

Vision par ordinateur - Reconnaissance Visuelle

Cours 5 : Reconnaissance d'objets et de scènes.

Vers l'interprétation

D'après plusieurs tutoriaux ICCV, NIPS
Torralba, Li, Hoeim...

Céline Hudelot - Mention IA - CentraleSupélec

2022-2023

Plan

1 Introduction

2 Description d'images

- Descripteurs basiques
- SIFT
- Gist

3 Petite parenthèse : problème du panorama

- Bow

4 Classification

5 Conclusion

Les différentes tâches de reconnaissance visuelle



Les différentes tâches de reconnaissance visuelle

Classification : Est-ce que cette image contient un immeuble ? [Oui/Non]



Les différentes tâches de reconnaissance visuelle

Classification : Est-ce que cette image représente une scène de plage ?
[Oui/Non]



Les différentes tâches de reconnaissance visuelle

Applications de la classification : recherche d'images - organisation de collections d'images



Google Images search results for "street".

Results 19 - 26 of about 46,200,880 for street [Details] (0.04 seconds)

Image Preview	Description	Size	Source
	Street preview	345 x 357 - 17n.jpg	www.town.toronto.ca.us
	Street Maintenance	407 x 107 - 15n.jpg	www.town.toronto.ca.us
	Main Street Station	360 x 238 - 70n.jpg	www.mazatlana.org
	GPO Vivian Copeland at Main Lambar Street, words crookedest See Street View (D570 4A) Details	410 x 314 - 43n.jpg	http://parks.ca.gov
	Lambard Street, words crookedest See Street View (D570 4A) Details	500 x 287 - 59n.jpg	www.2fotours.com
	bashan.alibaba.com	360 x 269 - 38n.jpg	

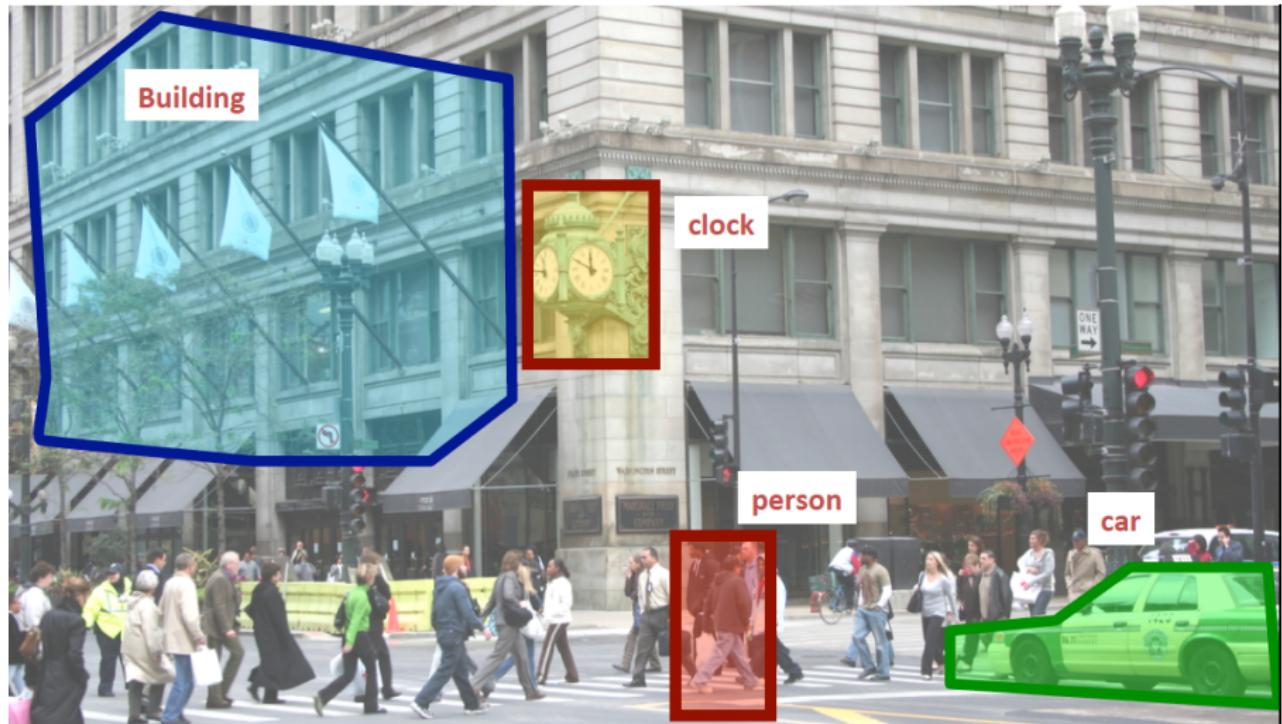
Les différentes tâches de reconnaissance visuelle

Détection : Est-ce que cette image contient une voiture ? [Où ?]



Les différentes tâches de reconnaissance visuelle

Détection : Quels sont les objets présents dans cette image ? [Où ?]



