Images descriptors with SIFT and Bag of Words (1-a and 1-b)

Section 1 - SIFT

1.1) Computing the gradient of an image

1. Show that kernels Mx and My are separable, i.e. that they can be written Mx = hyhx.T and My = hxhy.T with hx and hy two vectors of size 3 to determine.

$$Mx = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} \qquad Mx = h_y h_x^T$$

$$My = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix} \qquad My = h_x h_y^T$$

$$hx = \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix} \qquad hy = \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}$$

2. Why is it useful to separate this convolution kernel ?

It is useful to find *hx* and *hy* to **reduce the number of multiplications** performed on each of the image matrix values during convolution. We go from 9 to 6 operations per value by using the hx and hy matrices instead of the *Mx* and *My* matrices: there is **computational interest**.

1.2) Computing the SIFT representation of a patch

3. What is the goal of the weighting by gaussian mask ?

The **Gaussian mask** centered at a point is used to **reduce the noise** around this point by considering its neighbors : the distant points will have less importance.

4. Explain the role of the discretization of the directions

We discretize **very large** direction vectors into encoder vectors of size 128 in order to **reduce the size** of the encoder vectors and to **find the dominant orientation** for a given point by considering the orientations of its neighbors and thus construct **orientation histograms**. The discretization allows it to be **more robust** to the slight change of orientation.

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

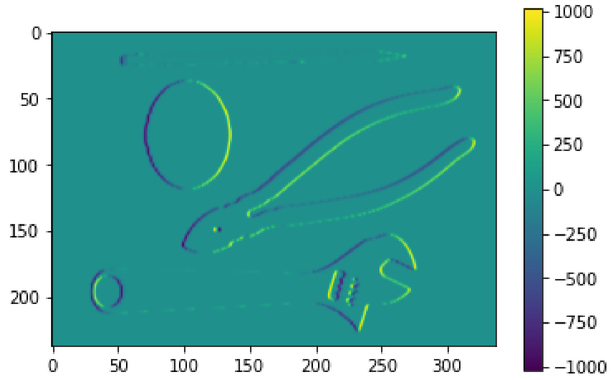
Descriptors whose norm is lower than 0.5 are replaced by null vectors in order not to consider those which are **not interesting**, which have only little contrast. The different post-processing steps will be used to **compare** different feature descriptors between them. The L2 standard works particularly well on float descriptors. Finally, we avoid **loss of information** due to one or several directions being **too dominant** and overwriting other directions that could be interesting.

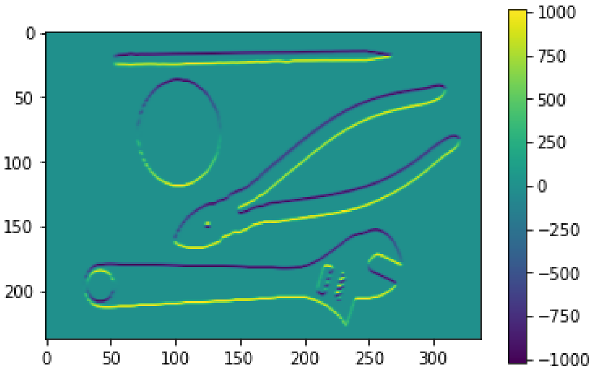
6. Explain why SIFT is a reasonable method to describe a patch of image when doing image analysis

SIFT is a reasonable method to describe a patch of an image because it is a method that can detect corners, circles, similarity between two images, ... while being **robust** to changes in scale, rotation, brightness or viewpoint.

