefully both histogram of car and ship show high probability around same cluster position for example, cluster 290, 370 and so on. It is intuitive because when we check out the images of car and ship we find out most images share similar pattern, for example long structure of objects -cabin and car body. Moreover, histogram of horse shows that histogram of horse is scattered over all 400 clusters in histogram and each clusters have near the same probability. The possible explanation is that the extracted features in clusters are quite general and in this way each images of horse can share nearly the same number of features in each clusters.

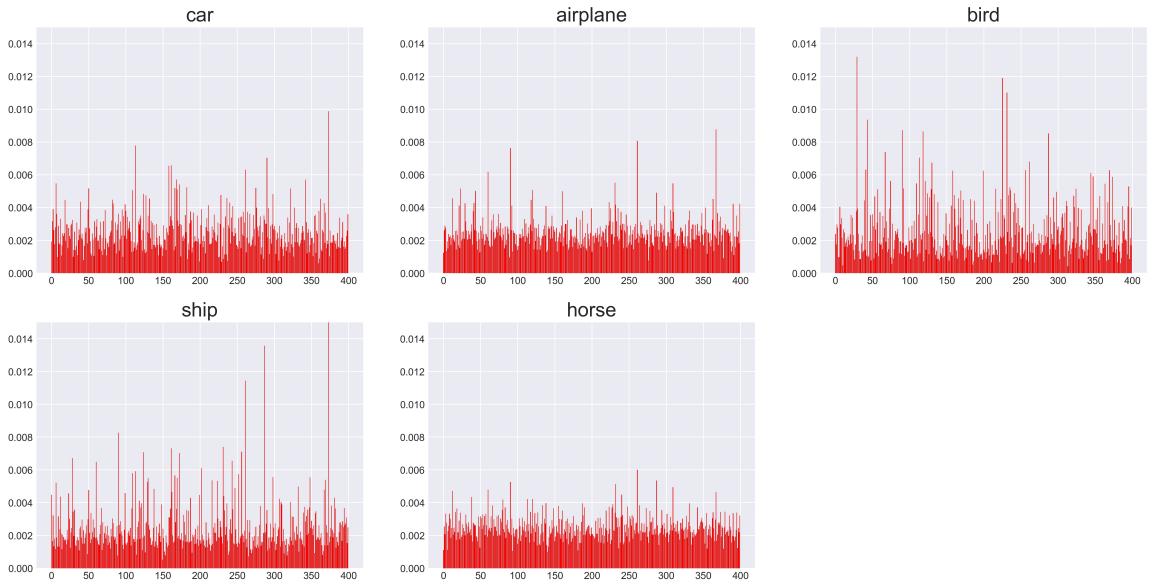


Figure 2: Histogram of 5 images from 5 class in training dataset. (a-e) are histograms from car, airplane, gird, ship and horse classes. And the size of histogram is 400.

2.5 Classification

In this part we use Support Vector Machine(SVM) from *sklearn* to train classifiers. We train 5 classifiers for each class and each class has rest of 300 images. In this way each classifier is binary prediction which means target class is labeled as *1* and others are labels as *0*. Furthermore, SVM is sensitive to the scale of input and we need to standardize our input features. This process is achieved by built-in function *sklearn.preprocessing.StandardScaler*. After standardization the input features' mean is 0 and variance is 1.

2.6 Evaluation

After training 5 classifiers we use total 5000 test images to evaluate them. We use same process to extract features and make histograms. We compare the results for three different setting. First setting is different number of clusters, which are 400, 1000 and 4000 based on gray scale SIFT. The second setting is different color space, which are gray scale and RGB based on SIFT and 400 clusters. The last one is different descriptors, which are SURF and HOG based on gray scale images and 400 clusters.

Besides, for each setting, we use Average Precision(AP), Accuracy for a single class and also Mean Average Precision(MAP), Mean Accuracy over all classes and accuracy to evaluate performance of classifier. Moreover, we also plot the top-5 and the bottom-5 ranked test images for each classifier.