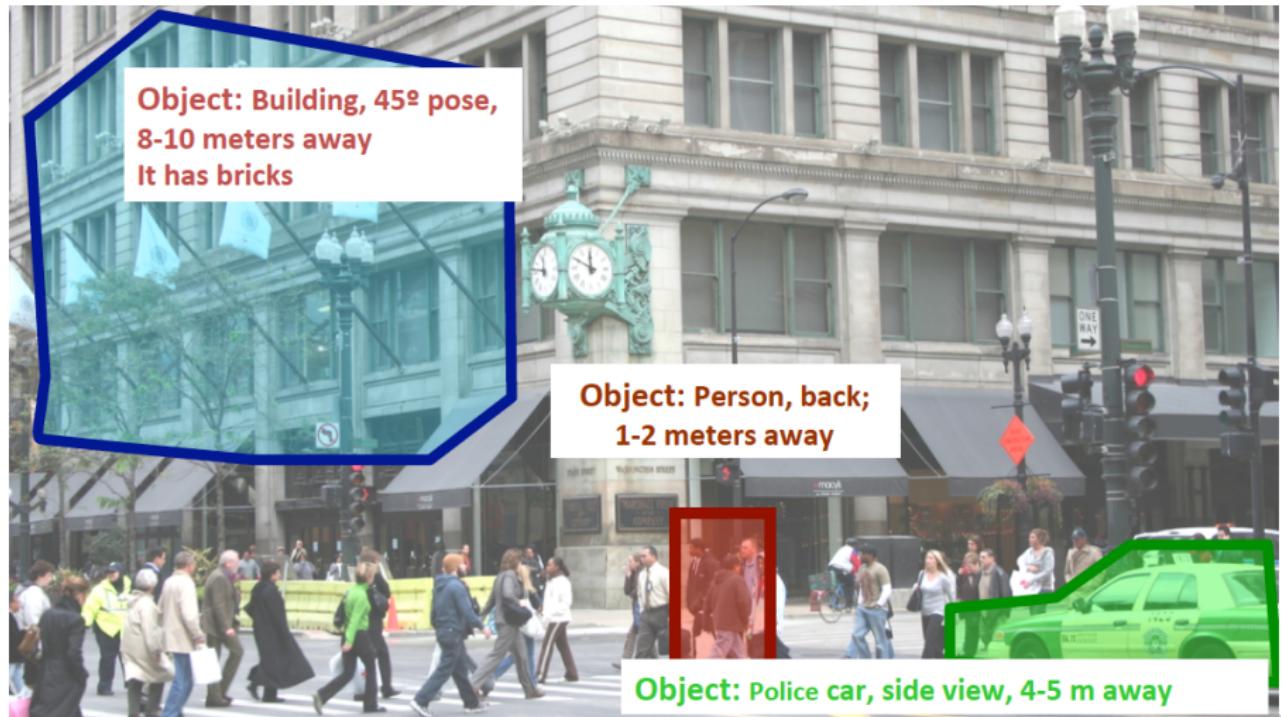
Les différentes tâches de reconnaissance visuelle

Détection : Localisation précise (segmentation sémantique)



Les différentes tâches de reconnaissance visuelle

Détection : Caractérisation sémantique (catégorisation sémantique) et géométrique.



Les différentes tâches de reconnaissance visuelle

Détection : de nombreuses applications



Computational photography



Assistive technologies



Surveillance



Security



Assistive driving

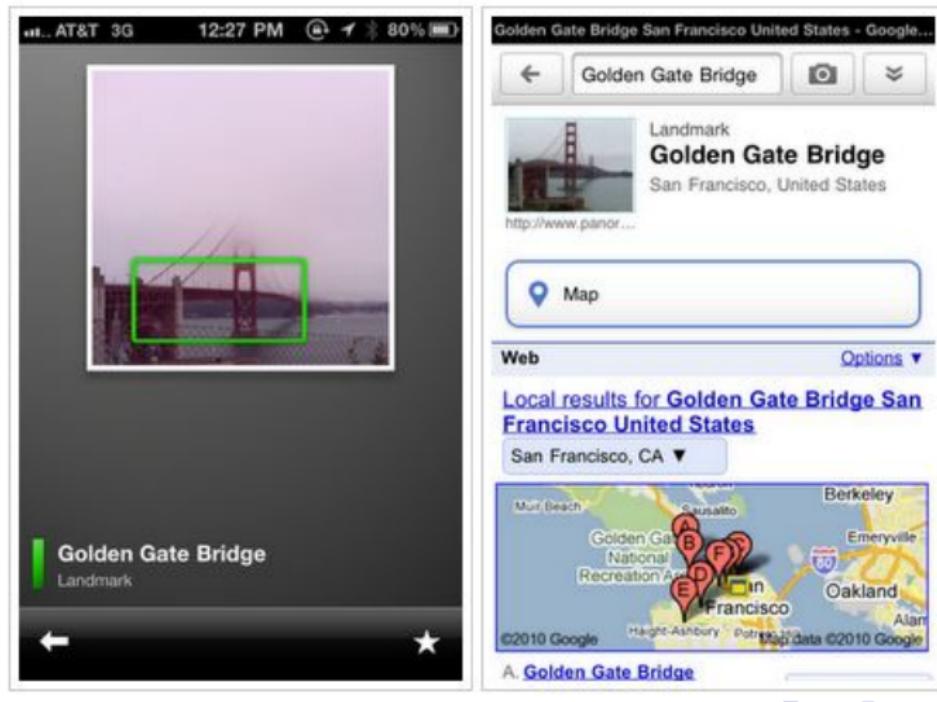
Les différentes tâches de reconnaissance visuelle

Reconnaissance d'instance (vs catégorisation) : est-ce que cette image contient le magasin Macy de Chicago ?



