Descripteur SIFT et BoW

ENCADRE PAR THOMAS ROBERT

Louise Marchal et Alexia Bourmaud
SORBONNE UNIVERSITE SCIENCE - RDFIA |

SIFT & BoW

1. Explication de la démarche expérimentale

Pendant les TME1 et 2 nous avons appris à représenter des images avec la technique du BoW. Pour cela nous avons dans un premier temps utilisé des descripteurs visuels locaux, appelé SIFTs (Scale-Invariant Feature Transform), afin de caractériser numériquement des petits patches de l'image. Puis nous avons créé un dictionnaire visuel de descripteurs-types les plus fréquents parmi tous les SIFTs apparaissant dans le jeu d'images. Et enfin nous avons utilisé le dictionnaire visuel pour représenter chaque image du jeu d'apprentissage de façon condensée.

Dans un premier temps, nous avons cherché à caractériser numériquement des petits patchs (16*16) des images à l'aide des descripteurs SIFTs. Pour cela, pour chaque pixel du patch nous avons calculé le gradient afin d'obtenir à la fois une matrice d'orientation des gradients et une matrice de norme des gradients. Ensuite nous avons discrétisé la matrice d'orientation des gradients en 8 valeurs, et pondéré la matrice de norme des gradients à l'aide d'une gaussienne. Ensuite pour chaque région 4*4 on a calculé l'histogramme des orientations des gradients, puis on a concaténé ces 16 histogrammes pour obtenir un vecteur de taille 128. Finalement nous avons seuillé puis normalisé le vecteur des histogrammes afin d'obtenir le SIFT du patch.

Dans un deuxième temps nous avons créé un dictionnaire visuel contenant 1000 descripteurs SIFTs types présents dans la base d'images. Pour cela nous avons calculé tous les SIFTs de chaque image du jeu d'apprentissage, dans le but d'obtenir tous les SIFTs qui existent dans le jeu d'images. Puis nous avons utilisé l'algorithme des k-means afin de séparer l'ensemble de SIFTs en 1000 cluster. Chaque cluster contenant des SIFTs que se ressemblent. Enfin nous avons utilisé les 1000 centroïdes des clusters pour créer les 1000 descripteurs-types du dictionnaire visuel.

Pour finir, nous avons utilisé le dictionnaire visuel créé, afin de représenter de façon condensée chaque image de la base. Cette représentation de l'image permet de la caractériser et sera réutilisée plus tard (TME 3) pour la classer. Pour obtenir cette représentation nous avons commencé par chercher tous les descripteurs SIFTs de l'image. Puis pour chaque descripteurs nous avons cherché avec quel mot du dictionnaire visuel il était le plus proche (étape de coding). Ensuite nous avons créé un vecteur de taille 1000, qui à chaque mot du dictionnaire visuel associait le nombre de descripteurs de l'image qui lui était le plus proche.

2. Réponses aux questions

1.

$$M_x = \frac{1}{4}\begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} = \frac{1}{2}\begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix} * \frac{1}{2}\begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}^T$$

Donc $M_x = h_y * h_x^t$

Même chose pour M_y

$$M_y = \frac{1}{4}\begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix} = \frac{1}{2}\begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix} * \frac{1}{2}\begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}^T$$

Donc $M_y = h_x * h_y^t$

2. Séparer le filtre de convolution permet d'avoir moins de calcul lorsqu'on applique le masque sur l'image (6 calculs contre 9).