3 Result

3.1 vocabulary size

Firstly, we evaluate the results with different vocabulary size. From Table 1 and Table 1 it is clear to see that 4000 vocabulary size shows the worst performance. The possible explanation is that 4000 vocabulary size includes many clusters that are not general enough, which make histograms sparse. In this way it is difficult for SVM to classify. Results of 400 and 1000 vocabulary size show small difference. 400 vocabulary size has larger MA, while 1000 vocabulary size has larger MAP. Thus, we can conclude that when vocabulary size is large than 400 the cluster in visual vocabulary become less common, which shows small influence on classification. Thus, we use 400 vocabulary size for the rest of two setup.

Furthermore, we find out for three different vocabulary sizes all prediction of bird shows the worst performance among five classes. The explanation is that SIFT can not extract the general features from bird class, which means histogram can not successfully represent the images from bird class . Figure **??** also verifies our explanation. Because histogram of bird is sparser than other four classes.

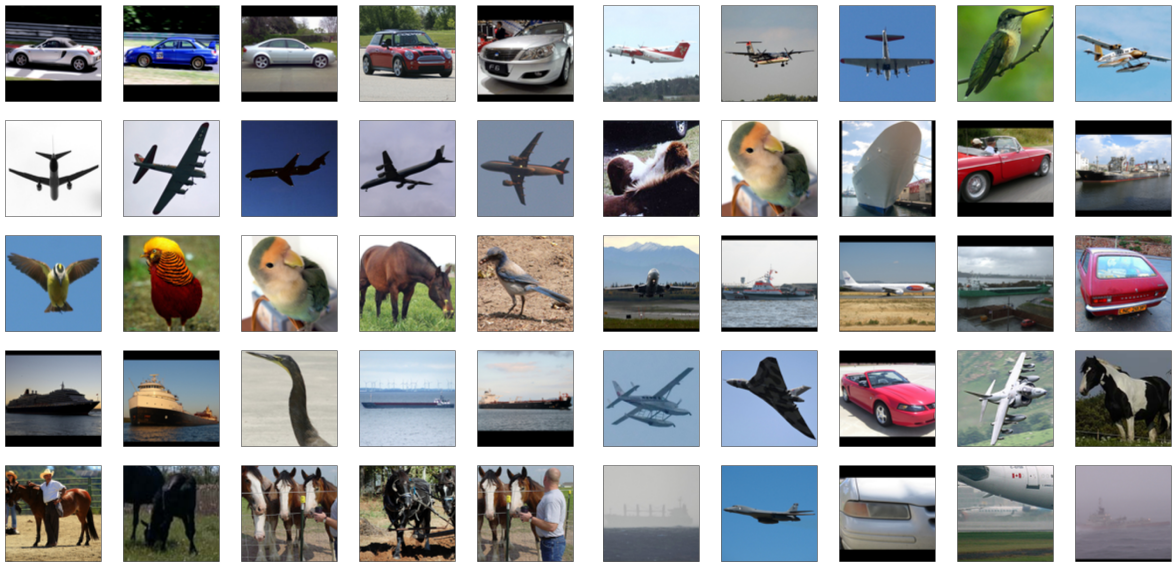
Figure 4 shows results of five classifiers with 1000 vocabulary size. The first row is result of car classifier and the rest are airplane, bird, ship and horse classifiers. Figure 4(a) shows top-5 ranked images and we can see that most classifiers successfully predict the target class. But we can still observe that there are false positive in bird classifier for all three vocabulary size, which verifies the result of MAP and MA of bird class . It is because the feature of chicken is quite similar to bird's. Figure 4(b) shows bottom-5 ranked images and the images seems random, which verifies the effectiveness of classifier because there are all not images of target class. Figure 3 and 5 shows result of 400 and 4000 vocabulary size. We can see that there are more wrong prediction in 4000 vocabulary size. Result of 400 vocabulary size is close to 1000 vocabulary size.

Vocabulary size	MAP	AP Car	AP Airplane	AP Bird	AP Ship	AP Horse
400	0.5840	0.6416	0.6169	0.4664	0.5836	0.6119
1000	**0.5824**	0.6347	0.6319	0.4574	0.5931	0.5953
4000	0.5523	0.6063	0.6182	0.4417	0.5937	0.5018

Table 1: MAP and AP results of different vocabulary sizes based on gray scale SIFT. Vocabulary size are 400,1000 and 4000.

