**Computer Vision and Image Processing: Homework 1**

**Spatial Pyramid Matching for Scene Classification**

**Name**: Manpreet Dhanjal

**UBIT Name**: DHANJAL

**Person Number**: 50248990

**Introduction**:

In this project we have developed a scene recognition system using Bag of Words approach. In order to create Bag of Words, few filters are applied on the image in Lab color format which gives us the image features like details about the color intensity and edges in the image. Finally features are extracted from all the training images and visual words (clusters) are prepared using kmeans clustering algorithm.

For the features matching, a word map which is a representation of the number of words found in the image, is calculated for all the training images and the histogram is generated for it. The histograms are generated using Spatial Pyramid so that the image is not misclassified because of some mixed scene structure. Finally, histogram matching of the test image with all the training images results in the recognition of the scene.

**Analysis & Results:**

**1.1** The filters provided for the assignment can be broadly classified into 4 categories. All these categories have filters of varying sizes. These categories are:

1. **Gaussian** – This filter, in general, removes noise from the image or in other words, smoothens the image. Since we are applying the filter in L a b color space, it does the work of edge preservation. This filter compresses the minute details in the image by smoothening the small gaps and puts more focus on the edges of the bigger objects in the image.

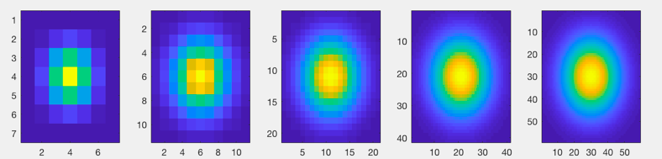


Figure: Gaussian Filters

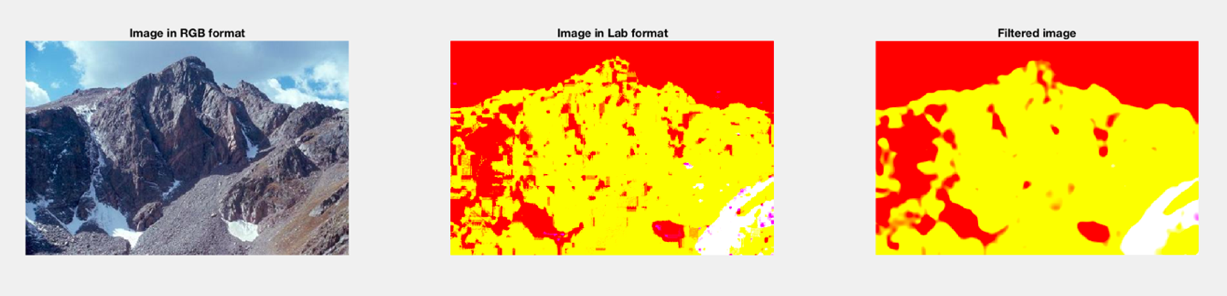


Figure: Effect of Gaussian Filter on image

1. **LoG (Laplacian of Gaussian)** – This filter, detects all the edges (which represents sharp changes in intensity) in the image. As shown in the image below, this kind of filter returns black pixel where the intensity is uniform and sharp edges where there is a change in intensity in the image.