Les différentes tâches de reconnaissance visuelle

Reconnaissance d'instance (vs catégorisation) : de nombreuses applications



Les différentes tâches de reconnaissance visuelle

Reconnaissance d'instance (vs catégorisation) : de nombreuses applications

WHAT MAKES PARIS LOOK LIKE PARIS?



Carl Doersch, Saurabh Singh, Abhinav Gupta, Josef Sivic, and Alexei A. Efros. What Makes Paris Look like Paris ? ACM Transactions on Graphics (SIGGRAPH 2012), August 2012, vol. 31, No. 3

<http://graphics.cs.cmu.edu/projects/whatMakesParis/>

Les différentes tâches de reconnaissance visuelle

Reconnaissance d'activités ou d'évenements : que font ces gens ?



La reconnaissance : un problème complexe



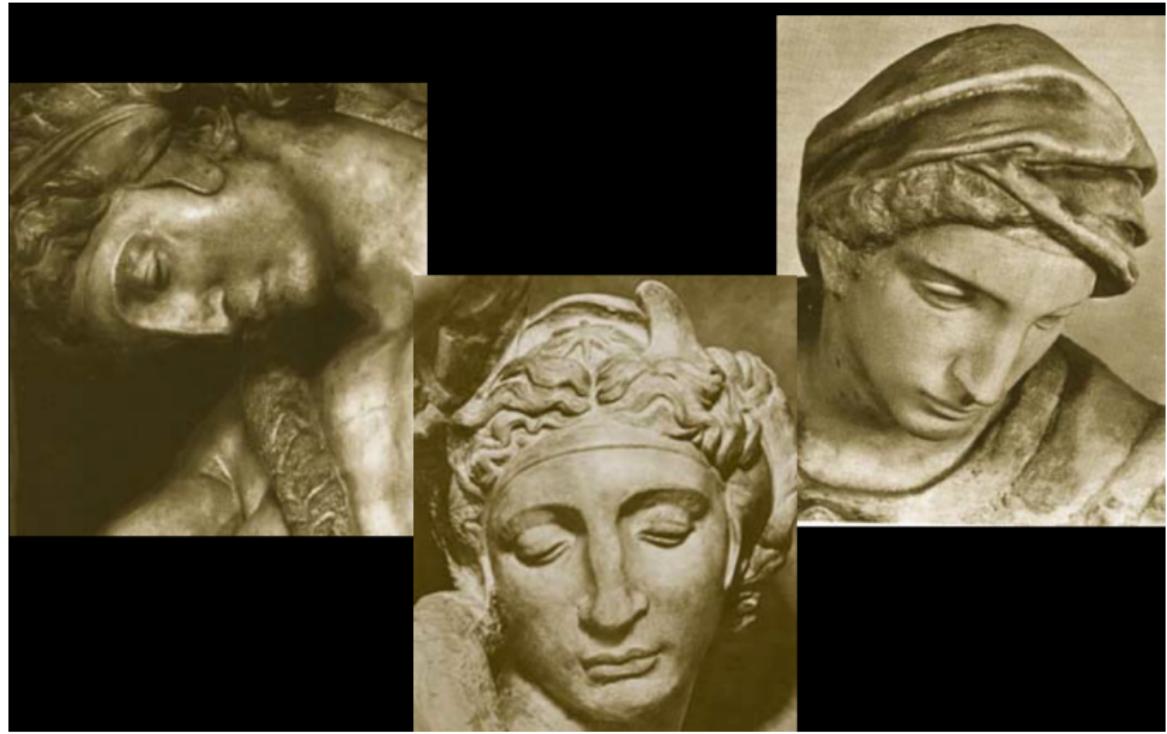
ImageNet Large Scale Visual Recognition Challenges



<http://www.image-net.org/>

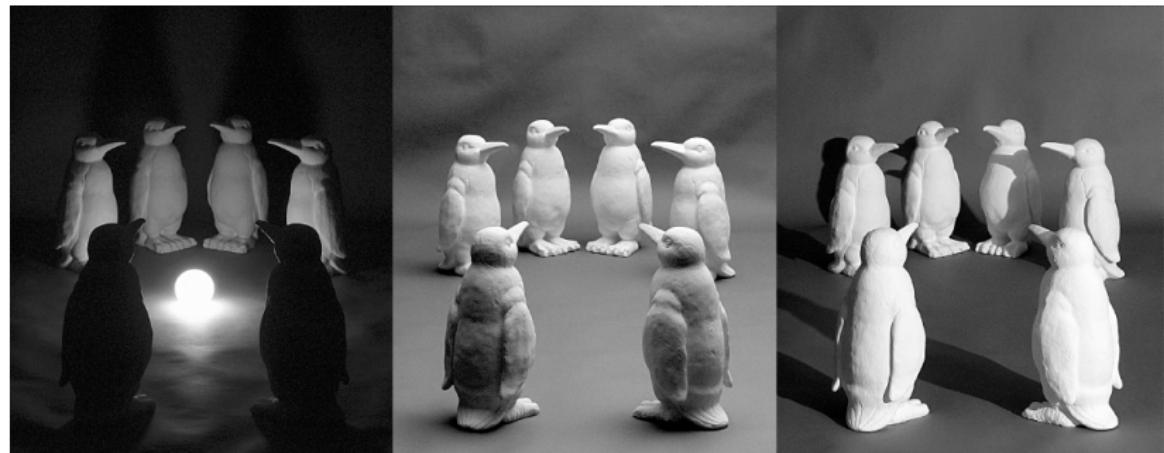
La reconnaissance : un problème complexe