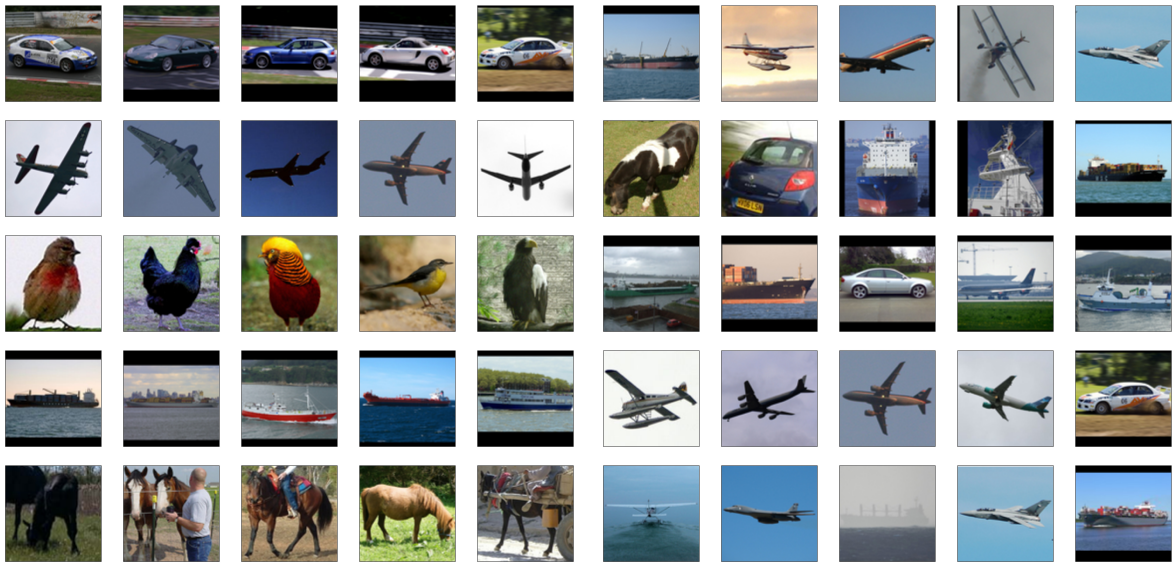
Vocabulary size	MA	Car	Airplane	Bird	Ship	Horse
400	**0.8234**	0.8230	0.8437	0.8005	0.8215	0.8247
1000	0.8146	0.8073	0.8430	0.8000	0.8160	0.8065
4000	0.8049	0.8000	0.8247	0.8000	0.8000	0.8000

Table 2: MA and Accuracy results of different vocabulary sizes based on gray scale SIFT. Vocabulary size are 400,1000 and 4000.



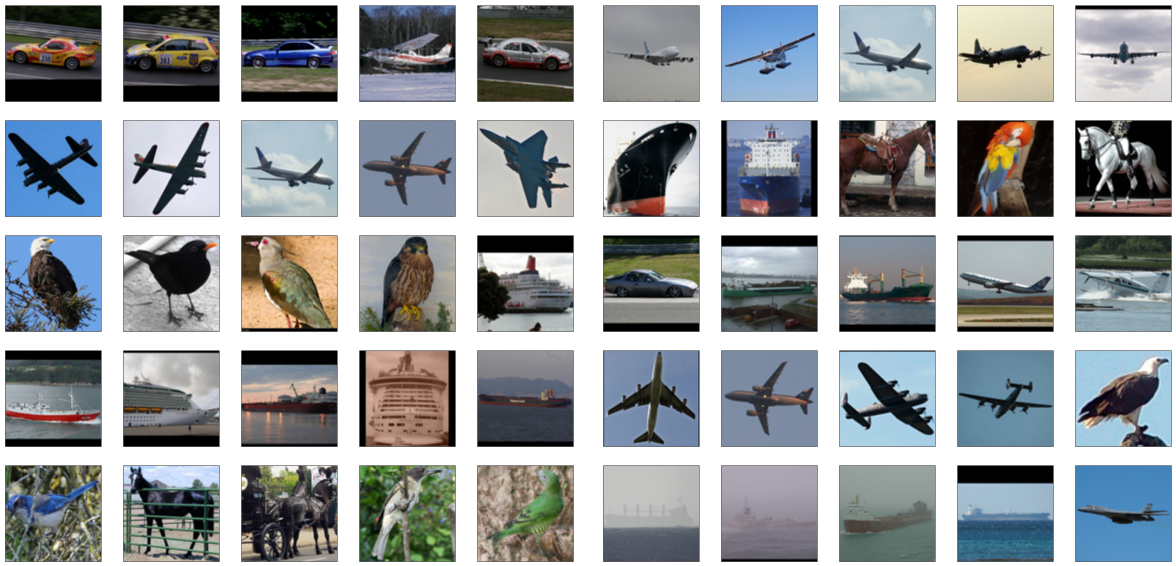
(a) Top-5 ranked images for each class with 400 vocabulary size (b) Bottom-5 ranked images for each class with 400 vocabulary size

