Compte rendu TME 1-2

RDFIA

Célina Hanouti
Rym Kaci

Master 2 DAC
2020-2021

Table des matières

1 Introduction

Dans le cadre de ce tp nous allons déterminer une représentation concise et pertinente permettant de synthétiser les informations pour un ensemble d'images. Cette représentation permettra de pouvoir reconnaître des motifs similaires sur des images.

2 Processus général de prétraitement

Avant de pouvoir réaliser un modèle de Machine Learning permettant d'associer une image à une classe donnée, une série d'étapes de *pré-traitements* est nécessaire.

Sur une image, seules certaines zones sont porteuses d'informations pertinentes (les zones fortements contrastées, les contours, les coins…). Le but est de parvenir à extraire l'information contenue dans ces zones afin de synthétiser l'information contenue dans une image. Il faut donc arriver à détecter les zones d'intérêt et les caractériser de manière synthétique.

Pour ce faire nous allons utiliser un descripteur (SIFT dont nous détaillerons le fonctionnement plus bas). De manière générale, un descripteur permet de résumer par un vecteur numérique l'information contenue dans une région de l'image en essayant de rester le plus robuste possible à toutes les transformations possibles sur cette dernière.

Une fois que les régions de l'image sont représentées par un ensemble de vecteurs, un critère de seuillage permettra de garder les zones les plus porteuses d'information. A cette étape chaque image de notre ensemble sera représentée par un *BoF (Bag of Features)* : un ensemble de descripteurs caractérisant les zones d'intérêt. Le but est maintenant d'arriver à regrouper ces features indépendemment des images dans lesquelles ils apparaissent.

Les différents vecteurs sont partitionnés de manière aléatoire, et par un processus de clustering le but et d'arriver à un partitionnement idéal regroupant les vecteurs par similarité. Une fois ce partitionnement atteint -ou approché- chaque groupe sera représenté par un vecteur qui servira de *descripteur type* ou de *prototype*.

Enfin, chaque image sera représentée dans une modélisation *BoW (Bag of Words)*. Un *BoW* est une matrice sparse où les colonnes représentent les descripteurs types, et les lignes les différentes images. Chaque image est donc représentée par un vecteur résumant le degré d'implication de chaque descripteur dans cette dernière.

Cette modélisation BoW est une représentation pertinente pouvant servir par la suite à résoudre différentes problématiques de Machine Learning (clustering, classification d'images).

3 Partie 1 - SIFT

SIFT est un descripteur permettant de traiter une image par patchs (de taille 16*16) dans le but de rester le plus indépendant possible aux changement de dimensions ou de cadrage et à la luminosité. L'information contenue dans un patch est représentée par un vecteur numérique (de taille 128).

Question 1 Les masques M_x et M_y sont séparables, ils peuvent s'écrire sous la forme $M_x = h_y.h_x^T$ et $M_y = h_x.h_y^T$ avec $h_x = (-1, 0, 1)^T$ et $h_y = (1, 2, 1)^T$

Question 2 L'intérêt de séparer les filtres de convolution est de gagner en complexité computationnelle lors de la réalisation du produit de convolution. Au lieu d'effectuer pour chaque pixel une boucle imbriquée (3 x 3 opérations) on effectue 2 boucles (2 x 3 opérations). On passe donc de $O(n^2)$ à $O(n)$.

Question 3 L'intérêt d'utiliser un masque Gaussien est de lisser l'image de telle sorte à estomper les détails trop petits mais aussi de donner plus de poids aux régions à proximité du point, ce qui permet d'avoir plus de robustesse aux changements locaux.

Question 4 Le rôle de la discrétisation des directions du gradient est d'avoir une représentation fixe et de taille raisonnable des orientations du gradient dans une région d'un patch. Discrétiser en 8 directions permet de rester invariant au rotations faibles de l'image (45°).

