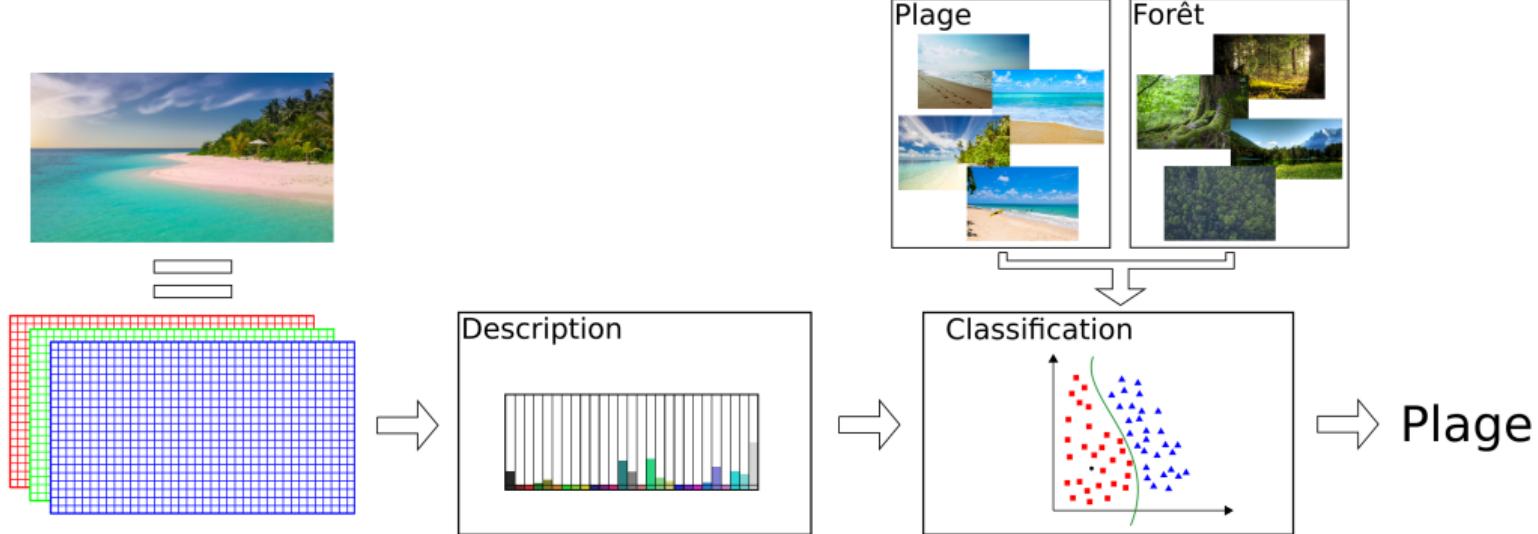


AIV2 – Description locale des images statiques

Pierre Tirilly

Master Informatique, parcours RVA – Université de Lille

Dans le dernier épisode d'AlV2...



Retour sur les descripteurs visuels...

Propriétés souhaitables :

- ▶ *Invariance aux transformations des objets :*
 - ▶ Illumination
 - ▶ Translation
 - ▶ Rotation
 - ▶ Pose
 - ▶ Occultations
- ▶ *“Discriminabilité”* : Aptitude à différencier les différentes classes



Retour sur les descripteurs visuels...

Analyse des descripteurs utilisés :

Descripteur	Invariances	“Discriminabilité”
Hist. de couleurs	Translation Rotation Illumination (limitée)	Pas de motifs précis
Hist. de LBP	Translation Illumination	Peu de motifs possibles



Description locale des images

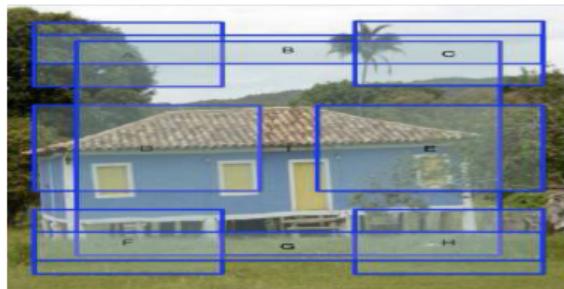
Information globale vs. information locale

Problématique : L'information visuelle est variable selon les positions dans l'image



Description locale : options

A priori géométriques

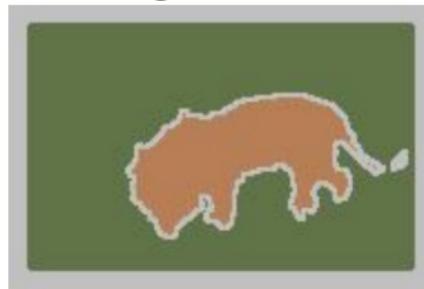


Grille (*dense sampling*)



[Tollari et al., 2008]

Segmentation



Points / régions d'intérêt



[Barnard et al., 2001]