Figure 3: Top-5 and bottom-5 ranked images for each classes with 400 vocabulary size. (a) is top-5 ranked images. (b) is bottom-5 ranked images. The first row images are result of car class and the rest are airplane, bird, ship and horse classes.



(a) Top-5 ranked images for each class with 1000 vocabulary size (b) Bottom-5 ranked images for each class with 1000 vocabulary size

Figure 4: Top-5 and bottom-5 ranked images for each classes with 1000 vocabulary size. (a) is top-5 ranked images. (b) is bottom-5 ranked images. The first row images are result of car class and the rest are airplane, bird, ship and horse classes.



(a) Top-5 ranked images for each class with 4000 vocabulary size (b) Bottom-5 ranked images for each class with 4000 vocabulary size

Figure 5: Top-5 and bottom-5 ranked images for each classes with 4000 vocabulary size. (a) is top-5 ranked images. (b) is bottom-5 ranked images. The first row images are result of car class and the rest are airplane, bird, ship and horse classes.

3.2 Color space

Additionally, we evaluate the result of different color space. Since SIFT can only process gray scale images we deal with three channels separately. Firstly, we use gray scale images to extract the key points and then we use these key points to extract descriptors from three channels. Then we stack these descriptors into one list.

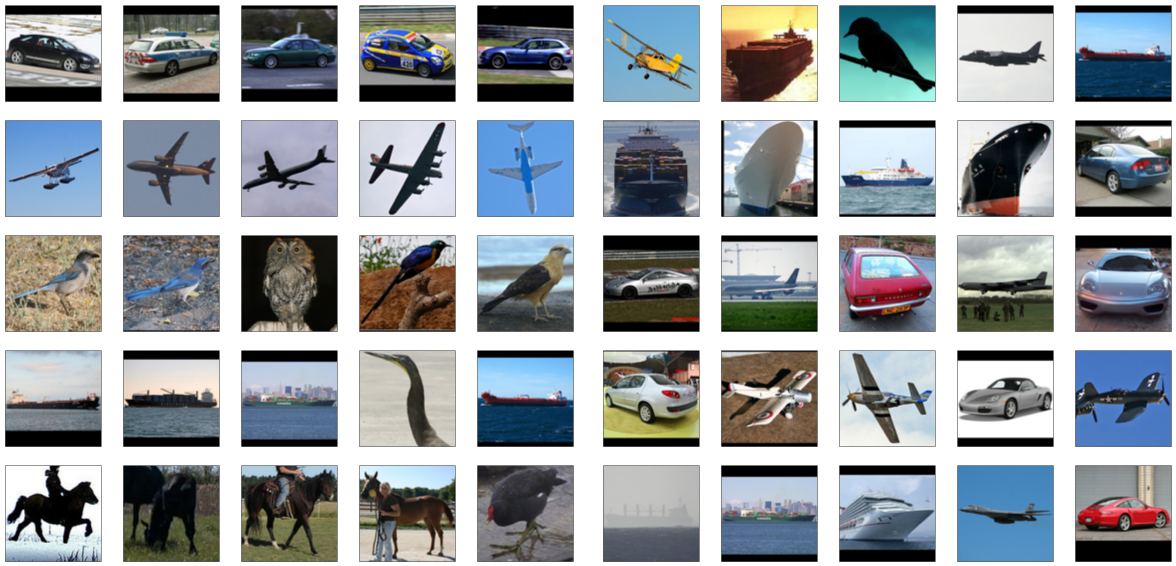
From Table 3 and Table 4 it is clear to see that the result of RGB is slightly better than result of gary scale. Figure 6(a) shows that most classifiers yield true positive for top-5 ranked images. It is intuitive that RGB color space can provide more accurate information when detecting certain object, for example, blue sky of airplane, blue sea of ship or green grass of horse.