Question 5 : L'intérêt des différents post-processing

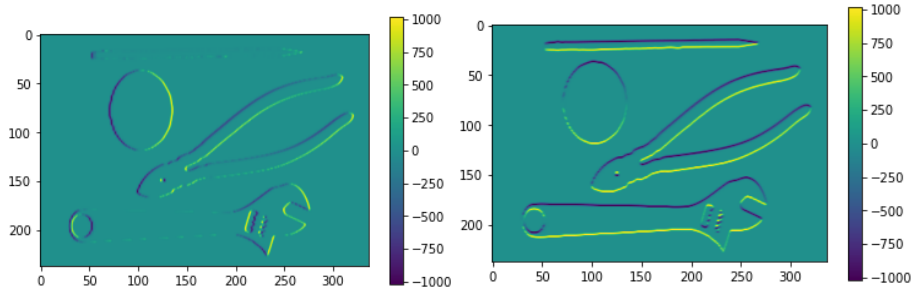
- L'intérêt de la normalisation est d'assurer la stabilité de notre représentation et son invariance aux tranformations photométriques (luminosité, contraste).
- Le seuillage à 0.5 des vecteurs permet de détecter les points d'intrêt, les encodings avec une norme basse étant considérés sans information pertinente.
- Le seuillage des valeurs supérieurs à 0.2 permet une tolérance aux artefacts de luminosité.
- Les valeurs de seuil (0.5 et 0.2) sont des hyper-paramètres.

Question 6

- Le SIFT est un descripteur de patch permettant d'apporter plus d'information sur ce dernier que les valeurs des pixels par exemple, qui elles n'apportent que des informations locales (couleur du pixel).
- Les orientations des gradients permettent de caractériser les points d'intérêt sur un patch.
- Il permet une représentation plus synthétique du patch (vecteur 128 au lieu de 16*16=256).

Question 7 : Interprétation des résultats

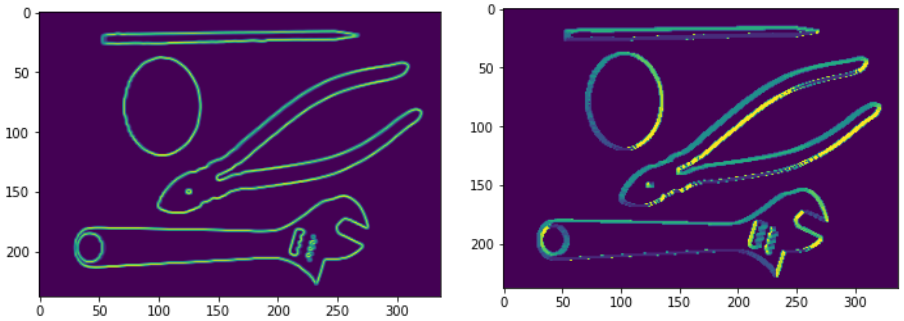
On remarque que le gradient I_x est fort sur les contours verticaux de l'image, un peu moins fort sur les contours en diagonale, et nul sur les zones plates et les contours horizontaux. Inversement, le gradient I_y est quant à lui fort sur les contours horizontaux, et



(a) Gradient I_x (b) Gradient I_y

FIGURE 1 – Gradients de I

nul sur les contours verticaux. Ceci s'explique par le fait que le gradient est par définition perpendiculaire au contour.



(a) Norme du gradient (b) Orientation du gradient

FIGURE 2 – Norme et orientation du gradient de I

Les normes de gradient sont relativement homogènes. Les orientations quant à elles dépendent de l'orientation du contour. Dans cette représentation les orientations ne sont pas encore discrétisées.

La représentation SIFT consiste à traiter l'image par patch de taille 16*16. Un patch est ensuite divisé en 16 régions (4 lignes et 4 colonnes). Pour chaque région on stocke l'histogramme des orientations du gradient normalisées (8 valeurs). La représentation d'un patch est donc un vecteur de taille 128 (16*8).

- Sur les régions plates (complètement blanches ou noires) le gradient étant nul, les 8 valeurs caractérisant la région seront nulles.
- Si la région est uniquement divisée par un trait vertical ou horizontal, une seule valeur dans l'histogramme des orientations va être mise en évidence. Elle correspondra à l'orientation du gradient en x ou en y, l'un des deux étant nul.
- Si la région est divisée par un trait en diagonale deux valeurs dans l'histogramme des orientations sont mise en évidence. L'orientation du gradient en x et l'orientation du gradient en y.