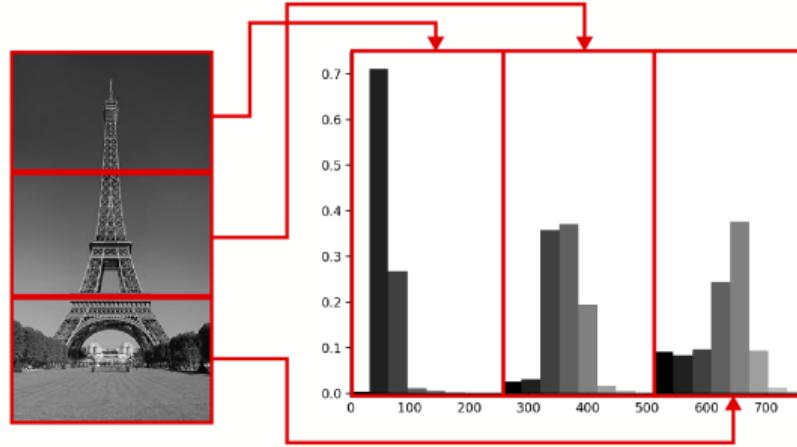
Description locale : découpage

Principe : Image découpée en régions régulières

- ▶ Découpage a priori : régions typiques de la composition des images
- ▶ Découpage en grille (*dense sampling*) : régions carrées

Description de l'image :

- ▶ Calcul d'un descripteur par région
- ▶ Concaténation des descripteurs dans un ordre fixé



Comparaison d'images :

- ▶ Comparaison des descripteurs concaténés
- Géométrie des régions conservée

Description locale : points d'intérêt

Points d'intérêt : Points contenant un motif singulier, représentatif de la présence d'information

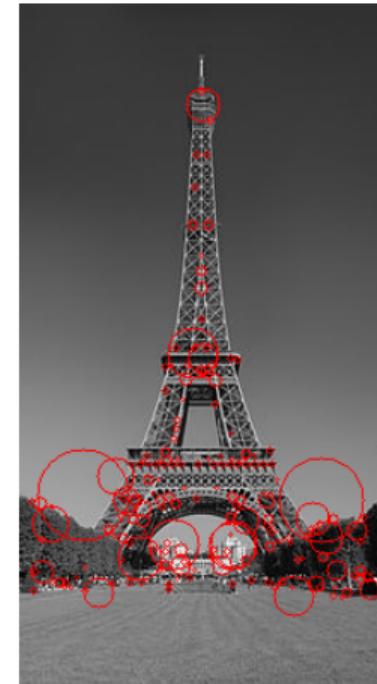
Motifs singuliers :

- ▶ Coins
- ▶ Arêtes
- ▶ Blobs
- ▶ ...

→ Dépendent de l'algorithme de détection utilisé

Questions :

- ▶ Comment trouver ces points d'intérêt ?
- ▶ Comment décrire les régions autour des points ?



Points d'intérêt : Coins de Harris

Coins de Harris : Détection des régions contenant des coins

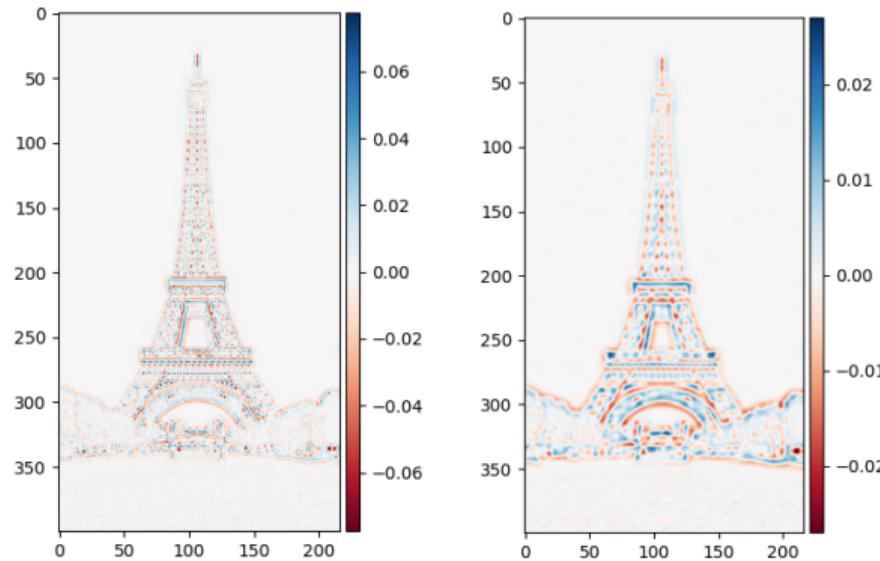
- ▶ Basés sur les vecteurs de gradient
- ▶ Aplats = normes faibles
- ▶ Arêtes = normes élevées, orientations uniformes
- ▶ Coins = normes élevées, orientations variées



Points d'intérêt : DéTECTEUR difference-of-Gaussians (DoG)

DéTECTEUR DoG : Maxima/minima locaux de contours

- ▶ Détection de contours par filtres DoG
- ▶ Point d'intérêt = Maximum/minimum local de la réponse aux filtres
- ▶ Variance des filtres gaussiens ↔ échelle de la région d'intérêt