Variation de points de vue



La reconnaissance : un problème complexe

Variation de conditions d'éclairage



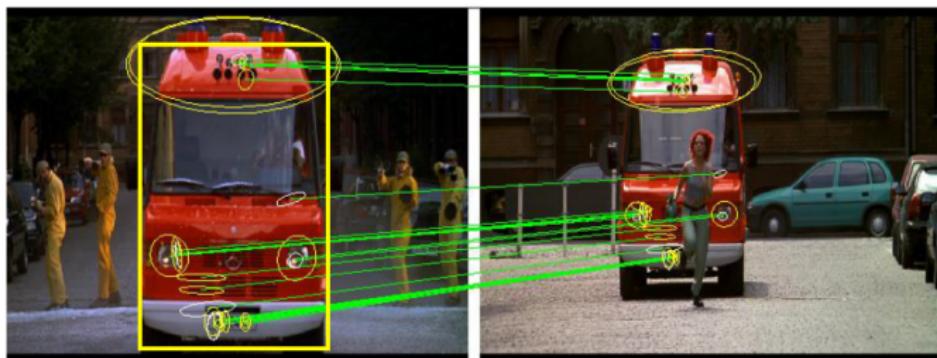
La reconnaissance : un problème complexe

Variation de l'échelle



La reconnaissance : un problème complexe

Occlusion



La reconnaissance : un problème complexe

Variété intra-classe



La reconnaissance : un problème complexe

Une réalité complexe



La reconnaissance : un problème complexe

Quel est le niveau de la reconnaissance ?

Car is an object composed of:

a few doors, four wheels (not all visible at all times), a roof,
front lights, windshield



If you are thinking in buying a car, you might want to be a bit more specific about your categorization.

La reconnaissance : un problème résolu ?

Résultat pour la tâche de classification pour challenge ILSVRC (1000 categories, 1000 images par classe pour l'entraînement, 100k images pour le test)

Classification Results (CLS)



S.o.t.a : (Hu et al, 2018) Squeeze-and-Excitation Networks, CVPR 2018
<https://arxiv.org/abs/1709.01507>

La reconnaissance : un problème résolu ?

Des modèles de plus en plus précis et de plus en plus complexes

ImageNet Challenge

IMAGENET

- 1,000 object classes (categories).
- Images:
 - 1.2 M train
 - 100k test.



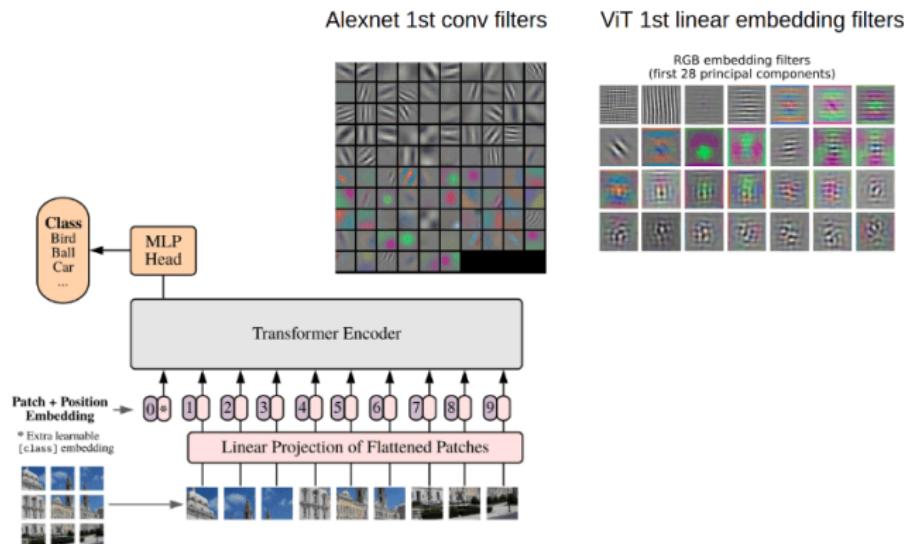
Source : <https://paperswithcode.com/sota/image-classification-on-imagenet>

(Yu et al, 2022) CoCa : Contrastive Captioners are Image-Text Foundation Models : une architecture encoder-decoder entraînée avec une **contrastive loss** et une (generative) captioning loss.

Utilise (ViT) transformer pré-entraîné sur ImageNet-21K¹ (14,197,122 images, divisées en 21,841 classes).

1. <https://arxiv.org/pdf/2104.10972.pdf>

La reconnaissance : un problème résolu ?