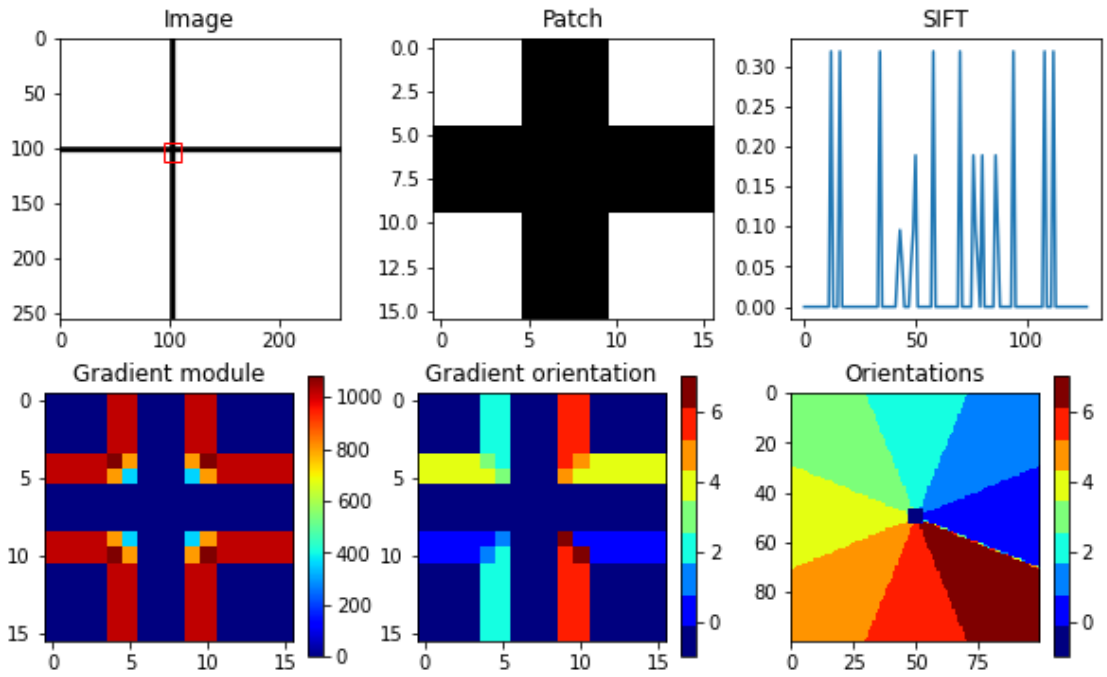


FIGURE 3 – Représentation SIFT

4 Partie 2 - Dictionnaire visuel

Après avoir extrait des descripteurs de chaque image de notre dataset, la deuxième étape consiste à construire un dictionnaire de mots visuels qui décrit au mieux l'ensemble des SIFT de notre dataset d'apprentissage. Un mot de ce dictionnaire est en fait un pattern "prototype" d'un ensemble de SIFTs.

Question 8 Le dictionnaire visuel nous permet d'avoir un espace latent dans lequel il est possible de représenter les images avec les mêmes mots visuels, et donc, de pouvoir les comparer entre elles. De plus, il n'est pas envisageable de représenter une image par l'ensemble des SIFTs la décrivant, en particulier pour la phase de classification, il faudra comparer chaque descripteur de notre image à tous les descripteurs de l'ensemble d'apprentissage, cela est peu pratique étant donné qu'on peut avoir un nombre considérable de SIFTs.

Question 9 Le dictionnaire visuel est construit à l'aide de l'algorithme des K-Means qui permet d'assigner un ensemble de SIFTs à un SIFT "moyen" ce qui permet de partitionner l'espace des SIFTs en regardant les endroits de fortes densités. en considérant le SIFT "moyen" à la place d'un SIFT, l'erreur commise est la distance entre ces derniers.

En considérant les points $\{x_i\}_{i=1..n}$ assignés à un cluster c, on montre de la manière suivante que le centre du *cluster* qui minimise la dispersion est bien le barycentre des points x_i :

$$f(c) = \sum_i \|x_i - c\|_2^2$$

La dérivée première et seconde de f nous donne :

$$\frac{\partial f}{\partial c} = \sum_i -2(x_i - c) \tag{1}$$

$$\frac{\partial^2 f}{\partial c^2} = 2 \times n \tag{2}$$

La dérivée seconde de f par rapport à c est positive donc f admet donc un minimum local.

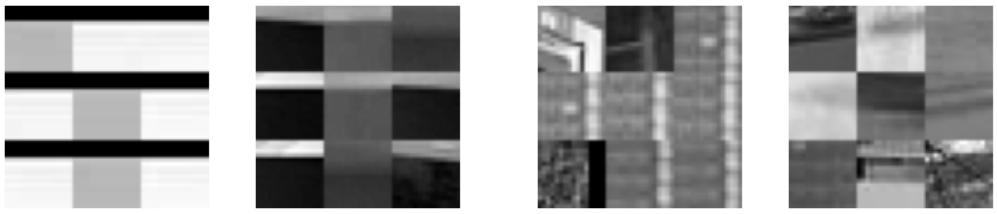
$$\frac{\partial f}{\partial c} = 0 \Rightarrow \sum_i -2(x_i - c) = 0 \Rightarrow nc = \sum_i x_i \tag{3}$$