Points d'intérêt : Exemples



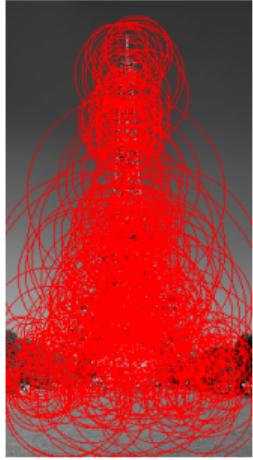
Harris corners

[Harris et Stephen, 1988]



DoG detector

[Lowe, 1999]



Hessian detector

[Bay et al., 2006]



MSER

[Matas et al., 2002]

Points d'intérêt : Propriétés

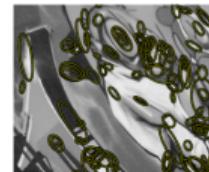
Répétabilité : Fait de retrouver les mêmes points d'intérêt sur un même objet sous différentes transformations

- ▶ de perspective (rotations, transformations affines, . . .)
- ▶ d'échelle
- ▶ de lissage / flou
- ▶ d'illumination
- ▶ de compression

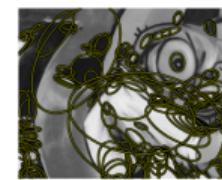
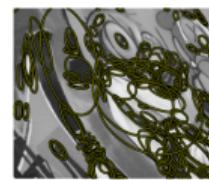
→ Invariance des régions détectées

Description d'une image : Calcul d'un descripteur par région autour de chaque point d'intérêt

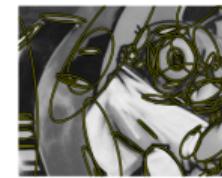
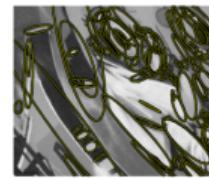
- ▶ Couleur
- ▶ Texture
- ▶ Forme



(a) Harris-Affine



(b) Hessian-Affine



(c) MSER

[Mikolajczyk et al., IJCV 2004]

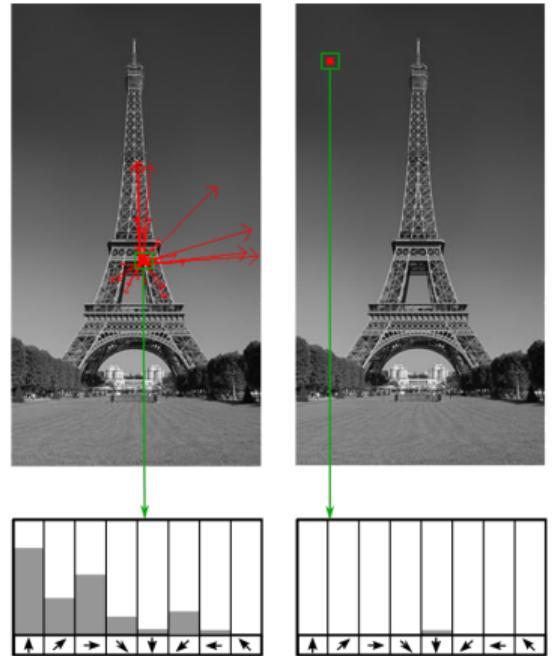
Description des formes : histogrammes d'orientation de gradient (HOG / SIFT)

Gradient : Information sur les contours

- ▶ Vecteurs normaux aux contours
 - ▶ Orientation du gradient → orientation du contour
 - ▶ Norme du gradient → intensité du contour
- Caractéristique de la forme des objets

Histogramme d'orientation de gradients (HOG / SIFT) :

- ▶ Quantification des orientations en n_o bins
- ▶ “Décompte” des vecteurs dans chaque bin, pondérés par leur intensité
- ▶ Normalisation de l'histogramme (plusieurs méthodes existent)



Conclusion sur les descripteurs locaux

Descripteurs locaux versus descripteurs globaux :

- ▶ Descripteur global : perte totale de la structure de l'image
- ▶ Découpage : forte structuration géométrique de l'image (utile à gros grain)
- ▶ Régions d'intérêt :
 - ▶ Structure locale (région) : plus ou moins forte selon le descripteur
 - ▶ Structure globale : distribution spatiale des points

Analyse des descripteurs utilisés :

Descripteur	Invariances	"Discriminabilité"
Hist. de couleurs	Translation Rotation Illumination (limitée)	Pas de motifs précis
Hist. de LBP	Translation Illumination	Peu de motifs possibles
SIFT	Translation Rotation Illumination Pose (limitée)	Motifs très spécifiques

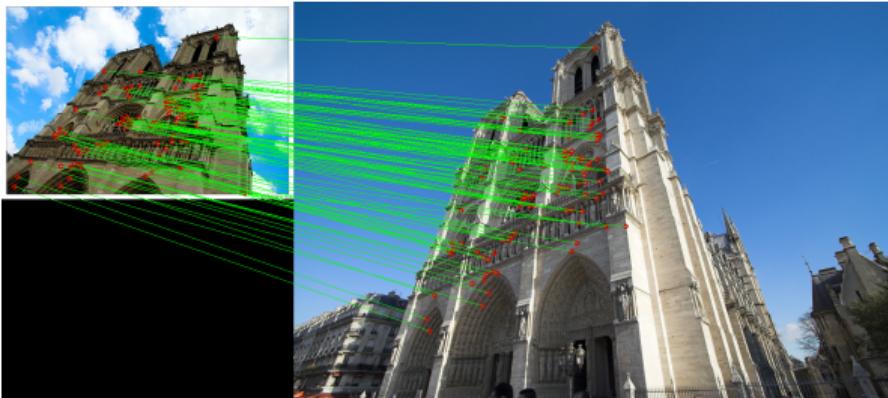
Description par points d'intérêt : problématique