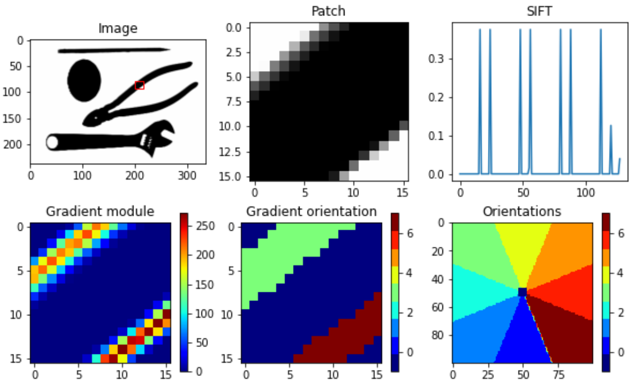
7. Interpret the results you got in this section

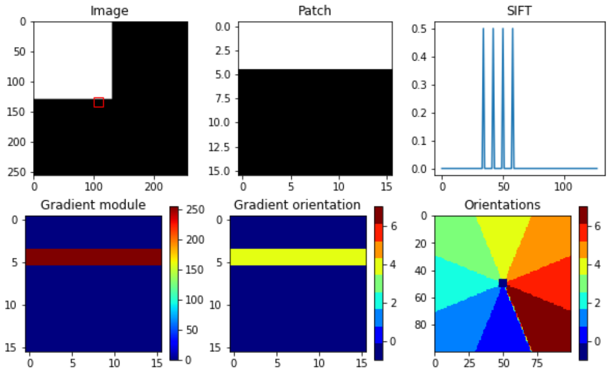
Image gradients :





The first image was obtained from the *Mx* matrix and the second with the *My* matrix. We can observe on the first image that the **horizontal** contours are highlighted, while the **vertical** contours are highlighted on the second image, which is coherent with the **Sobel filters**.





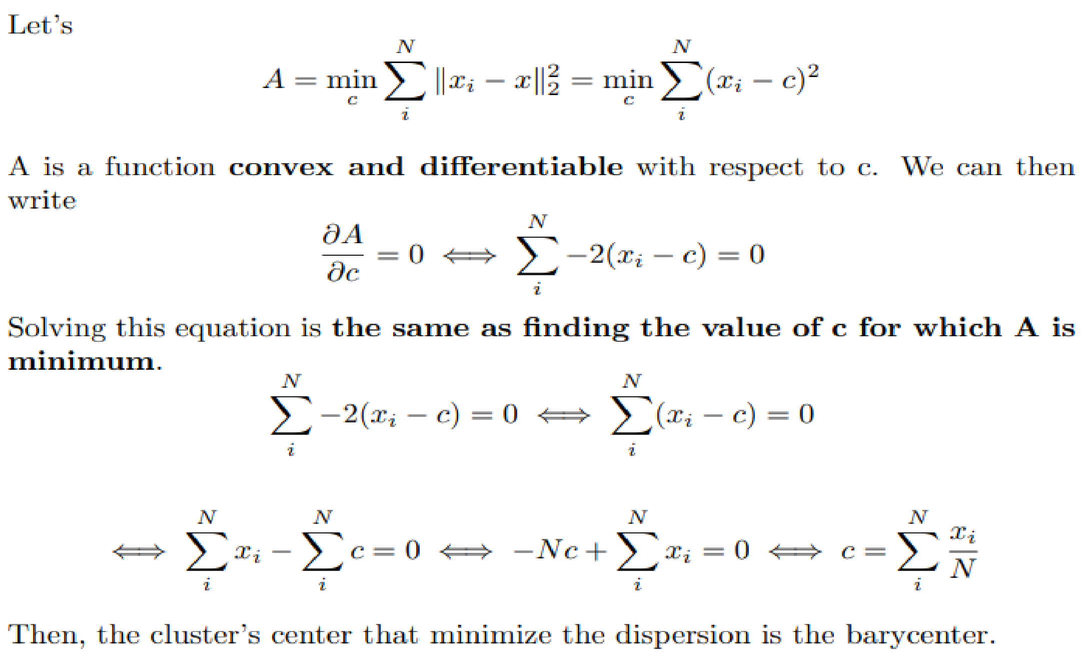
We can observe different images including **histograms of SIFTs** whose peaks correspond to the **dominant orientations**, or **gradient images** that indicate the **variation** in intensity of pixels (the more red it is, the higher the gradient on the gradient image module) and the **orientation** of gradients.