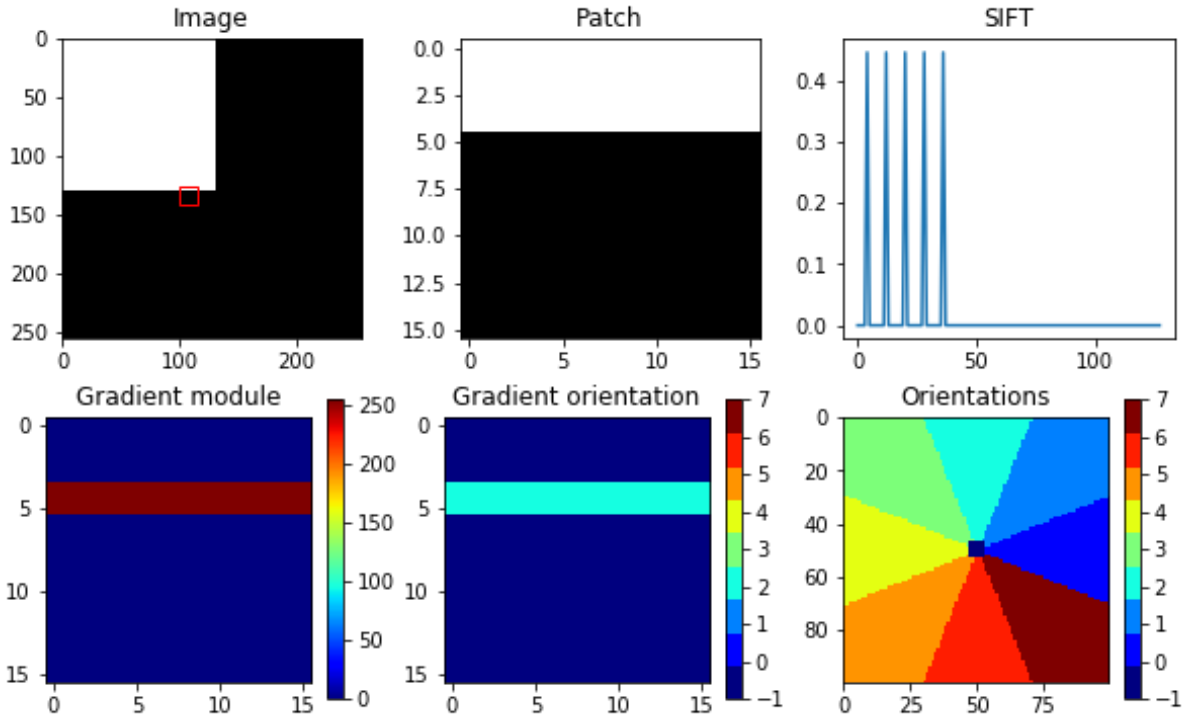
3. La pondération par masque gaussien permet un lissage afin d'atténuer le bruit.

4. La discrétisation permet de considérer comme similaire deux points qui ont une orientation de gradient légèrement différente.

5. Le post-processing permet de rendre le descripteur SIFT invariant aux changements de contraste en le divisant par sa norme euclidienne et de le rendre insensible aux changements d'exposition en le seuillant à 0,2.

6. Le SIFT est robuste aux mouvements de l'image.

7. On observe la propriété principale du SIFT c'est-à-dire sont invariance à la rotation de l'image. En effet, les patchs suivant sont similaires et sont simplement sous des orientations différentes, le deuxième est donc tourné de 90°. Cependant le graphique de SIFT et celui des orientations sont identiques pour les deux patchs.