Points d'intérêt : 1 image = n descripteurs

$$x_i \in \mathbb{R}^d$$

Problématique : Comment utiliser ces descripteurs en entrée d'un classifieur ?

- ▶ Nombre de descripteurs variable d'une image à l'autre
- ▶ Pas d'a priori sur la position des descripteurs dans les images
- ▶ Descripteurs peu discriminants individuellement



[EPFL]

Options invalides :

- ▶ Concaténation des vecteurs
- ▶ Classification individuelle + vote / fusion

Descripteurs locaux : modèle en sacs de mots visuels

Sac de mots visuels : Motivations

Descripteurs précédents : Histogrammes

- ▶ Image = ensemble d'éléments discrets (pixels de couleurs, LBP)
- ▶ Description = distribution de ces éléments dans l'image

Inspiration supplémentaire : Classification de textes / recherche d'information

- ▶ Texte = ensemble d'éléments discrets (termes)
- ▶ Description = distribution de ces termes dans le texte (*sac de mots*)

Problème : Descripteurs locaux = vecteurs dans \mathbb{R}^d

- ▶ Image = ensemble de descripteurs locaux = ensemble de vecteurs dans \mathbb{R}^d)

→ Comment transformer ces descripteurs en éléments discrets ?

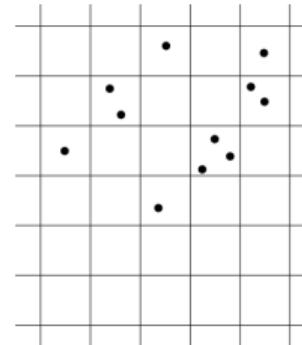
Quantification vectorielle

Quantifieur (vectoriel) : Fonction $q : \mathbb{R}^d \rightarrow [0, K]$

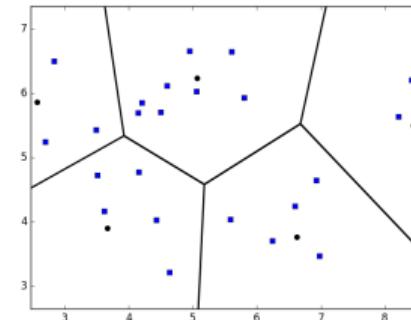
- ▶ Associe à chaque vecteur $x \in \mathbb{R}^d$ une valeur discrète $q(x) \in [0, K]$

Exemples de quantifeurs :

- ▶ Quantifieur uniforme : “grille”, projections aléatoires, réseaux (*lattices*)
- ▶ Partitionnement de données (*clustering*) : k-means, classification ascendante hiérarchique...



Quantifieur uniforme ($d = 2$)



Quantifieur obtenu par k-means
($d = 2$)

Quantification de descripteurs : vocabulaire visuel

Vocabulaire visuel : Quantifieur vectoriel $q(\cdot)$ sur les descripteurs visuels

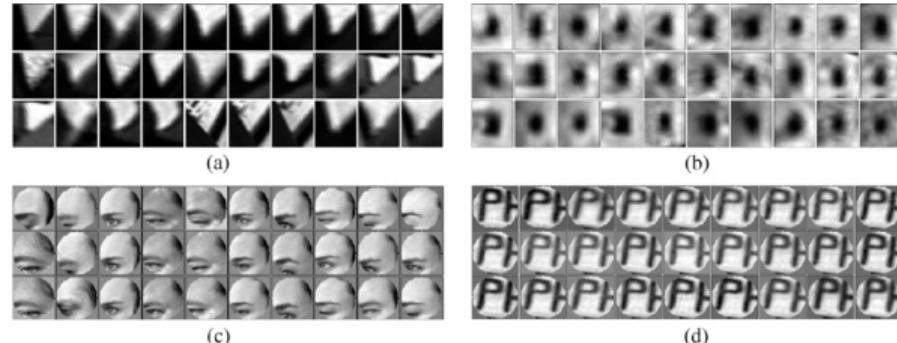
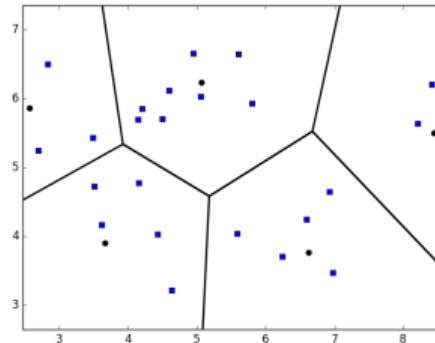
- ▶ Désigne aussi le partitionnement de \mathbb{R}^d induit par $q(\cdot)$