(Dosovitskiy et al, 2021) AN IMAGE IS WORTH 16X16 WORDS : TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE²

2. <https://arxiv.org/pdf/2010.11929.pdf>

La reconnaissance : pourquoi ce n'est pas un problème résolu !

- Les modèles état de l'art sont souvent des modèles entraînés de matière supervisée : **nécessite beaucoup de données labellisées.**
 - ▶ Couteux en temps, en argent, travail fastidieux, souvent fait à base de crowdsourcing.
 - ▶ Pas possible pour des données propriétaires.
 - ▶ Pour certaines tâches, les données ne sont tout simplement pas disponibles !

Justifie le fort intérêt pour l'apprentissage avec peu de données.

Retour sur le problème de classification



Synset: mushroom

Definition: any of various fleshy fungi of the subdivision Basidiomycota consisting of a cap at the end of a stem arising from an underground mycelium.

Popularity percentile: 84%

Depth in WordNet: 7



Synset: mushroom

Definition: mushrooms and related fleshy fungi (including toadstools, puffballs, morels, coral fungi, etc.).

Popularity percentile: 82%

Depth in WordNet: 8



Synset: mushroom

Definition: fleshy body of any of numerous edible fungi.

Popularity percentile: 82%

Depth in WordNet: 6



Synset: stuffed mushroom

Definition: mushrooms stuffed with any of numerous mixtures of e.g. meats or nuts or seafood or spinach.

Popularity percentile: 69%

Depth in WordNet: 8



Synset: mushroom sauce

Definition: brown sauce and sauteed mushrooms.

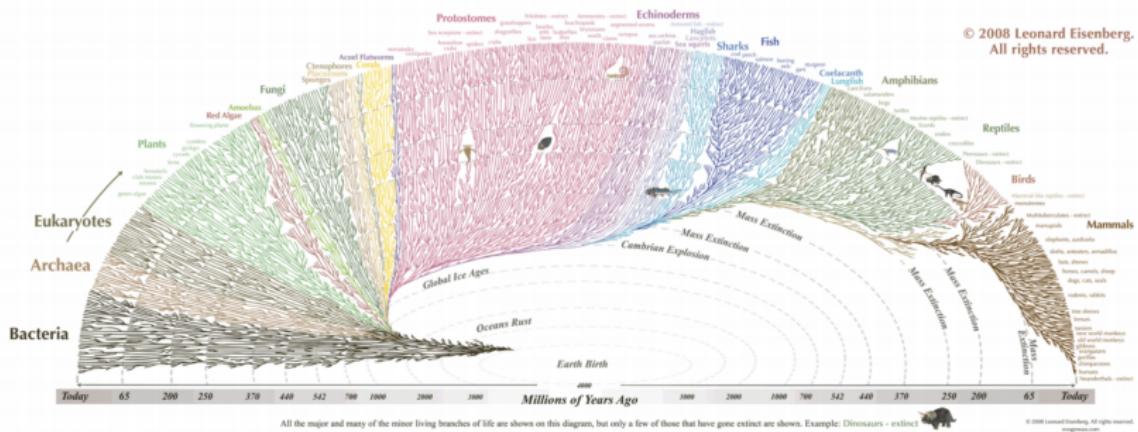
Popularity percentile: 69%

Depth in WordNet: 9

ImageNet has 30 mushroom synsets, each with \approx 1000 images.

Slide credit: Christoph Lampert

Retour sur le problème de classification



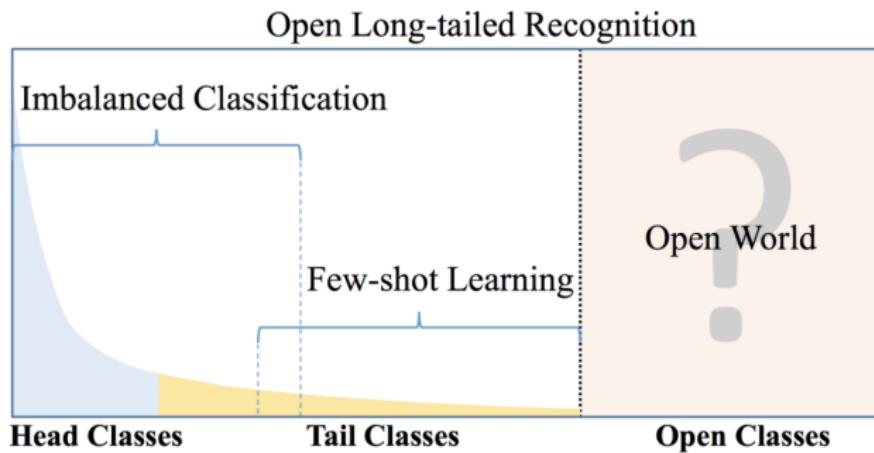
In nature, there are ≈14,000 mushroom species !

Many fine-grained visual categorization tasks may have classes with few or no training examples at all.

