RDFIA - Homework 1

Shubhamkumar Patel 21113255 - Johan Pardo 21110842

October 2021

1 Practical Work 1-a and 1-b

Question 1: Show that kernels M_x and M_y are separable, i.e. that they can be written $M_x = h_y h_x^t$ and $M_y = h_x h_y^t$ with $h_x =$ two vectors of size 3 to determine

With the following,

$$h_x = \frac{1}{2}\begin{pmatrix}-1\\0\\1\end{pmatrix}, h_y = \frac{1}{2}\begin{pmatrix}1\\2\\1\end{pmatrix}$$

M_x and M_y can be written in the form $M_x = h_y h_x^t$ and $M_y = h_x h_y^t$ so M_x and M_y are separable.

Question 2: Why is it useful to separate this convolution kernel

By going from a matrix of size 3x3 (Sobel Filter) to 2 vectors of size 3 we limit the computational work.

Question 3: What is the goal of the weighting by gaussian mask?

The goal of the weighting gaussian mask is to put more weight to pixels in the center than the one in the border. This leads to reduction of noise in the image.

Question 4: Explain the role of the discretization of the directions?

We have a 4x4 size patches and to get the gradient orientation we locally assign to an array of dimension 8, 1 gradient value for each 8 directions of 45°. Using this array we generate a histogram that we can use to get a representation of the gradient of the patch. While doing so we are regrouping directions which also helps reducing direction noises.

The role of the discretization of the 8 directions is to limit the importance of a direction, and therefore making our algorithm more robust to rotation.

Question 5: Justify the interest of using the different post-processing steps.

The post-processing steps are used to clean different artifacts and noise from the image.

First Step : We discard variations that are below a certain threshold (L2 Norm less than 0.5), by simply replacing the values by zeros. This is useful to remove noise.

Second Step : With the normalization step plus the use of a threshold by 0.2, makes our SIFT descriptor invariant to brightness change.

Question 6: Explain the SIFT is a reasonable method to describe a patch of image when doing image analysis

While doing image analysis, we want to be able to deal with all kinds of input. Because the SIFT method is invariant to rotation, brightness, zoom and much more, it can be used to deal with a variety of images, captured in a variety of conditions. Therefore, this method can be useful for image analysis.

Furthermore, the SIFT method is also useful because it generates a “fingerprint” of the image which is reproducible, it is also a memory and computationally efficient method.

Question 7: Interpret the results you got in this section

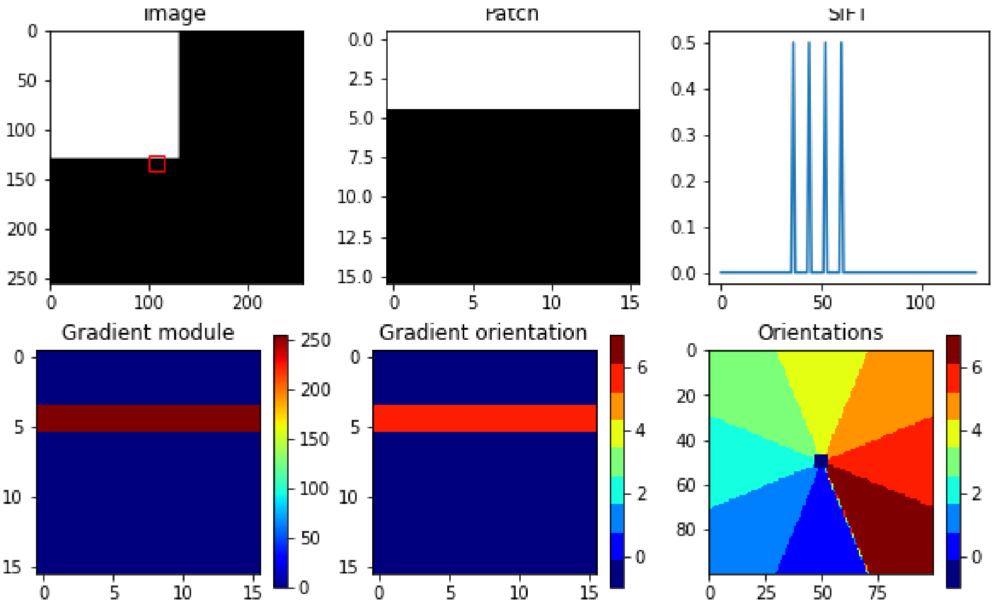


Image: It shows us the position of the patch relative to the image

Patch: It shows the local region that we want to describe.