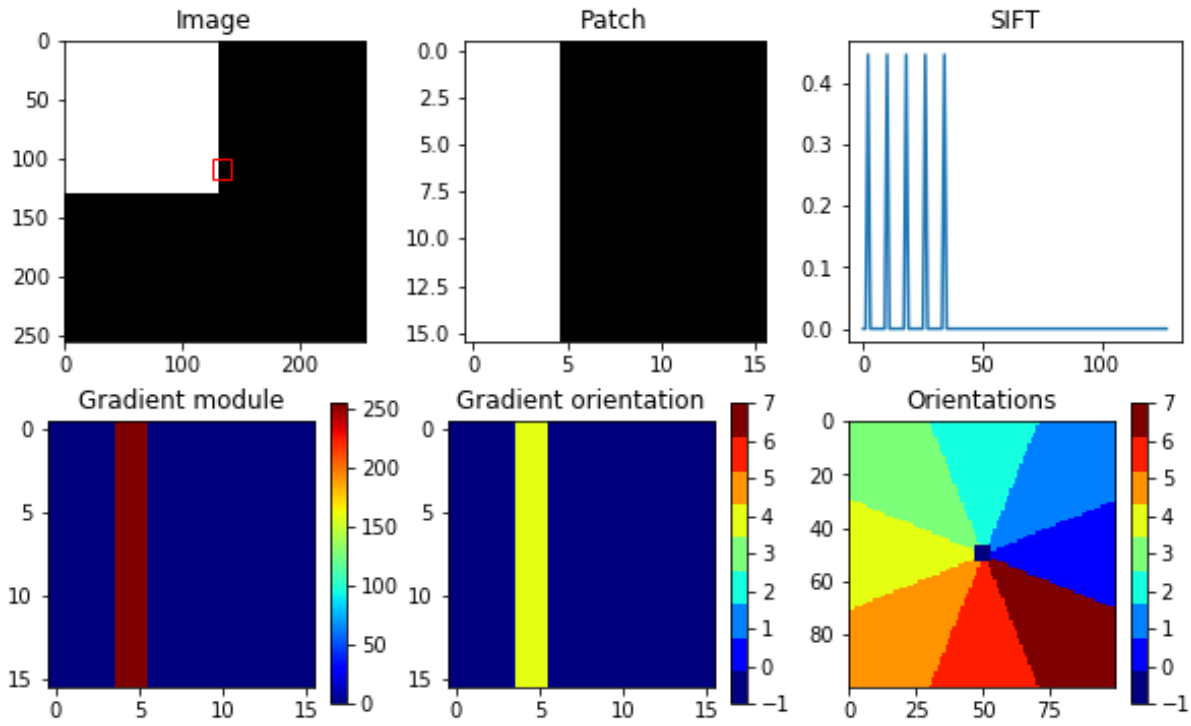


Figure 1 : Comparaison des SIFT suivant l'orientation d'un patch

Le patch suivant est différent des patchs précédents car il se situe autour d'un point anguleux de l'image, nous observons que le SIFT obtenu est également différents des précédents. De plus on observe une variation de l'orientation et du module du gradient.

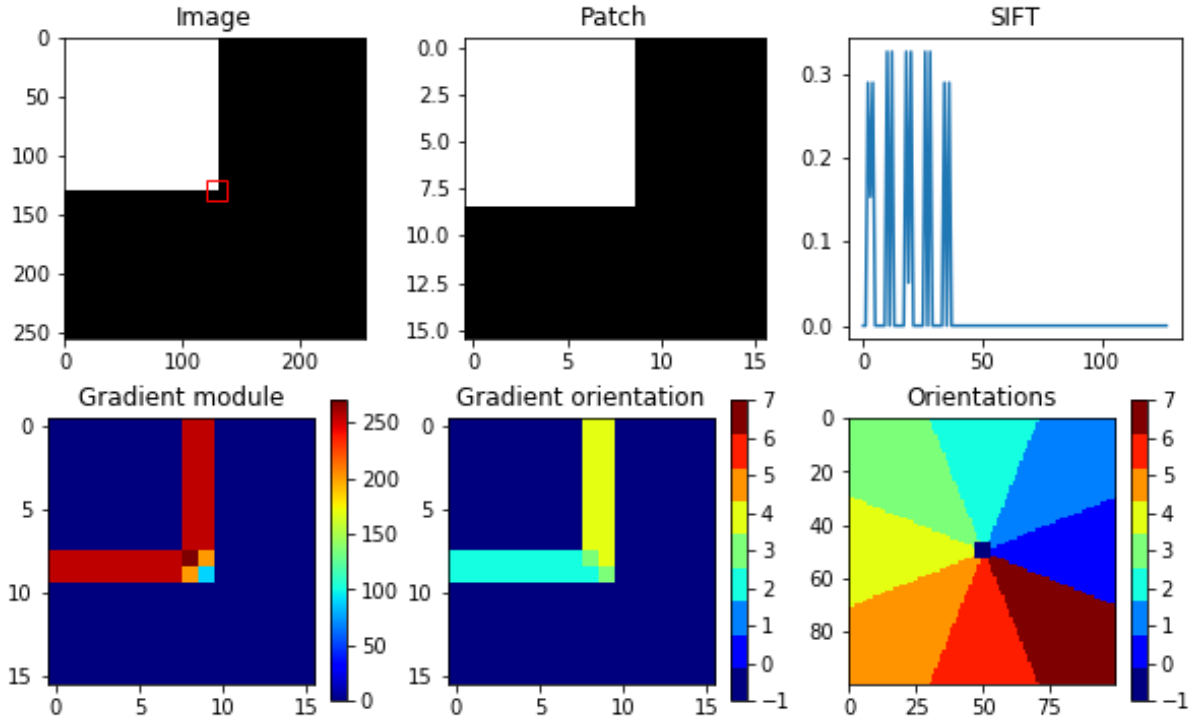


Figure 2 : SIFT d'un patch se positionnant autour d'un point anguleux

8. Le dictionnaire va permettre d'avoir un ensemble de features/points d'intérêts de références pour décrire sous forme de mots visuels les images.