Slide credit: Christoph Lampert

Disponibilité de données avec labels

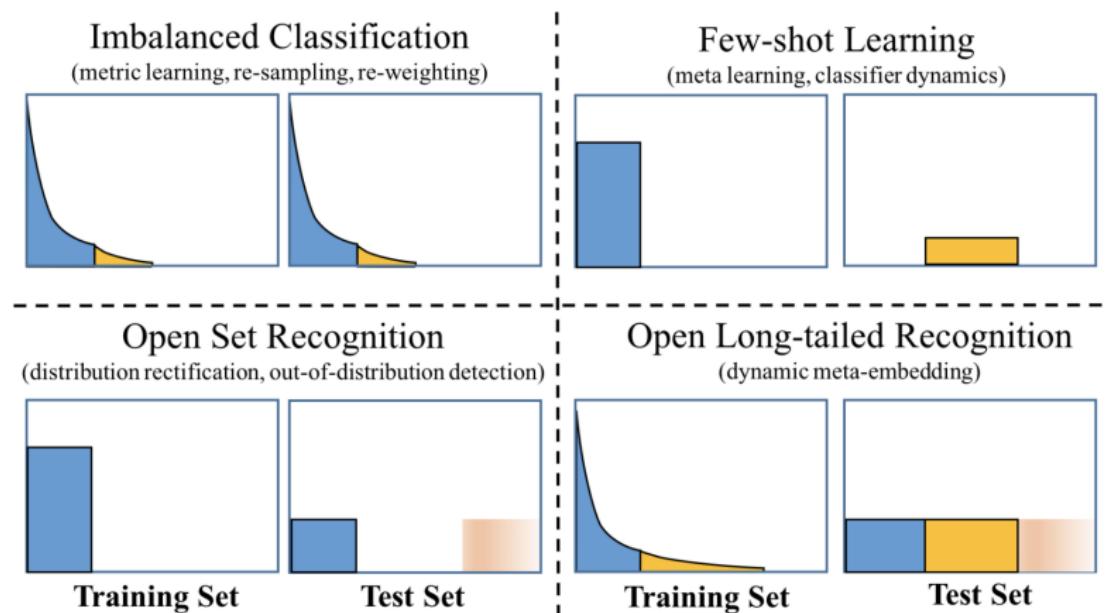
Reconnaissance à grande échelle, longue traîne dans un monde ouvert : existants versus scénarios réels.



Source : <https://bair.berkeley.edu/blog/2019/05/13/oltr/>

Disponibilité de données avec labels

Reconnaissance à grande échelle, longue traîne dans un monde ouvert : existants versus scénarios réels.



La reconnaissance : et pour les humains ?

Etudes cognitives sur la manière dont les humains catégorisent

- Rosch, Cognitive Psychology [Rosch] : basic-level categories.
- Jolicoeur, Kosslyn *et al*, Cognitive Psychology 1984 : entry-level categories.

Un petit test sur deux images



La reconnaissance : et pour les humains ?



- Superordinate : animal, vertebrate
- Basic level : bird
- Entry level : bird
- Subordinatel : American robin



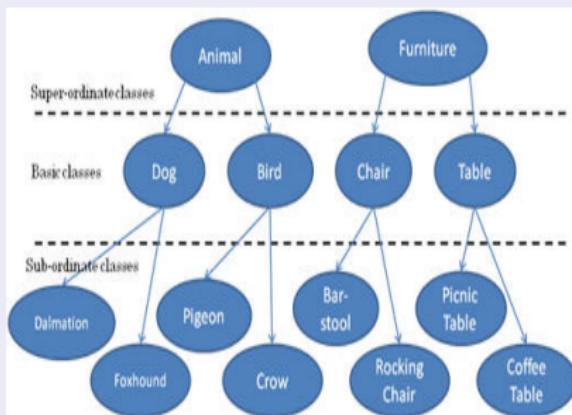
- Superordinate : animal, vertebrate
- Basic level : bird
- Entry level : penguin
- Subordinatel : Chinstrap penguin

Entry-level et basic-level : principalement utilisés par les humains pour décrire les objets.

La reconnaissance : un problème complexe

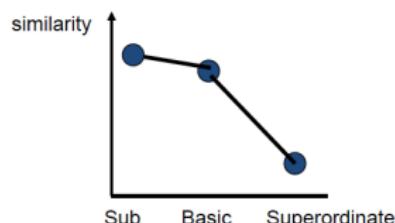
Quel est le niveau de la reconnaissance ?

Les niveaux de catégorisation de Rosch



Niveau basique :

- Même forme
- Même interactions motrices.
- Attributs communs.



Rosch, E. (1973). Natural categories. Cognitive Psychology.

Rosch, E., Mervis, C., Gray, W., Johnson, D., & Boyes-Braem, P. (1976). Basic objects in natural categories. Cognitive Psychology.

La reconnaissance : un problème complexe

Quel est le niveau de la reconnaissance ?

Rosch *et al.* ont montré que :

- Les gens tendent à prédire les catégories basiques (*dog*) avant les catégories plus génériques (*animal*) ou plus spécifiques (*golden retriever*).
- Les gens peuvent dire si un objet appartient à une catégorie basique plus vite.

Des travaux qui ont inspiré la vision par ordinateur

Comment reconnaître les objets au niveau basique ?