Color space	MAP	AP Car	AP Airplane	AP Bird	AP Ship	AP Horse
Gray scale	0.5819	0.6399	0.6233	0.4544	0.5713	0.6200
RGB	**0.6013**	0.6468	0.6401	0.4980	0.5992	0.6226

Table 3: MAP and AP results of different color space based on SIFT and 400 vocabulary size. Color space are gray scale and RGB.

Color space	MA	Car	Airplane	Bird	Ship	Horse
Gray space	0.8234	0.8263	0.8423	0.8003	0.8220	0.8280
RGB	**0.8266**	0.8267	0.8507	0.8007	0.8305	0.8245

Table 4: MA and Accuracy results of different color space based on SIFT and 400 vocabulary size . Color space are gray scale and RGB.



(a) Top-5 ranked images for each class of RGB

(b) Bottom-5 ranked images for each class of RGB

Figure 6: Top-5 and bottom-5 ranked images for each classes of RGB. (a) is top-5 ranked images. (b) is bottom-5 ranked images. The first row images are result of car classifier and the rest are airplane, bird, ship and horse classifiers.

3.3 Descriptors

Finally, we discuss the results of different descriptors. From Table 5 and 6 show that though the SURF's MA is better than SIFT, the MAP is worse than SIFT especially for bird class. From Figure 7(a) all top-5 ranked images of bird classifier are all false positive. That's because according to SURF [3] it is faster than SIFT but less robust. Because SURF use square-shaped filters as an approximation of Gaussian filters, which may make detection of features less accurate than SIFT.

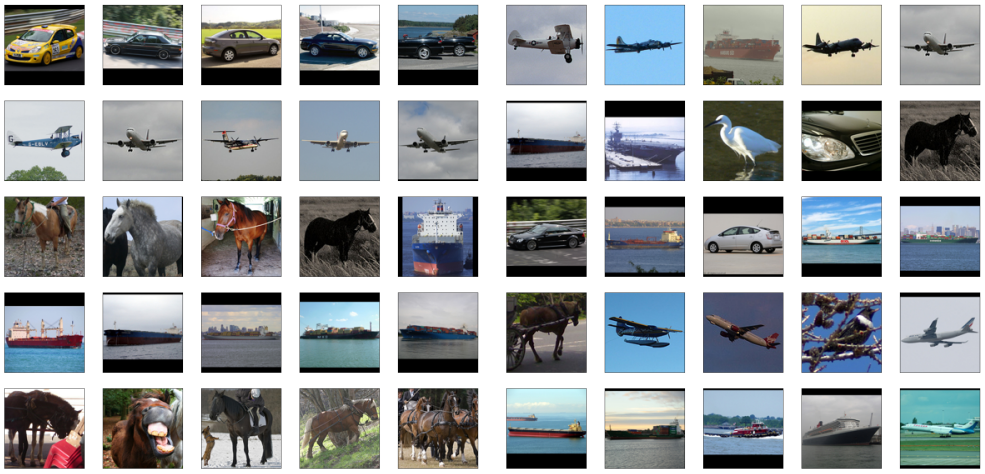
HOG shows the best performance among three descriptors. Table 5 and 6 show that HOG has highest MAP and MA and especially for bird class there is a huge improvement. And Figure 8(a) shows that it is remarkable to see that all classes yield true positive. The possible explanation is that HOG represent images in a denser form than SIFT because HOG stack images' features into one dimension vector. In this way each image can be represented in a more general way in histogram than SIFT.