Mot visuel (*visual word*) : une valeur prise par $q(\cdot)$

- ▶ Désigne aussi une des partitions de l'espace contenant les descripteurs regroupés ensemble par $q(\cdot)$

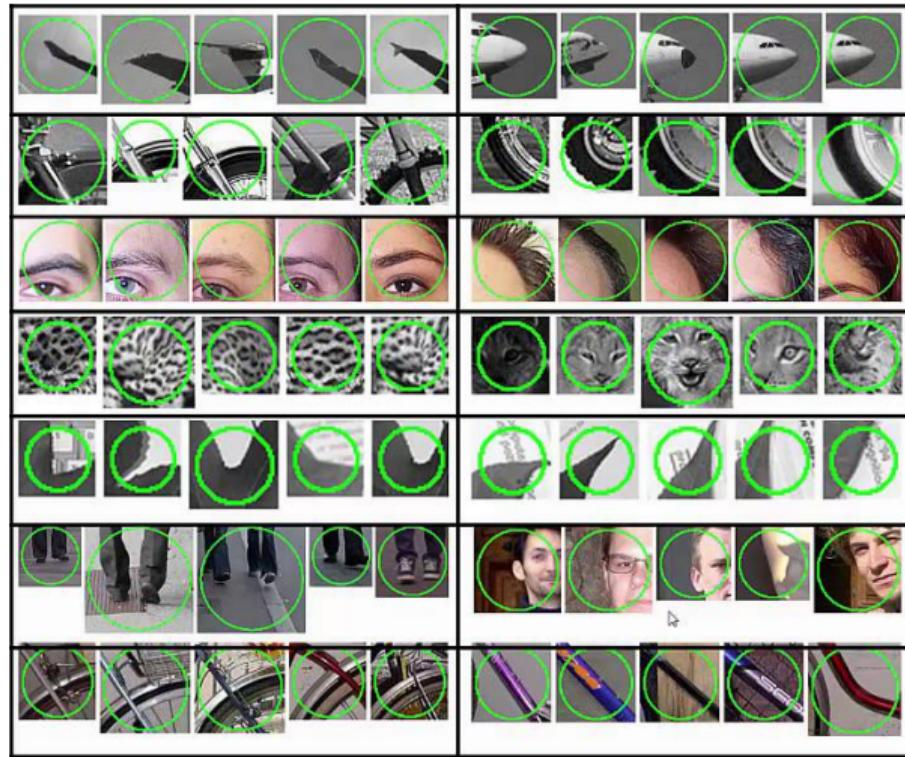
Quantifieurs utilisés : En général, k-means

- ▶ Alternatives : k-means hiérarchique, forêts aléatoires, *lattices*...



[Sivic et Zisserman, 2003]

Mots visuels : exemples



Sacs de mots visuels (*bag of (visual) words, bag of features*)

Construction du vocabulaire : $q(\cdot)$ obtenu par k-means

- ▶ Entraînement de $q(\cdot)$ sur un ensemble de descripteurs distinct
- ▶ Nombre de clusters / mots visuels arbitraire ($10^4 - 10^5$)
- ▶ Assignation des x_j au centroïde le plus proche



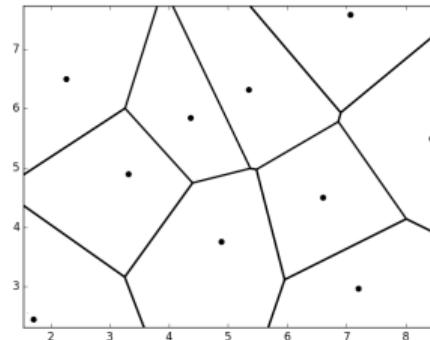
Sac de mots visuels : Ensemble de couples $\{(w_i, \text{tf}_i)\}_{i=1}^m$

- ▶ $w_j = q(x_j)$ est le mot visuel associé au descripteur x_j dans le vocabulaire $q(\cdot)$
- ▶ tf_i est le nombre de descripteurs $x_j \in I$ tels que $w_i = q(x_j)$