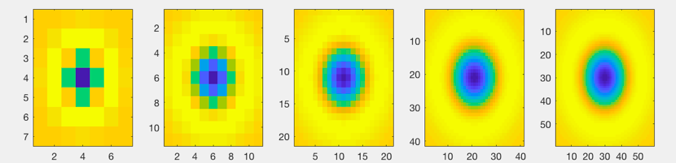


Figure: LoG Filters

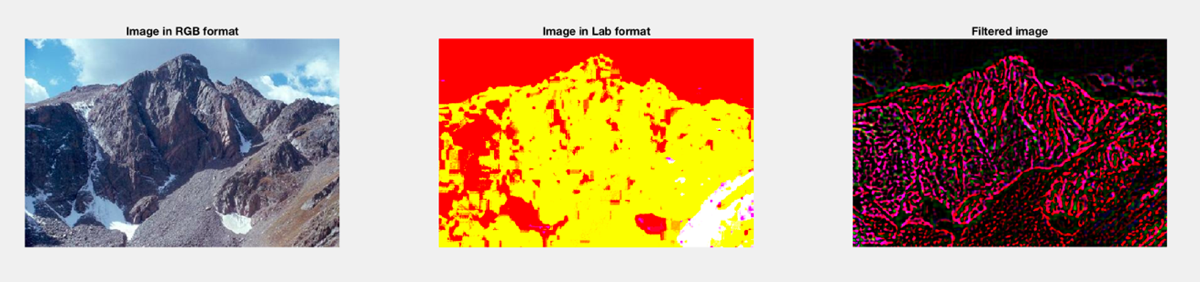


Figure: Effect of LoG Filter on image

1. **Horizontal Edge** – These set of filters detect horizontal edges in the image. This filter returns zero pixel value when the intensity is uniform top to down and detects an edge when strong intensity change is identified in the vertical direction.

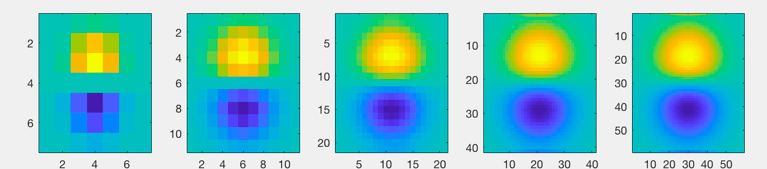


Figure: Horizontal Edge Filter

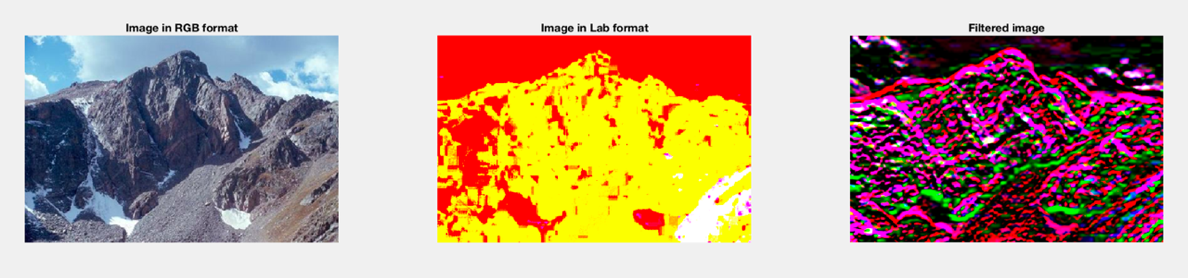


Figure: Effect of Horizontal Edge Filter

1. **Vertical Edge** – These set of filters work similar to the horizontal edge filters but they detect the vertical edges in the image. These filters return zero pixel value when the intensity is uniform in horizontal direction and detects an edge when strong intensity change is found in the aforementioned direction.