9. Pour chaque cluster on choisit de définir un point au centre de tous les autres points afin de minimiser la distance euclidienne entre tous ces points, c'est le barycentre.

10. Afin de trouver le nombre de clusters idéal, il faut effectuer des tests et observer les résultats.

11. Les centroïdes du dictionnaire sont des SIFTs qui ne sont associés à aucun patch. Pour pouvoir observer un des centroïdes sous forme d'image, il faut donc observer le patch le plus proche.

12.

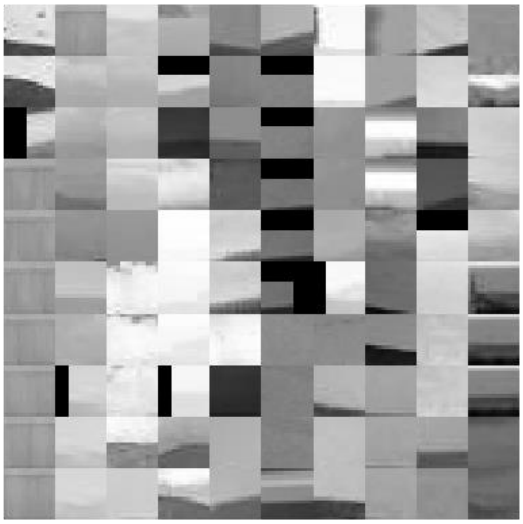


Figure 3 : Patchs les plus proches du cluster 6

On observe que les patchs sont similaires, ils ont une même nuance. Cela nous montre que le clustering fonctionne bien.

Pour pouvoir observer visuellement les différents mots du dictionnaire nous avons cherché le patch le plus proche de chaque centroïde.

Figure 4 : Mot visuel associé à chaque centroïde

On voit que certains patchs se ressemblent fortement, cela signifie peut-être que certains centroïdes sont très proches. Il serait intéressant de diminuer le nombre de cluster pour essayer de regrouper ces patchs qui se ressemblent.

13. Le vecteur z représente un histogramme de la proportion de patch de l'image appartenant à chaque cluster.

14.