Descriptors	MAP	AP Car	AP Airplane	AP Bird	AP Ship	AP Horse
SIFT	0.5819	0.6399	0.6233	0.4544	0.5713	0.6200
SURF	0.5696	0.5796	0.6803	0.4301	0.5528	0.6052
HOG	**0.6778**	0.6875	0.6092	0.6902	0.6251	0.7771

Table 5: MAP and AP results of different descriptors based on gray scale and 400 vocabulary size. Descriptors are SIFT, SURF and HOG.

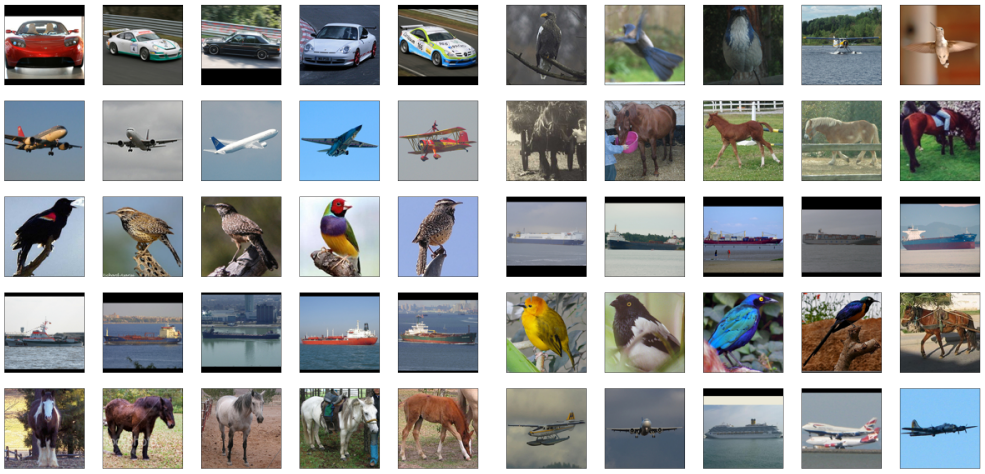
Descriptors	MA	Car	Airplane	Bird	Ship	Horse
SIFT	0.8234	0.8263	0.8423	0.8003	0.8220	0.8280
SURF	0.8261	0.8285	0.8583	0.8013	0.8238	0.8188
HOG	**0.8495**	0.8373	0.8388	0.8505	0.844	0.8770

Table 6: MA and Accuracy results of different descriptors based on gray scale and 400 vocabulary size. Descriptors are SIFT, SURF and HOG.



(a) Top-5 ranked images for each class of SURF (b) Bottom-5 ranked images for each class of SURF

Figure 7: Top-5 and bottom-5 ranked images for each classes of SURF. (a) is top-5 ranked images. (b) is bottom-5 ranked images. The first row images are result of car class and the rest are airplane, bird, ship and horse classes.



(a) Top-5 ranked images for each class of HOG (b) Bottom-5 ranked images for each class of HOG

Figure 8: Top-5 and bottom-5 ranked images for each classes of HOG. (a) is top-5 ranked images. (b) is bottom-5 ranked images. The first row images are result of car class and the rest are airplane, bird, ship and horse classes.

4 Conclusion

In this project we follow the steps to build BoW model. And we test different setting such as different color space, different descriptors and different vocabulary size. Firstly, we find out RGB color space outperforms gray scale with SIFT descriptors since RGB images can provide more information. Secondly, small vocabulary size has higher MAP and MA because the features in histogram are more general. Finally, we find out that HOG is the best descriptor because it provide denser and more general features for training classifiers.

References

[1] Sift. `https://en.wikipedia.org/wiki/Scale-invariant_feature_transform`.

[2] Stldataset. `http://ai.stanford.edu/~acoates/stl10/stl10_binary.tar.gz`.

[3] Surf. `https://stackoverflow.com/questions/11172408/surf-vs-sift-is-surf-really-faster`.

5 Appendix