- Ordonez et al : entry-level categories[Ordonez et al, 13,15]



Recognition Prediction

grampus griseus



What should I Call It?

dolphin

- Deng et al : optimiser précision et spécificité [Deng et al,12]

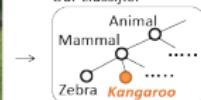
"Easy" image



Conventional classifier



Our classifier



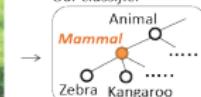
"Hard" image



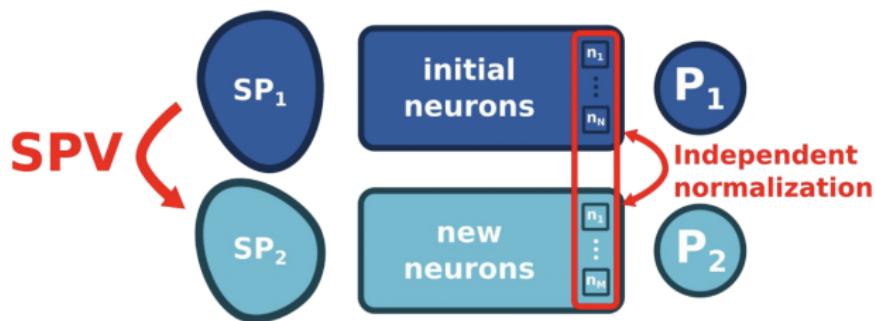
Conventional classifier



Our classifier



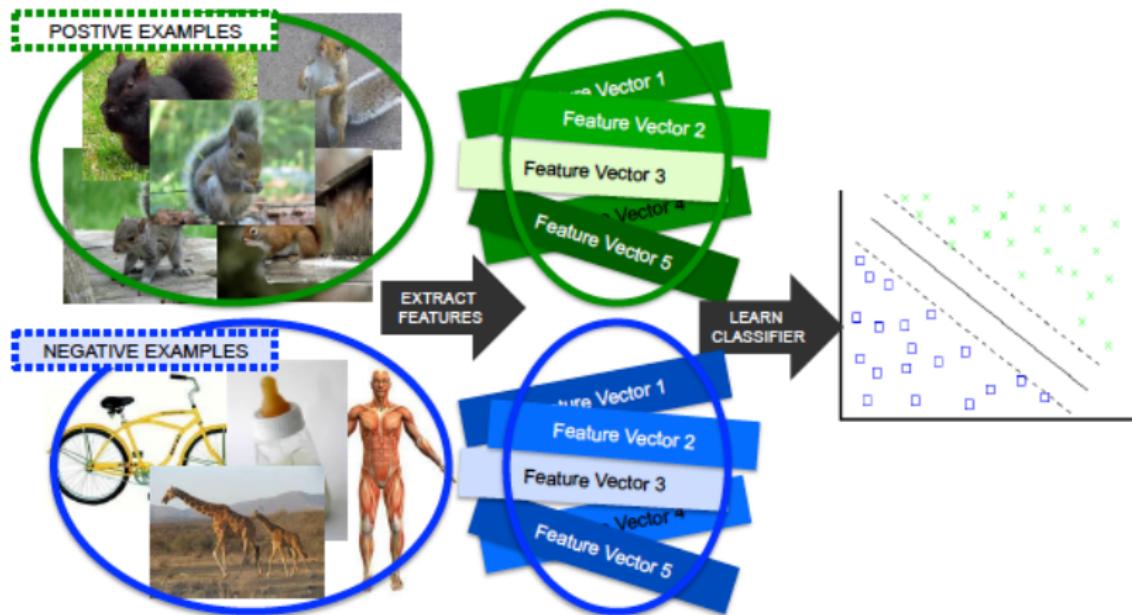
Des travaux qui ont inspiré la vision par ordinateur



- **Source Problem Variation (SPV)**
- **Train new neurons**
- **Representation**
 - Independent normalization
 - Combination (concatenation) + dimensionality reduction (FSFT)

Travaux avec Y. Tamaazousti et H. Le Borgne

La reconnaissance : vue générale de la chaîne de traitements



La reconnaissance : vue générale de la chaîne de traitements

Les principales questions

- Quelles caractéristiques ? Quels descripteurs ? (c.f. cours précédent)
- Quels modèles ?
- Comment modéliser le problème de reconnaissance ?

La reconnaissance : famille d'approches

Les différents modèles

Bag of words models



Csurka, Dance, Fan, Willamowski, and Bray 2004
Sivic, Russell, Freeman, Zisserman, ICCV 2005

Voting models



Viola and Jones, ICCV 2001
Heisele, Poggio, et. al., NIPS 01
Schniederma, Kanade 2004
Vidal-Naquet, Ullman 2003

Shape matching Deformable models



Berg, Berg, Malik, 2005
Cootes, Edwards, Taylor, 2001

Constellation models



Fischler and Elschlager, 1973
Burl, Leung, and Perona, 1995
Weber, Welling, and Perona, 2000
Fergus, Perona, & Zisserman, CVPR 2003