SIFT: We have a 128 bins histogram representing the gradient values at each pixels of our 16x16 patch. We can see that the gradient is Zero where there is no variance whereas peaks are shown around pixels where the variance in gradient increases.

Gradient Module: It represents the amplitude of local variance of gradient in the patch.

Gradient Orientation: It represents the orientation of each gradient in the patch.

Orientations: It shows us in which direction is our patch going on average.

Question 8: Justify the need of a visual dictionary for our goal of image recognition that we are currently building

Visual dictionary is needed for our goal of image recognition that we are currently building because it let us identify the SIFT's descriptors and compare them with the features of the Visual dictionary.

It is a global way of describing the images and comparing them with the others. At the step before we only had SIFT descriptors to describe our images now we can use the dictionary to represent images using visual words.

Question 9: Considering the points assigned to a cluster c, show that the cluster's center that minimize the dispersion is the barycenter mean of the points x_i
Let $x_{ii} = 1...n$ points assigned to a cluster c and

$$f(c) = \sum_{i=1}^{n} ||x_i - c||_2^2$$

We apply the gradient operator as follows :

$$\nabla f = \sum_{i=1}^{n} 2(x_i - c)$$

Because f is a convex function we have the following property : $\nabla f = 0$, We now have,

$$\nabla f = \sum_{i=1}^{n} 2(x_i - c) = 0$$

After simplifying :

$$\sum_{i=1}^{n} 2(x_i - c) = 0 \Rightarrow \sum_{i=1}^{n} x_i = n * c$$

By futher developing previous result we get the what is the definition of a barycenter:

$$\boxed{c = \frac{1}{n} \sum_{i=1}^{n} x_i}$$

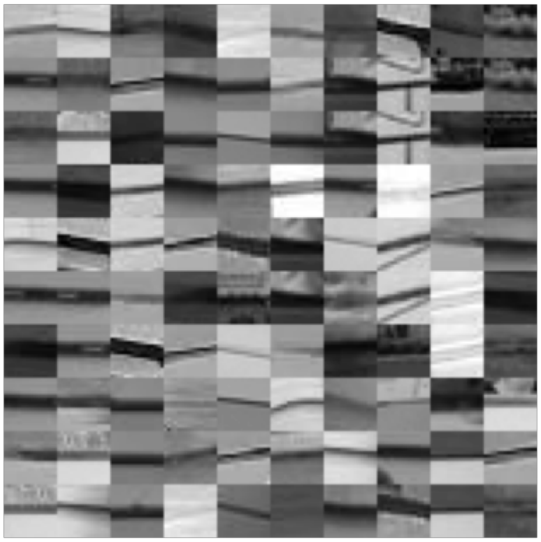
Question 10: In practice, how to choose the optimal number of cluster

In practice, we choose the number of cluster by using the elbow or silhouette methods.

Question 11: Why do we create a visual dictionary from the SIFT's and not directly on the patches of raw image pixels?

We create a visual dictionary from the SIFTs because this way we can see the relation of each patch with each visual words.
Let us recall that we cannot visualize the elements of the dictionnary they are centroids that best represent the categories or classes that we extracted from some dataset of images. There is no way using the raw pixel to compare them with the visual words. That is why we seek the euclidean distance between patches and the visual words to compare them and extract information from those that are closely related to the vectors or centroids, and then only can we visualize them.