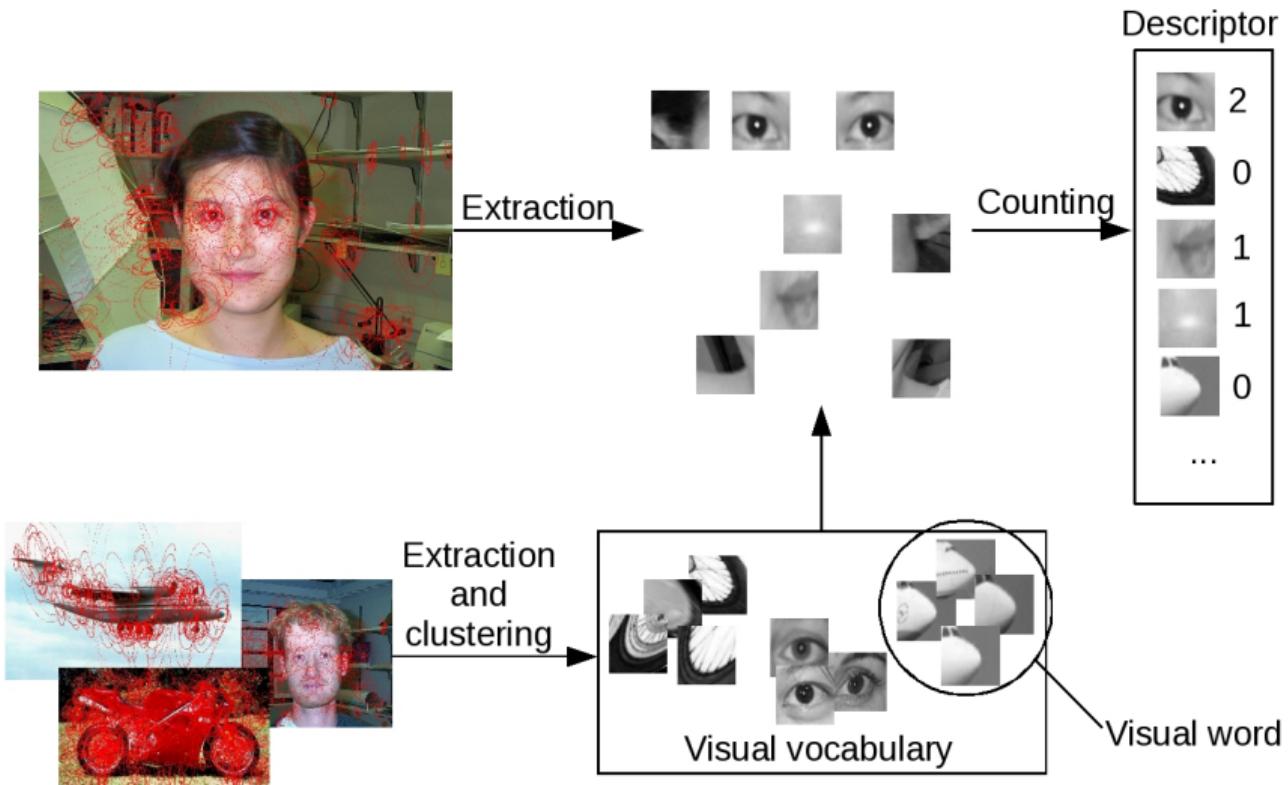
Description de l'image : Histogramme des w_j

- ▶ $h_i = [\text{tf}_0, \text{tf}_1, \text{tf}_2, \dots, \text{tf}_K]$
- ▶ Normalisation de l'histogramme \rightarrow invariance au nombre de points d'intérêt

$\rightarrow 1$ image $= 1$ vecteur de \mathbb{R}^K



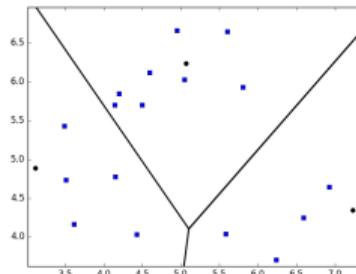
Sacs de mots visuels (*bag of (visual) words, bag of features*)



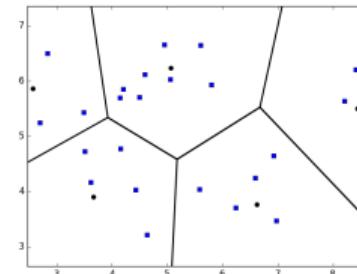
Taille du vocabulaire : analyse

Impact du vocabulaire : Le vocabulaire visuel est une approximation du matching de descripteurs

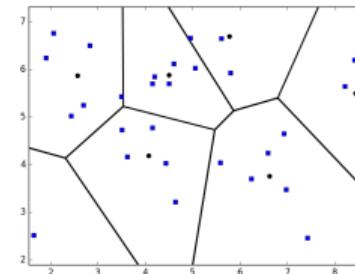
- Vocabulaire plus grand → approximation plus fine