$$c = \frac{1}{n} \sum_i x_i \tag{4}$$

Question 10 Afin de choisir le nombre de clusters k idéal, on peut utiliser la méthode *Elbow* qui consiste à lancer plusieurs fois l'algorithme des K-means avec plusieurs valeurs de k, et pour chacune des valeurs, on calcule le ratio entre les distances intra-clusters et les distances inter-clusters et on choisit le k qui minimise cette quantité. Il est important de noter l'importance du choix de la valeur de k, en effet, si on se retrouve avec autant de clusters que de SIFTs, il y'a un risque d'overfitting du set d'apprentissage, chaque SIFT serait son propre "prototype". À l'inverse, si on a très peu de clusters, nous disposerons de peu de mots pour caractériser correctement l'ensemble d'apprentissage.

Question 11 L'analyse des éléments du dictionnaire doit se faire à travers les SIFTs et en aucun cas, à travers les pixels des images, car si on appliquait l'algorithme des K-means uniquement sur les intensités des pixels, cela correspondrait à un partitionnement sur ces dernières, on aura donc un cluster pour les pixels rouges, un autre pour les pixels bleus etc.., de plus, un descripteur est bien plus informatif que les couleurs des pixels et comme nous voulons représenter des patterns, un descripteur est tout à fait adéquat, du fait qu'il nous apporte bien plus d'informations sur un pattern (son orientation, son gradient etc..).

Question 12 La figure 4 montre les 9 régions les plus proches de chaque mot visuels parmi 4 choisi au hasard. Sans surprise, on remarque que les régions sont bel et bien similaires entre elles. Sur la figure 5, nous pouvons distinguer différentes régions dont certaines sont uniformes, d'autres correspondent à des toits ou des changements radicaux d'intensité.



(a) (b) (c) (d)

FIGURE 4 – Visualisation des 9 régions les plus proches de 4 mots visuels différents

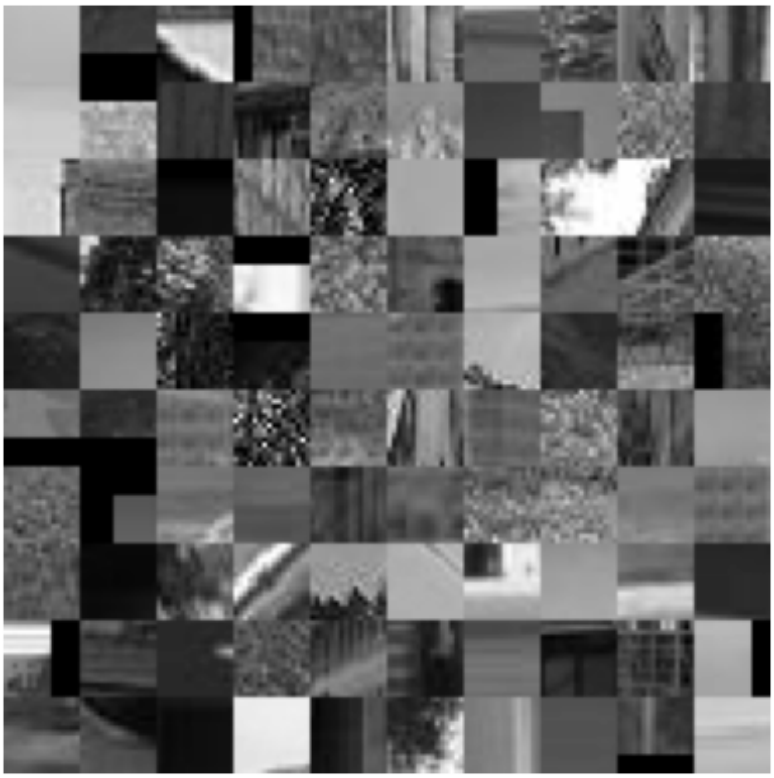
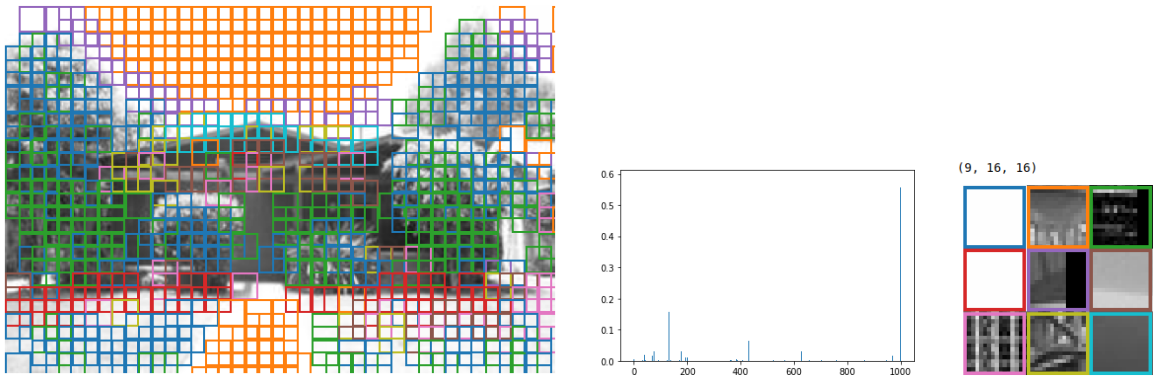