Question 12: Comment the results you get

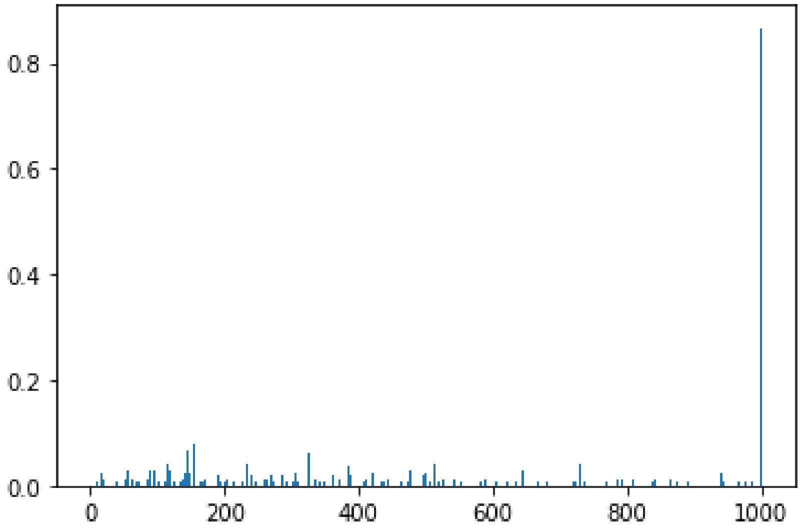


We applied KMeans on a part of our dataset with K = 1001. We then assigned clusters to patches or regions of images by computing the euclidean distance between the centroids and patches. We get the result below. The results that we got are a visualization of the visual dictionary by showing us images who represent the same feature. In this case we can see images which have contain similar structures.

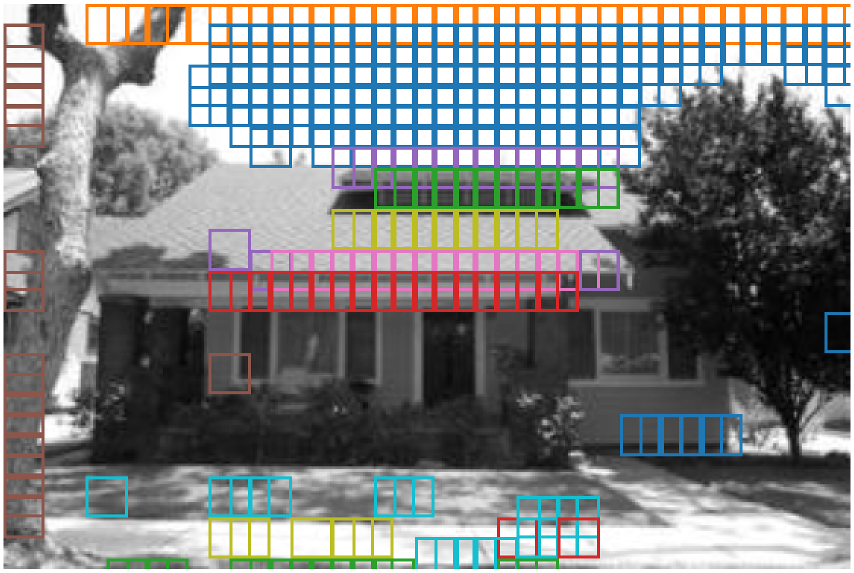
Question 13: Concretely, what does the vector z represent of the image

The vector z represents the number of similarity between the image and the BoVW. It contains the frequency of each clusters (Visual Words) in the patches extracted from an image.

Question 14: Show and discuss the visual results you got



This image represents the frequency of each features present in the image.



We can see the regions of the image that are labelled using different colors (grass, sky, walls and porch of house).

Question 15: What is the interest of the nearest-neighbours encoding? What other pooling could we use (and why)

The interest of the nearest-neighbours encoding is that it's simple to implement and lightweight because 1 patch belongs to only 1 cluster (One-Hot encoding) and therefore limit computation work. Another approach is to use the VLAD (Method similar to Soft Assignment method) which can help us get a distribution of similarity of a patch to all classes.

Using which we can have a better stability and expressiveness (in terms of visual words) to identify features from an image.

Question 16: What the interest of the sum pooling? What other pooling could we use (and why)?

The interest of the sum pooling is to factorize information to limit the number of computation. It also had a tendency to give more advantage to visual words with greater frequency of appearing in an image. There are other type of pooling such as the average pooling which can be better because we still have a trace of other parts.

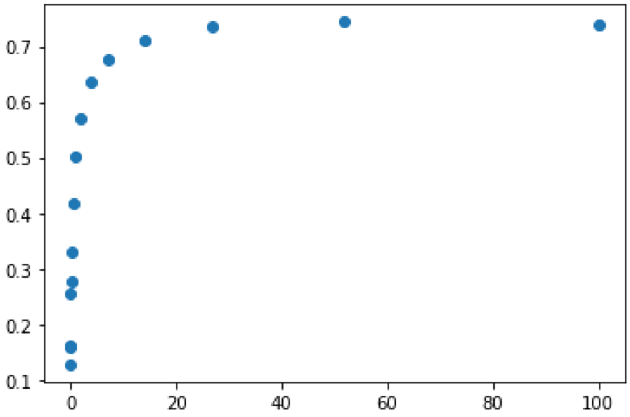
Question 17: What is the interest of the L_2 normalization? What other normalization could we use (and why)?