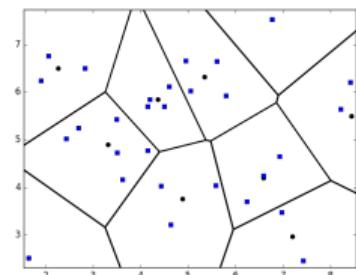
3 mots visuels



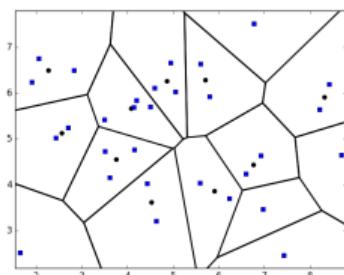
5 mots visuels



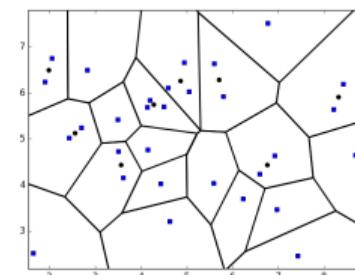
7 mots visuels



10 mots visuels



15 mots visuels

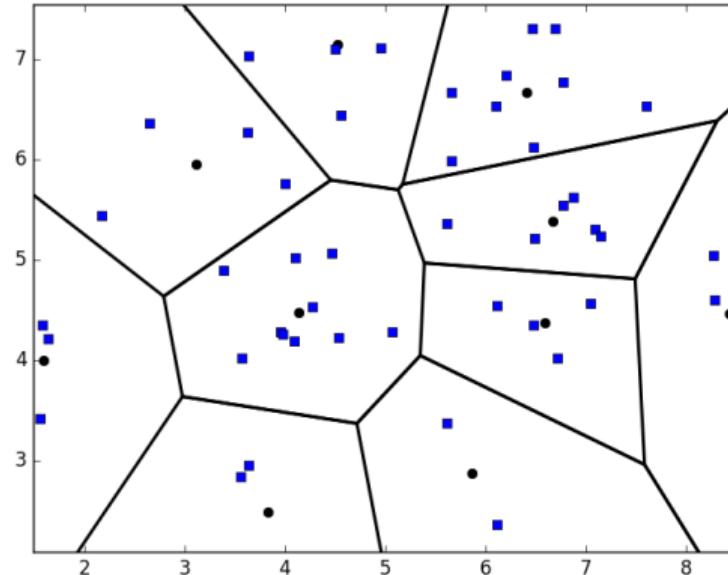


20 mots visuels

Limites des sacs de mots visuels : erreur de quantification

Rappel : Le vocabulaire visuel est une approximation du matching de descripteurs

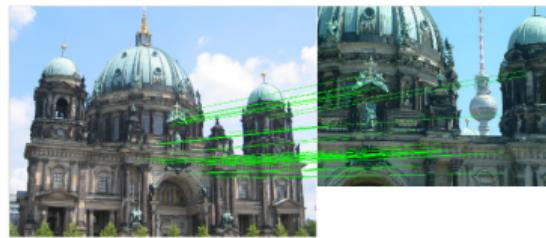
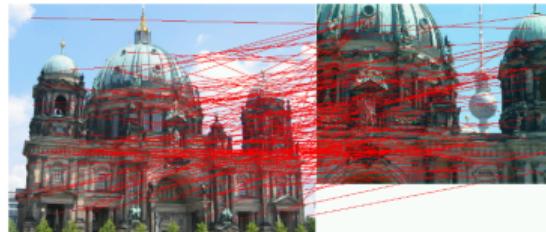
- ▶ Deux descripteurs d'un même mot visuel peuvent être éloignés
- ▶ Deux descripteurs proches peuvent être dans un mot visuel différent



Limites des sacs de mots visuels : perte de la géométrie

Cohérence géométrique : Les objets sont composés de motifs locaux organisés géométriquement

- ▶ Le modèle en sac de mots ignore cette composante géométrique
- ▶ Possibilité de l'intégrer à la description ou d'utiliser des post-traitements



[Raguram et al., 2012]

Descripteurs locaux : VLAD

Intégrer l'erreur de quantification

Rappel : La quantification de descripteurs est une approximation du matching exact

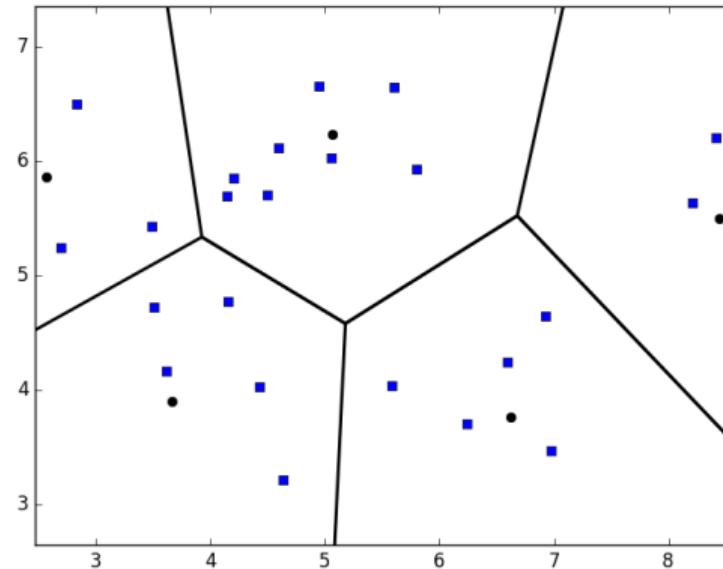
- ▶ Un descripteur est approché par le centroïde de son mot visuel
- Statistique de premier ordre (moyenne)

Erreur de quantification : Information perdue dans l'approximation

- ▶ Résidu : $r = x - c_q(x)$

Problématique : Comment intégrer ce résidu dans le descripteur

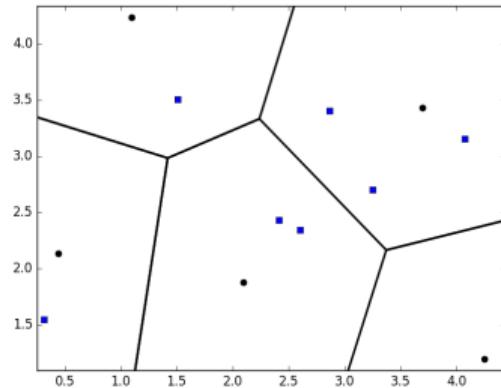
- Intégrer une statistique de second ordre (écart-type)