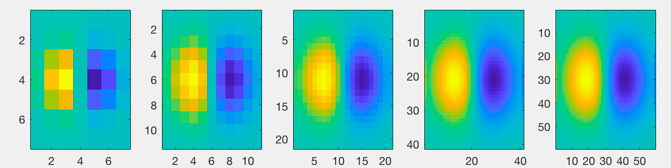


Figure: Vertical Edge Filter

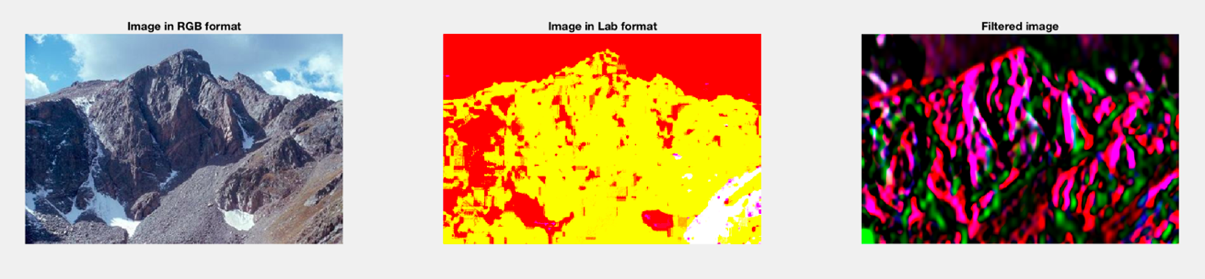


Figure: Effect of Vertical Edge Filter

**1.1** Following are the collages of the filter responses of 20 images from the training data set.



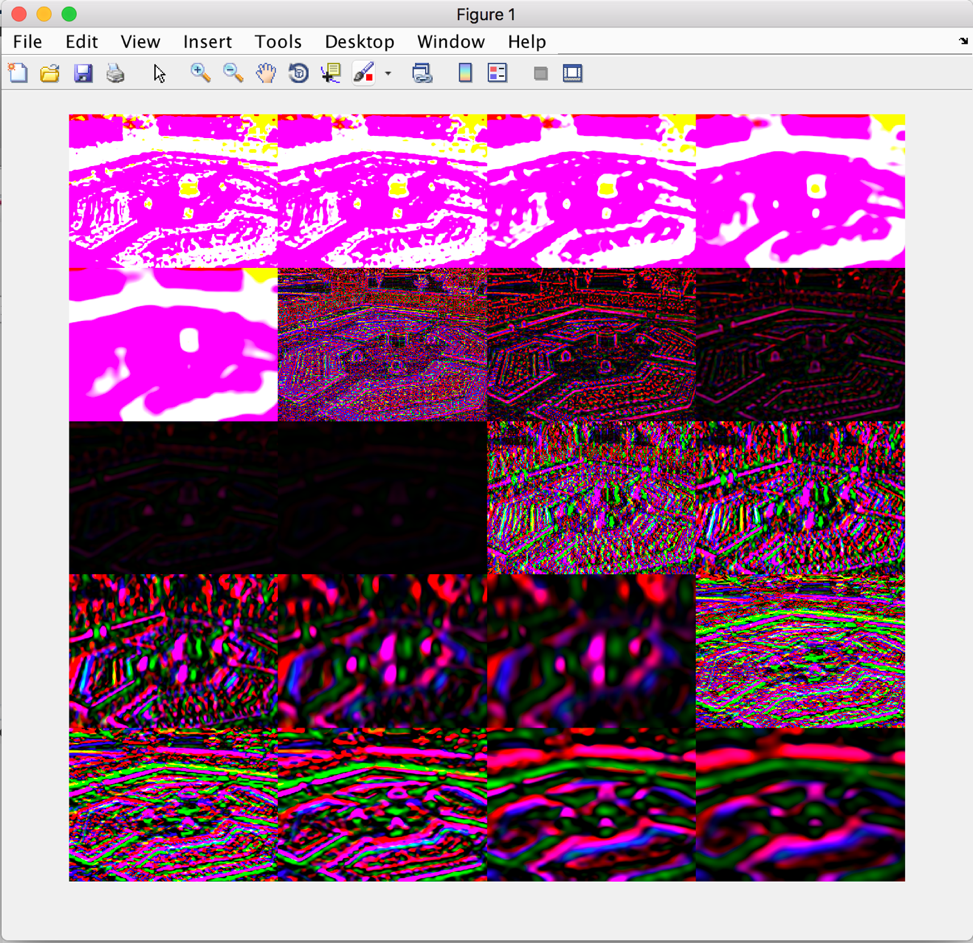


Figure: Collage

**1.3** Following are the images with their respective word maps from the computer room category:

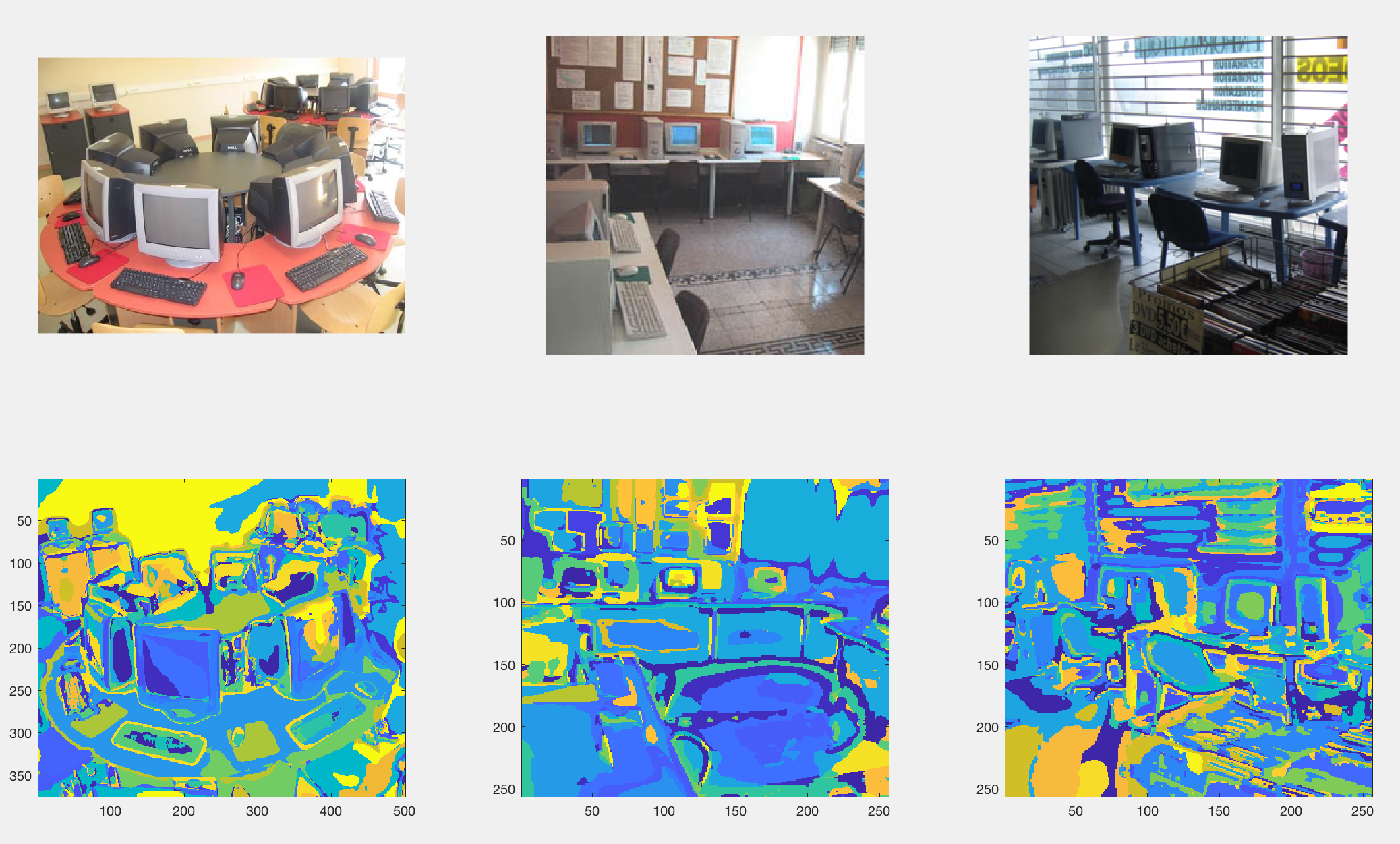
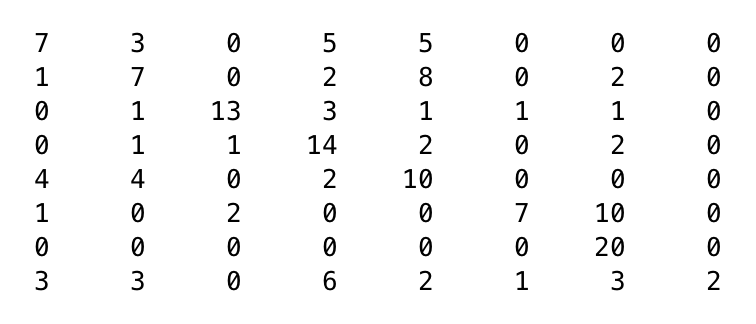


Figure: Word Map of 3 image from Computer Room

A word map basically is a 2D matrix, with the same dimensions as the source image, in which the similar pixels have the same value. To calculate the word map we have to find the distance between each pixel of the image with every word in the dictionary and assign the index of the word with the minimum distance to the word map. When viewed with the ‘imagesc’ command, it results in an image where all the pixels belonging to the same word category (that is, pixels having similar features) have the same color.

