Calcul de descripteurs locaux et de dictionnaires visuels

Chloé Dequeker, Octave Mariotti

4 octobre 2017

1 Extraction de descripteurs locaux : SIFT

1.1 Masques séparables

Montrons que les masques M_x (resp. M_y) sont séparables, c'est à dire qu'ils peuvent s'écrire comme $M_x = h_x \times h_y$, où h_x est un masque 1d sur les colonnes et h_y un masque 1d sur les lignes. On souhaite donc montrer qu'il existe h_x et h_y tels que :

$$\frac{1}{4} \times \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} = h_x \times h_y$$

Soit deux vecteurs tels que $h_x = \begin{pmatrix} \frac{-1}{4} & \frac{0}{4} & \frac{1}{4} \end{pmatrix}$ et $h_y = \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}$.

$$\begin{aligned} \begin{pmatrix} \frac{-1}{4} & \frac{0}{4} & \frac{1}{4} \end{pmatrix} \times \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix} &= \frac{1}{4} \times \begin{pmatrix} -1 & 0 & 1 \end{pmatrix} \times \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix} \\ &= \frac{1}{4} \times \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} \end{aligned}$$

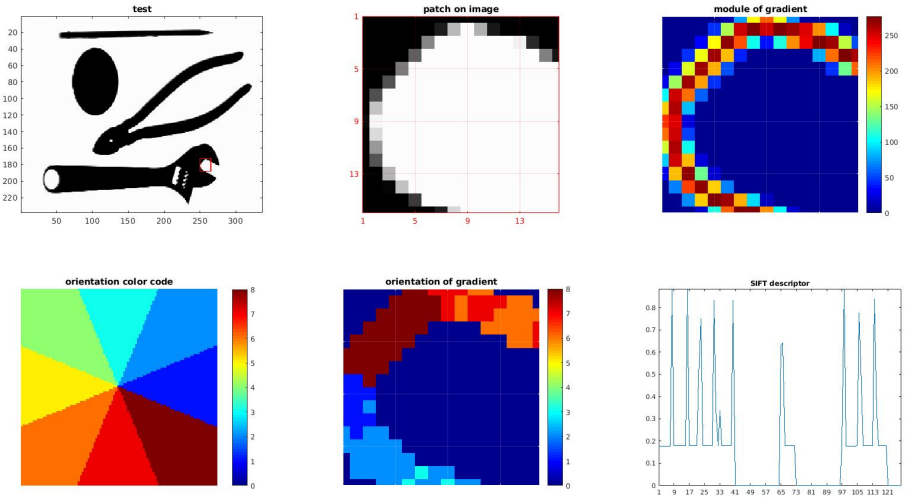
Réciproquement, on obtient la matrice M_y avec les deux vecteurs $h_x = \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix}$ et $h_y = \begin{pmatrix} \frac{1}{4} & \frac{2}{4} & \frac{1}{4} \end{pmatrix}$.

1.2 Descripteur SIFT - Patch

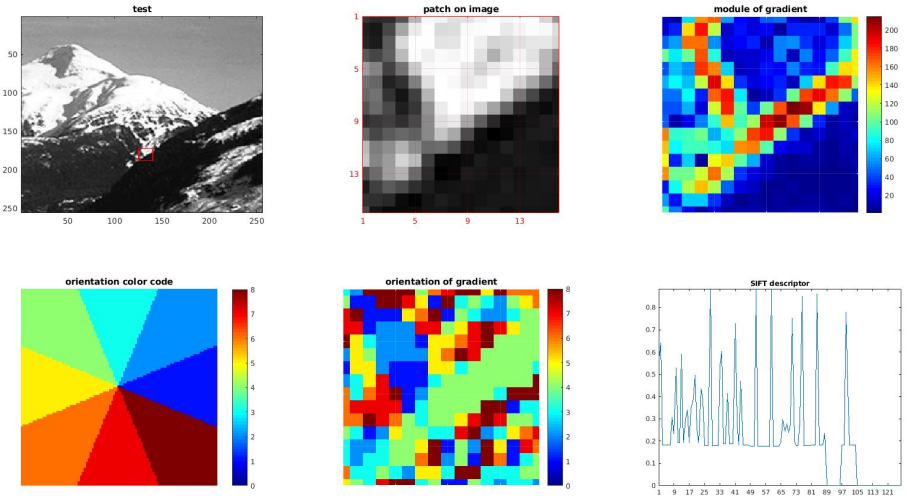
On obtient pour le patch [173 ; 250] la figure 1a un descripteur SIFT qui va décrire une courbe, représentant le contour de la clef anglaise. Les couleurs vont représenter la direction principale de chaque pixel, par rapport à ses voisins. Les descripteurs vont pointer vers l'intérieur de la clef, où l'on retrouve une zone blanche homogène. De la même façon on retrouve sur la figure 1b une description du contraste coloriée selon les huit orientations.