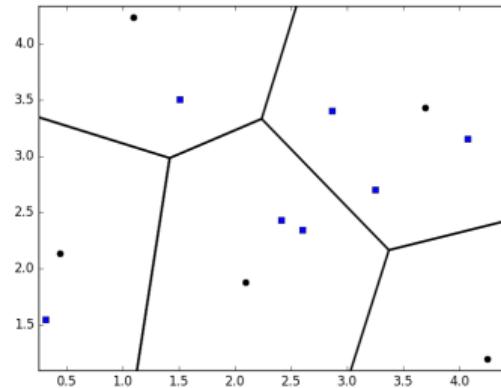
Vector of locally aggregated descriptors (VLAD) : principe

Idée : Intégrer les résidus au vecteur descripteur

- ▶ Accumuler les résidus au sein de chaque mot visuel
- \approx Calcul de l'écart-type des descripteurs du mot visuel pour l'image



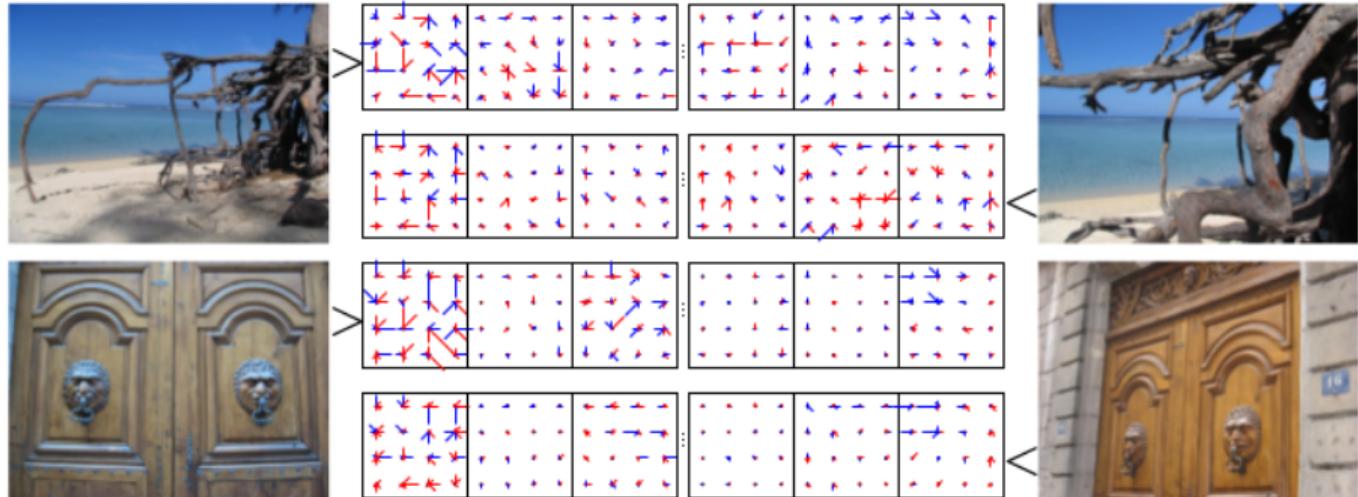
$$c_{q(x)} = \operatorname{argmin}_j \|x - c_j\|_2$$



$$v_i = \sum_{\{x | q(x)=i\}} x - c_i$$

$$v = [v_1 v_2 \dots v_K]$$

VLAD : exemples



[Jégou et al., 2010]

VLAD : Construction et propriétés

Construction : Pour une image I à n_I descripteurs locaux

1. Assigner chaque descripteur x à son centroïde le plus proche dans $q(\cdot)$
2. Calculer les résidus $r = x - c_{q(x)}$
3. Pour chaque mot du vocabulaire, sommer les résidus des descripteurs assignés :
$$v_i = \sum_{\{x|q(x)=i\}} x - c_i$$
4. Concaténer les vecteurs v_i sur les K mots visuels : $v = [v_1 v_2 \dots v_K]$

Normalisation : Les valeurs de v_i dépendent

- ▶ Du nombre de descripteurs x assignés à $q(x) \rightarrow$
$$v = \left(\text{signe}(v_1) \cdot \sqrt{|v_1|}, \text{signe}(v_2) \cdot \sqrt{|v_2|}, \dots, \text{signe}(v_{Kd}) \cdot \sqrt{|v_{Kd}|} \right)$$
- ▶ Du nombre de descripteurs dans l'image \rightarrow normalisation par la norme $\|v\|_2$

Taille du descripteur : $D = K \times d$

- ▶ Utilisation de vocabulaires de petite taille (de l'ordre de $10^1 - 10^2$)
- Possibilité de réduire la dimension à $D' < D$ par ACP (PCA)

TP

- ▶ Implémentation des sacs de mots visuels basés sur SIFT et K-means
- ▶ Implémentation des VLAD
- ▶ Implémentation des découpages géométriques
- ▶ Évaluation sur une tâche de reconnaissance d'objets