Section 2 - Visual Dictionary

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

The objective of these visual dictionaries is to represent SIFTs as closely as possible in a limited number of "words". **Frequent patterns** are extracted from the images, they allow to create **categories** and thus **reduce the representation space**. It also allows to describe in the same way the whole collection, which allows to standardize the patterns in categories and thus to **be able to compare the images reasonably**.

9. Considering the points {xi}i=1..n assigned to a cluster c, show that the cluster's center that minimize the dispersion is the barycenter (mean) of the points xi



Let's

$$A = \min_c \sum_i^N \|x_i - x\|_2^2 = \min_c \sum_i^N (x_i - c)^2$$

A is a function **convex and differentiable** with respect to c. We can then write

$$\frac{\partial A}{\partial c} = 0 \iff \sum_i^N -2(x_i - c) = 0$$

Solving this equation is **the same as finding the value of c for which A is minimum**.

$$\sum_i^N -2(x_i - c) = 0 \iff \sum_i^N (x_i - c) = 0$$

$$\iff \sum_i^N x_i - \sum_i^N c = 0 \iff -Nc + \sum_i^N x_i = 0 \iff c = \sum_i^N \frac{x_i}{N}$$

Then, the cluster's center that minimize the dispersion is the barycenter.

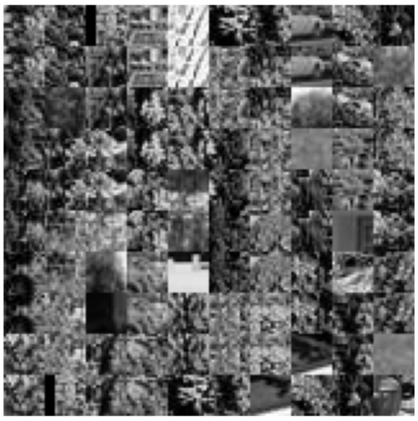
10. In practice, how to choose the "optimal" number of clusters ?

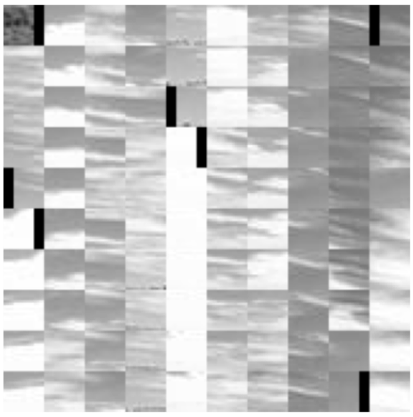
In practice we can use the **Elbow method** to determine the optimal number of clusters. The optimal number on an associated graph is the number k where the **inertia (or distortion)** starts to **decrease linearly**.

11. Why do we create a visual dictionary from the SIFTs and not directly on the patches of raw image pixels ?

As for discretization, the creation of visual dictionaries from *SIFTs* allows **robustness on image alterations**.

12. Comment the results you get.





We can see a **cluster** for each of these pictures. The first picture appears to be a cluster of tree leaves and the second, a cluster of clouds. We find **similar patterns** in the images and make clusters.

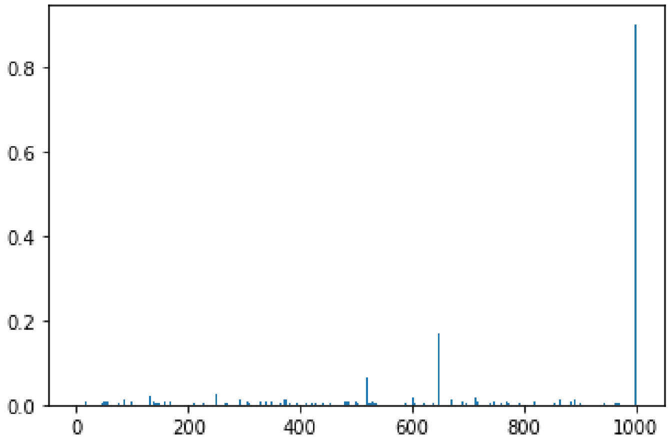
Section 3 - Bag of Words (BoW)