1.3 Cas des régions homogènes

Dans les cas où une région est parfaitement homogène, alors un vecteur SIFT aura une norme de 0 (aucune direction n'a pu être discrétisée et ajoutée à l'histogramme). Cependant, on retrouve rarement dans des photos des zones qui sont parfaitement homogènes, et on aura donc dans les cas de faible contraste un vecteur SIFT de norme faible (il n'y a pas une direction fortement marquée). Dans le calcul du vecteur SIFT on procède à la normalisation l_2, qui va placer l'ensemble des valeurs sur une même échelle et empêchera ainsi de distinguer les zones possédant un faible contraste des zones à fort contraste. Pour pallier ce problème, on a choisi un seuil (0.2) en dessous duquel la norme d'un vecteur SIFT est considérée comme nulle. On obtiendra ainsi la représentation des zones homogènes (bleues) et non-homogènes (rouges) pour la figure 2.



(a) Analyse du descripteur SIFT pour l'image tools.gif pour le patch [173 ; 250]



(b) Analyse du descripteur SIFT pour l'image MITmountain/image_0364.jpg pour le patch [173 ; 125]

1.4 Comparaison par rapport au descripteur de Lowe [1]

Le descripteur de Lowe va échantillonner des points autour du keypoint à considérer. Cet échantillonnage va permettre de dresser un spectre de 36 directions possibles, comparativement aux 8 présentées dans ce TD. La principale différence est que le descripteur de Lowe va ensuite assigner une (ou plusieurs) directions à ce keypoint. La (les) direction(s) associée(s) est (sont) décrite(s) par le pic le plus élevé ainsi que tous les autres pics de l'histogramme à 80% de sa valeur. Cette méthode permet d'orienter les descripteur afin de les comparer en dépit des rotations de l'image. Dans notre cas, cette technique n'est pas implémentée, notre algorithme n'est donc pas robuste aux rotations. Cependant, pour une tâche de classification d'image, il n'est pas forcément nécessaire d'être invariant par rotation : une photographie de montagne présentera toujours le même type de patch sur la ligne d'horizon, avec des voisinages semblables tout en risquant peu d'être incliné par rapport à la verticale, de même pour une maison, pour un intérieur. Cette sensibilité à la rotation est en outre gommée par la plus grande taille des fenêtres de l'histogramme. Comme tous les histogrammes sont conservés, et non uniquement ceux des keypoints, la densité des données par image est bien plus élevée, pouvant expliquer pourquoi cet algorithme est performant en classification. Il aura en revanche du mal à reconnaître des objets particuliers, ou des animaux, qui peuvent être orientés dans différentes directions.