FIGURE 5 – Région la plus proche pour 100 mots visuels pris au hasard

5 Partie 3 - Bag of Words

Question 13 Le vecteur z d'une image représente son encodage dans le dictionnaire visuel, autrement dit, z est le vecteur de représentation de l'image (sa signature) qui correspond à la fréquence d'apparition de chacun des mots visuels dans l'image avec une certaine pondération.

Question 14



(a) (b) (c)

FIGURE 6 – Représentation du vecteur z d'une image

On remarque que les différents mots visuels détectés sur la figure 6-a correspondent bien à des patterns qui sont visuellement proches. Les mots visuels représentent les parties texturées, plates, ou les contours de l'image (figure 6-c). La représentation BoW de l'image (figure 6-b) est une représentation où la probabilité d'apparition d'un mot est très faible, le dictionnaire visuel étant riche. On remarque un pic correspondant à un mot apparaissant beaucoup plus souvent que les autres, il s'agit du mot représentant le ciel (très présent sur l'image).

Question 15 Le codage au plus proche voisin permet d'affecter chaque SIFT à un unique mot visuel qui correspond à celui qui lui est le plus similaire et donc, le plus représentatif. Cependant, on peut considérer d'autres codages avec une affectation plus *soft*, par exemple, on peut utiliser un codage *Softmax* qui définit une distribution de probabilité sur les mots visuels ce qui permet notamment de traiter certains cas ambiguës entre deux mots différents. À la place du codage au plus proche voisin on aura :

$$h_i[j] = \frac{e^{\|x_i - c_j\|^2}}{\sum_k e^{\|x_i - c_k\|^2}}$$

Question 16 Après avoir associé chaque SIFT de l'image à un mot visuel, on agrège afin d'obtenir le vecteur z. Le *Sum Pooling* est une manière intéressante de faire cette agrégation, car elle nous permet d'avoir, pour une image, les fréquences d'apparition de chaque mot visuel. Il existe cependant d'autres méthodes, on peut en citer deux :

- *Max Pooling* : cette méthode nous permet d'avoir une représentation binaire du contenu visuel de l'image, autrement dit, pour chaque mot visuel, nous avons uniquement l'information qui nous dit si celui-ci est dans l'image ou non, on peut donc déduire que cette méthode est beaucoup moins pertinente que le *Sum Pooling*.
- *TF-IDF* : Fréquemment utilisé en Recherche d'information, cette méthode de pondération considère, en plus de la fréquence du mot visuel dans l'image, la fréquence inverse de ce dernier, c'est-à-dire, la proportion d'images dans l'ensemble d'apprentissage contenant ce mot. Ceci permet de donner un poids plus important aux mots visuels les moins fréquents, considérés comme plus discriminants.

Question 17 Les représentations des images (le vecteur z) peuvent avoir des plages de valeurs différentes, c'est pour cela que la normalisation des descripteurs est importante, car elle nous permet d'avoir ces vecteurs dans un même espace normalisé et donc, de pouvoir les comparer entre eux.

6 Conclusion

Ces deux TPs nous ont permis de voir une approche de représentation d'images utilisée en Computer Vision classique, qui est ensuite utilisée pour faire de la classification. Comme on peut le constater, l'accent est mis sur la partie descripteurs et pas sur le processus d'apprentissage. Un inconvénient de cette approche est que la Pipeline est relativement lourde et coûteuse, notamment au niveau de la construction des Bag of Words. À l'inverse, en utilisant des techniques modernes à base de Deep Learning, l'accent portera plus sur l'apprentissage étant donné que les descripteurs seront directement intégrés dans les architectures utilisées.