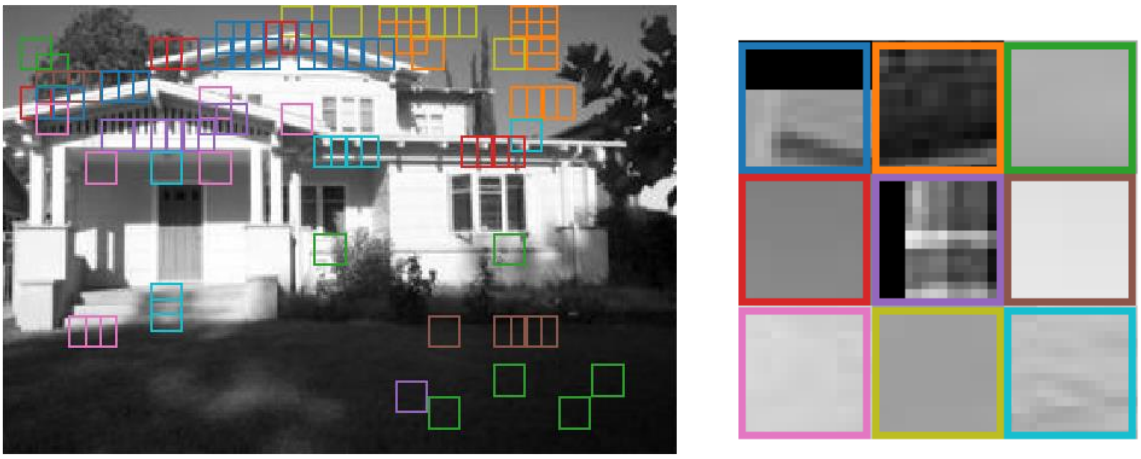


Figure 5 : Résultats image 7

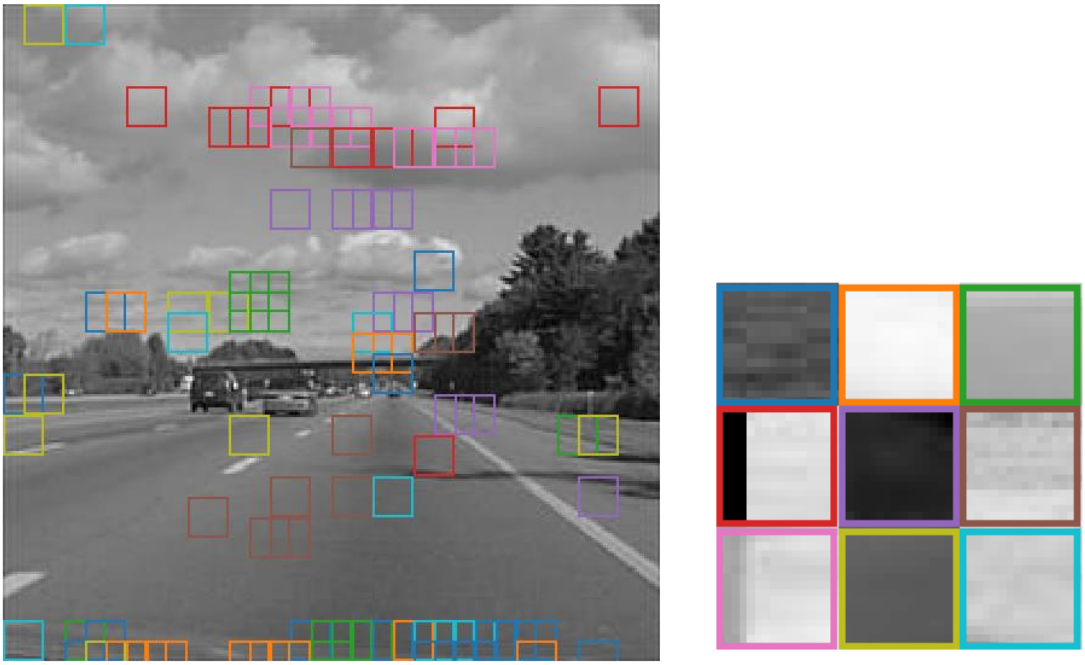


Figure 6 : Résultats image 1000

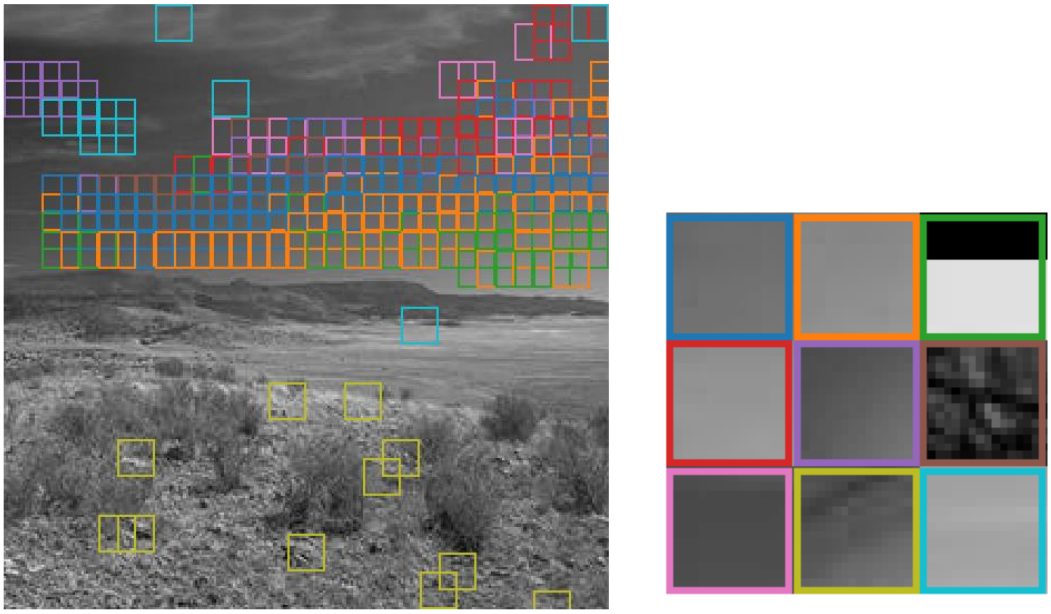


Figure 7 : Résultats image 2000

15. Cela permet d'associer à chaque descripteur le centroïde du dictionnaire visuel qui est le plus proche et donc qui lui est le plus similaire.

On peut aussi utiliser le soft-coding ou le sparse-coding.

16. Le pooling calcul le nombre de patch associé à chaque cluster. Il existe aussi le max-pooling.