Here, in the first word map of the figure shown above, we can see that the whole surface of the table has been classified into one word (blue color), the keyboards have been classified into one word (greenish blue color) and the wall at the back of the room has been assigned a single word (yellow color).

**2.5** Confusion Matrix



After evaluating the recognition system, the confusion matrix is as mentioned above. The accuracy of the system comes to be 50%.

**2.6**

Looking at the confusion matrix, the most misclassified class is the ‘Tennis Court’ and the least misclassified class is the ‘Ocean’. Also computer room was also wrongly classified as library many times.

**Following categories were the most misclassified in our data set:**

**Tennis Court**: The tennis court has been misclassified into almost all of the other categories, this is because there are too many variables in the tennis court which matches with other scenes. It has a big hall kind of structure hence it could be classified as art gallery or ice skating. From the confusion matrix we see that tennis court was misclassified as ice skating 6 times, this could be attributed to the similarity in both the scenes like people holding sticks, boundary markings on the ground, flags of various teams, sitting area for the audience etc.

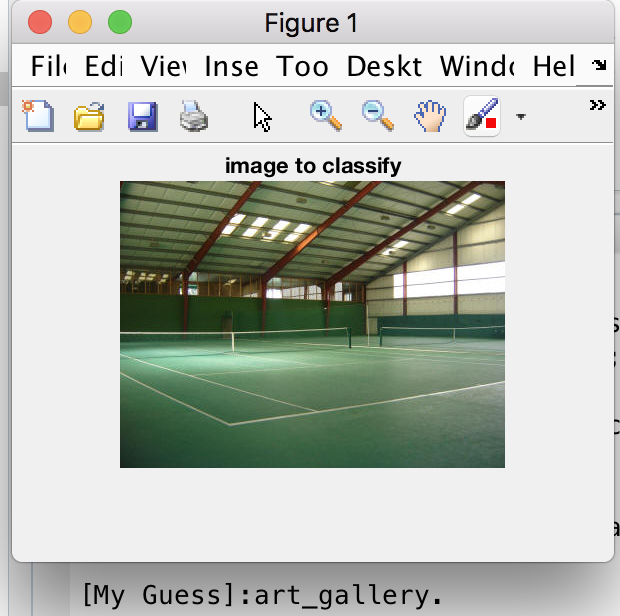


Figure: Tennis Court recognized as Art Gallery

**Art Gallery, Computer room, Library**: These scene categories also have similar kind of features like windows, chairs, tables etc. Books can also be recognized as a painting by our recognition system. Hence, the chances of misclassification are more within these 3 categories.