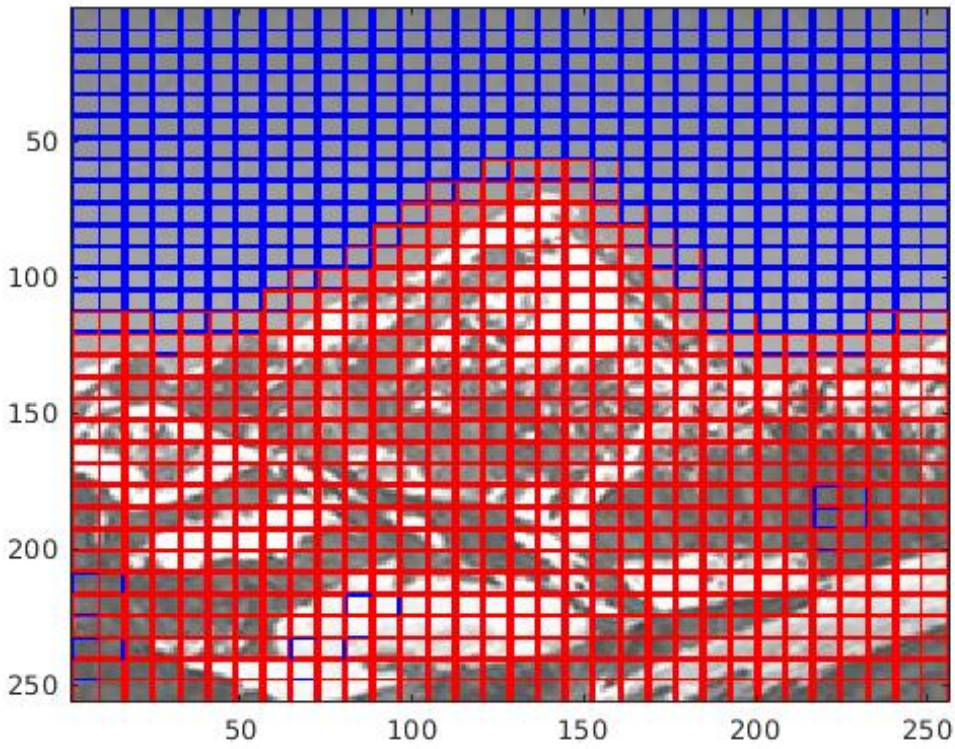


FIGURE 2 – Représentation des régions homogènes pour l'image MITmountain/image_0363.jpg.

2 Génération de dictionnaire visuel

2.1 Quantification des K-Means

Montrons que le barycentre défini par $\frac{1}{|C_m|}\sum_{x_i \in C_m} x_i$ minimise la distorsion des données, en calculant son gradient. On note x_{ij} la j-ième composante de x_i selon le vecteur $\mathbf{e_j}$.

$$\begin{aligned}
\nabla . \sum_{x_i \in C_m} ||x_i - c_m||_2^2 &= \sum_{x_i \in C_m} \nabla . ||x_i - c_m||_2^2 \\
&= \sum_{x_i \in C_m} \nabla . \sum_j (x_{ij} - c_{mj})^2 \\
&= \sum_{x_i \in C_m} \sum_j \nabla . (x_{ij} - c_{mj})^2 \\
&= \sum_{x_i \in C_m} \sum_j (2x_{ij} - 2c_{mj})\mathbf{e_j} \\
&= \sum_j \sum_{x_i \in C_m} (2x_{ij} - 2c_{mj})\mathbf{e_j} \\
&= \sum_j (-2Mc_{mj} + \sum_{x_i \in C_m} 2x_{ij})\mathbf{e_j}
\end{aligned}$$

Cette somme s'annule lorsque, pour tout j :

$$\begin{aligned} -2Mc_{mj} + \sum_{x_i \in C_m} 2x_{ij} = 0 &\Leftrightarrow 2Mc_{mj} = \sum_{x_i \in C_m} 2x_{ij} \\ &\Leftrightarrow Mc_{mj} = \sum_{x_i \in C_m} x_{ij} \\ &\Leftrightarrow c_{mj} = \frac{1}{M} \sum_{x_i \in C_m} x_{ij} \end{aligned}$$

Ainsi, le point défini par $\frac{1}{M} \sum_{x_i \in C_m} x_i$ annule le gradient de la dispersion des données. En vertu de la convexité de la somme, de la fonction carré, et de la distance euclidienne, il s'agit d'un minimum global.

2.2 Calcul matriciel

Afin d'exploiter les fonctions de calcul matriciel de matlab, on peut reformuler l'algorithme des k-means en termes d'additions et de multiplication de matrices. En effet, calculer la distance du point x_i au centre c_m peut se voir en terme de produit scalaire :

$$\begin{aligned} ||x_i - c_m||_2^2 &= \langle x_i - c_m | x_i - c_m \rangle \\ &= \sum_j x_{ij}^2 - 2x_{ij}c_{mj} + c_{mj}^2 \\ &= \sum_j x_{ij}^2 - 2\sum_j x_{ij}c_{mj} + \sum_j c_{mj}^2 \\ &= ||x_i||_2 - 2\sum_j x_{ij}c_{mj} + ||c_m||_2 \end{aligned}$$

Ainsi, on peut calculer la distance des x_i aux centres c_m en effectuant le calcul :

$$D = N_X - 2X * C + N_C$$

Où :
- $N_{X\,(i,j)} = ||x_i||_2$
- $N_{C\,(i,j)} = ||c_j||_2$
- $X_{(i,j)} = x_{ij}$
- $C_{(i,j)} = c_{j\,i}$

Il suffit ensuite de chercher l'élément minimal sur chaque ligne de D.

Références

[1] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2) :91–110, Nov 2004.