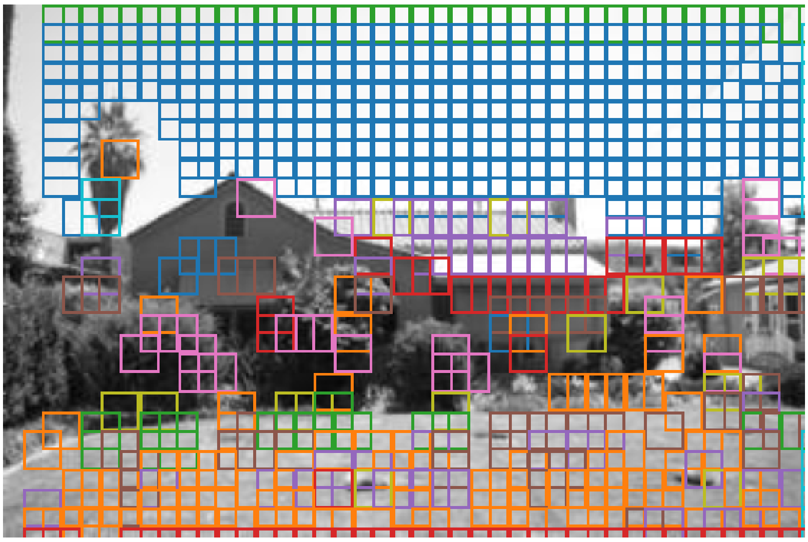
13. Concretely, what does the vector z represent of the image ?

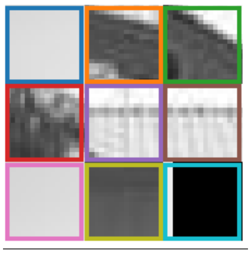
A vector *z* describes/represents globally our image. This vector contains for each cluster, the **number of times a pattern** in the image is associated with it.

14. Show and discuss the visual results you got.

The graph below represents on the x-axis the 1001 clusters and on the y-axis a value that quantifies the occurrence of the cluster (the normalized sum pooling).







Most areas of the image are associated with clusters. We can find different clusters in this image, for example, the blue and pink clusters represent a **uniform area**. **A color represents a cluster** : we can note that the blue cluster corresponds to the cluster whose peak is very high on the graph of clusters.

15. What is the interest of the nearest-neighbors encoding? What other encoding could we use (and why)?

The interest of the “nearest neighbor” encoding is to associate a patch with the word in the dictionary to which it is closest. Another encoding could be **a linear combination of nearest neighbors** by adding a weight according to the distance. This encoding allows to describe a space in a more **expressive** and **stable** way.

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

The interest of sum pooling is to **be able to quantify** the different clusters present on the image. Moreover, in some cases it is possible that we only want to know if a cluster is present in the image, we can then use a **1-0 encoder** with 1 if the cluster is present, 0 otherwise.

17. What is the interest of the L2 normalization? What other normalization could we use (and why)?

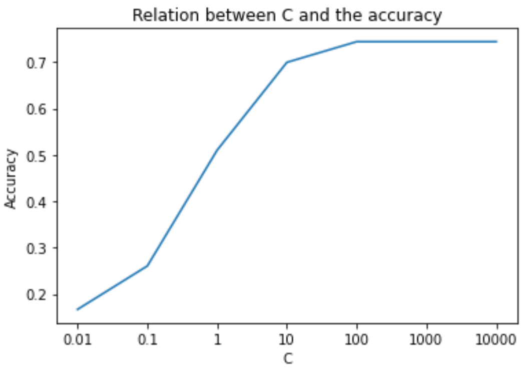
For example if two images have the same content but in one of the two images the content takes more space on the image, it will have more patch than in the other image. This normalization will allow us to **compare** different feature descriptors and thus **image representations**.