The interest of the L_2 normalization is to have results in terms of probability (result in form of a unitary vector), so we can compare it to other values of the same type. Another normalization that we could use is by dividing the value by the biggest value. It will not maintain a probability approach, but will increase the range of values and by doing so increase the accuracy.

2 Practical Work 1-c

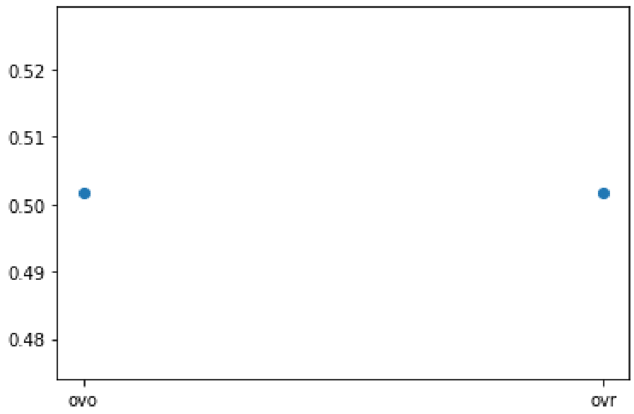
Question 1: Discuss the results, plot for each hyperparameter a graph with the accuracy in the y-axis.

Hyperparameter - C:



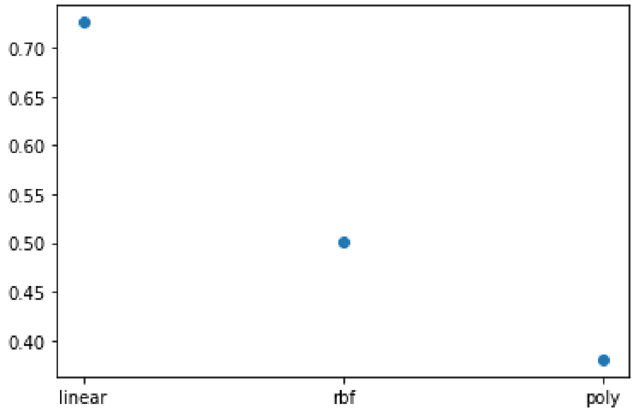
The more we increase C the more we give weight to the loss and by doing so increase the accuracy.

Hyperparameter - Desicion Function:



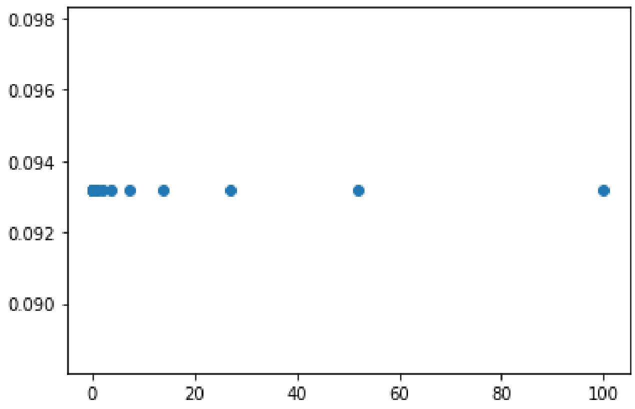
The decision, in this case, is not a huge factor.

Hyperparameter - Kernel:



The linear kernel is better suited for this case than the rbf or polynomial ones.

Hyperparameter - Gamma:



The gamma, in this case, is not a huge factor.

Question 2: Explain the effect of each hyperparameter.

- C: Regularization parameter. The strength of the regularization is inversely proportional to C. Must be strictly positive. The penalty is a squared l2 penalty.

- Kernel: Specifies the kernel type to be used in the algorithm. It must be one of 'linear', 'poly', 'rbf', 'sigmoid', 'precomputed' or a callable. If none is given, 'rbf' will be used. If a callable is given it is used to pre-compute the kernel matrix from data matrices; that matrix should be an array of shape.

- Gamma: Kernel coefficient for 'rbf', 'poly' and 'sigmoid'.

- Decision function: Whether to return a one-vs-rest decision function of shape as all other classifiers, or the original

To get further understant each hyperparameters, the scikit-learn documentation provides us with those details.

Question 3: Why the validation set is needed in addition of the test set ?

The validation set is needed in addition of the test set to check that our program doesn't fall in over/under-fitting