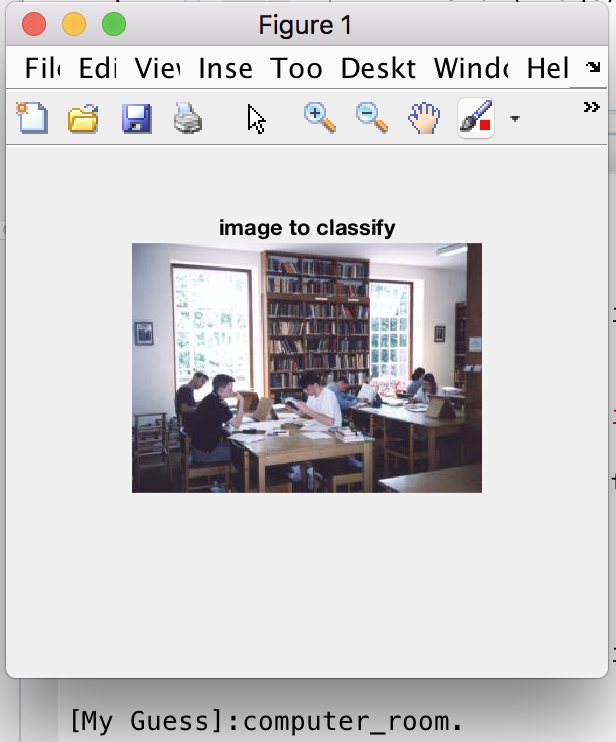
 

Figure: Library recognised Figure: Mountain recognised

as Computer Room as Ocean

**Mountain**: This category has been misclassified 10 times into the Ocean category. This could be attributed to the similarity between these scenes. Both of the scenes can be divided into 2 big structures, water and sky in the case of ocean and mountain and sky in the case of mountain and hence, the filter response of both of the images would be similar.

**Overall, following scenes are difficult to classify:**

Thus, its difficult to classify the following with the ‘bags of words’ approach:

**Indoor scenes**: In most of the indoor scene, many words will be common to all. For instance, a chair and table can be seen in library and computer room. Similarly, the big room structure is common to the tennis court and the art gallery and hence the chances of misclassification is more.

**Scenes with high variability**: If in an image of the computer room, the computers are in focus, then it can be easily classified. However, chairs and doors are the main component of that image, it might be misclassified as library. So, computer room has a large number of words associated with it which are common to other indoor scenes as well.

**Partial images**: If the image to be classified is just a part of the whole scene, it has more chances of getting misclassified. For instance, an image with flags of the tennis court might be misclassified as ice skating since the flags are not the main component of the scene of a tennis court.

**Mixed scenes**: A single image might have components from the garden as well as components form mountain scene. Hence, it will have words from both of the scenes and it will be difficult for the recognition system to recognize the image.

**2.7**

**Improving Performance:**

The recognition system submitted by me has 200 words (number of clusters) and 150 alpha value. This configuration results in 50% accuracy. But the accuracy can still be improved. Our recognition system works on four main factors that we can change in order to improve the accuracy of the system:

1. **Filters:** Currently, we are using just 4 types of filters. In order to fetch more features, we can add few more filters like the difference of Gaussian filters and filters which responds on the color in the image.
2. **Alpha:** We are picking up alpha random pixels from the training image. These pixels represent the features that were extracted from the image. Picking low values of alpha might lead to loss in important features in the image. So, increasing the value of alpha would lead to improved accuracy.
3. **Clusters (k):** The number of clusters used in the kmeans method define the number of words in our dictionary. Improving the number of clusters, improves the classification quality of the recognition system and results in improved accuracy.
4. **Histogram matching:** In order to find similarity between images, we are finding the intersection between the histograms of training and test image and classifying them. There are few more techniques that could be used instead of intersection of histograms, like merged histogram matching or matching just the high frequency values in the histogram.

According to the steps given in the assignment, we have to calculate the filter responses of all the pixels in an image and then take out alpha number of random pixels from it. This takes a lot of time. We can reduce this time by first taking out alpha random pixels from the image and then taking the filter responses of these pixels.

Apart from all these, increasing the size of training image is bound to increase the accuracy of the system.

**Conclusion**:

According to my analysis the ‘Spatial Pyramid Matching’ and ‘bags-of-words’ approach will be efficient in the following cases:

1. The recognition system is trained with all the possible images of a particular scene. For this we need to use a very large number of images for all the scene categories.
2. The images used for training and testing should not focus on any particular object which doesn’t contribute much value to the scene. For example, an image of tennis court will be misclassified if the focus is on flags rather than the court.
3. The categories of the natural scenes are different from each other. For example, a mountain is misclassified as ocean many times, this is because both the scenes are similar in terms of intensity values.
4. It works well for outdoor scenes. A lot of similarity could be found in different indoor scenes, therefore, the chances of misclassification is more in that. For example, a computer room is wrongly classified as library since they have similar components like chairs, tables, windows etc.

Improving the factors mentioned in section 2.7 would surely lead to improved performances in the system.