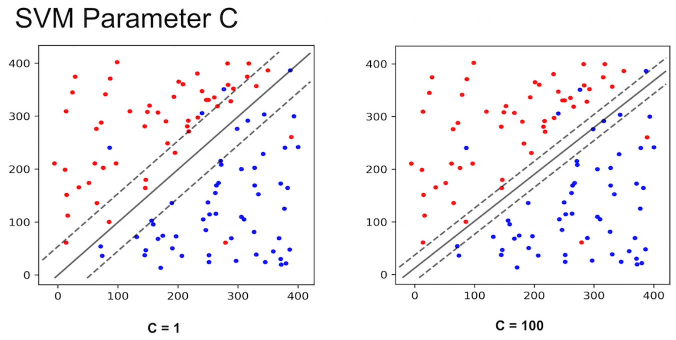
It returns a unit vector where only the **direction** is important. We could also use **L1 norm** or **L2sqr norm** but **L2** probably works best.

Learning a SVM classifier (1-c)

1. Discuss the results, plot for each hyperparameters a graph with the accuracy in the y-axis.
2. Explain the effect of each hyperparameter.

➔ Hyperparameter tuning : C

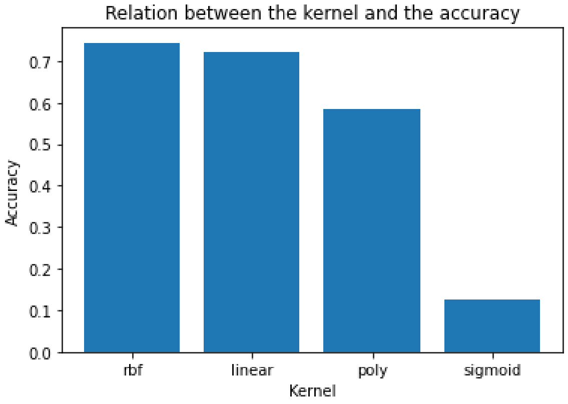


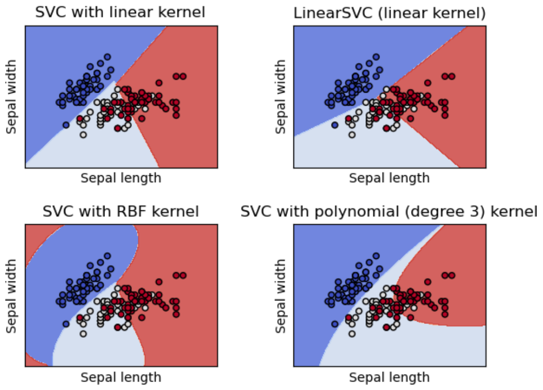


We observe that the larger the hyperparameter **C** is, the higher the accuracy will be. The accuracy does not seem to change from C=100.

The hyperparameter C is the **regularization parameter**. The smaller the value of C, the smaller the penalty for misclassified points (**large margin**). Conversely, the larger the value of C, the more we minimize the number of misclassified examples due to a high penalty (**small margin**).

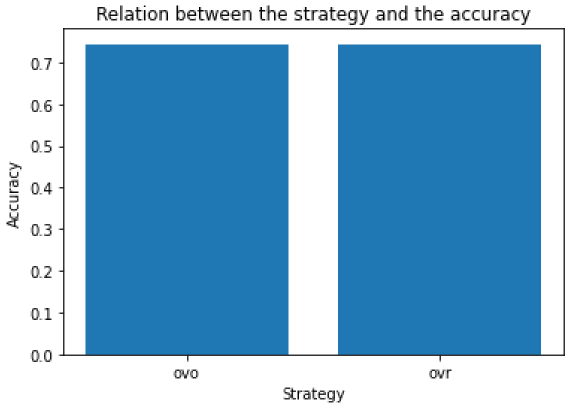
➔ Hyperparameter tuning : Kernel





On these data, the **rbf** and **linear** kernels seem to give the best results.
Kernels are used to define the **shape** of the dividing lines.

➔ Hyperparameter tuning : Strategy



The **One-vs-One** and **One-vs-Rest** methods seem to give similar performances.

With the One-vs-One method, we confront each class of images, **one by one**, with each of the other classes, unlike the second method which confronts each class with **all** the other classes at the same time.

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

The **test set** is not supposed to be known during the training of the models. The validation set is then used to **test/evaluate** the performance of our models **during training** without having to touch the **test set**.