Rigid template models

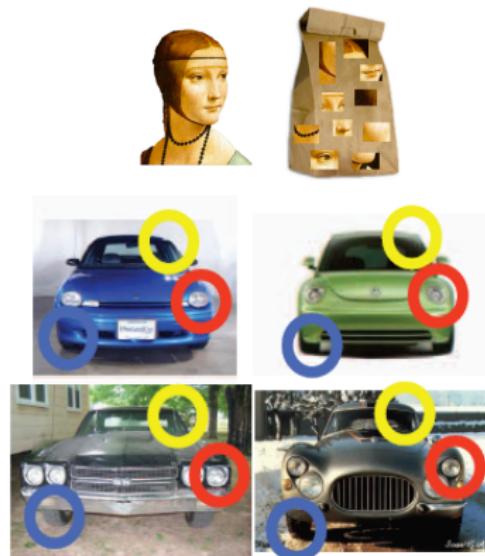


Sirovich and Kirby 1987
Turk, Pentland, 1991
Dalal & Triggs, 2006

La reconnaissance : famille d'approches

Choix du modèle

- Le modèle capture les relations spatiales entre les caractéristiques à des degrés divers.
 - Les objets = un ensemble non-ordonné de caractéristiques (ou avec des contraintes spatiales très faibles)
 - Les objets = un ensemble de parties contraintes spatialement.



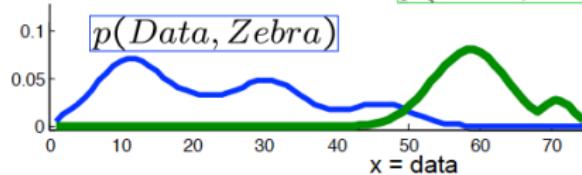
Les différents modèles co-existent et bénéficient les uns aux autres

La reconnaissance : famille d'approches

Approches génératives versus discriminatives

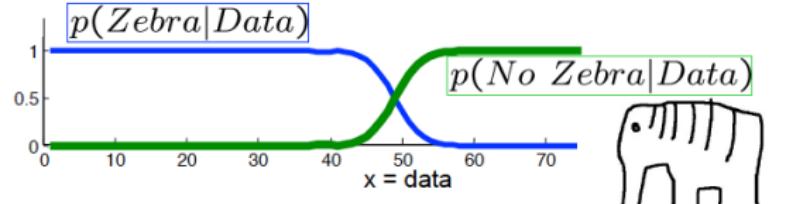
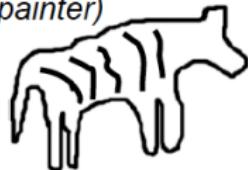
- Generative model

(The artist)



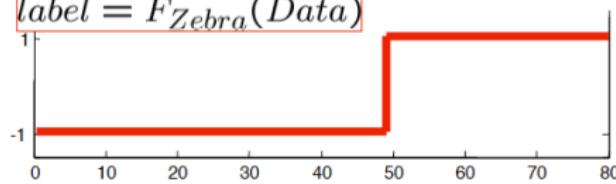
- Discriminative model

(The lousy painter)



- Classification function

$\text{label} = F_{\text{Zebra}}(\text{Data})$



La reconnaissance : Approches génératives versus discriminatives

Modélisation statistique du problème de reconnaissance visuelle.



$$p(\text{zebra}|\text{image})$$

vs

$$p(\text{nozebra}|\text{image})$$

Règle de Bayes

$$\frac{p(\text{zebra}|\text{image})}{p(\text{nozebra}|\text{image})} = \frac{p(\text{image}|\text{zebra})}{p(\text{image}|\text{nozebra})} \cdot \frac{p(\text{zebra})}{p(\text{nozebra})}$$

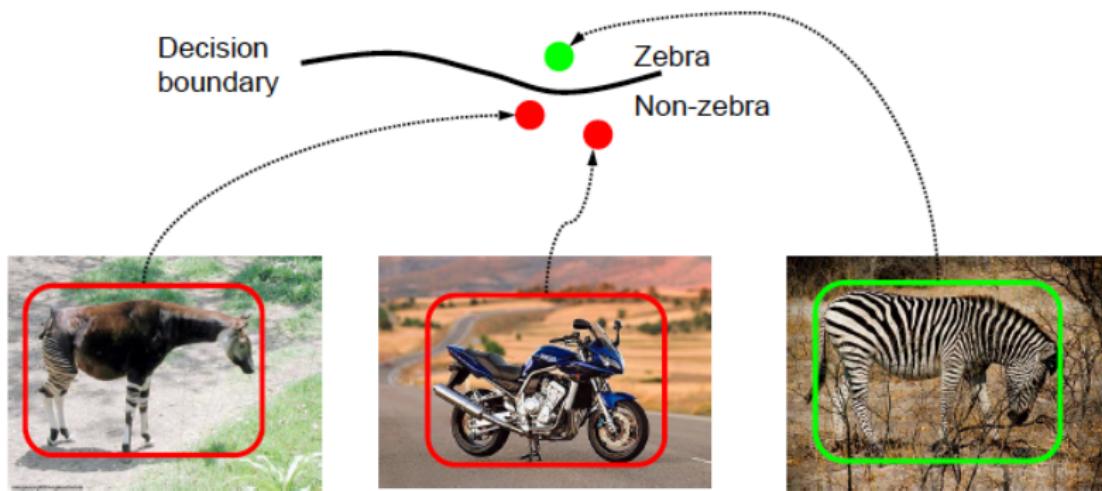
La reconnaissance : Approches génératives versus discriminatives

$$\frac{p(\text{zebra}|\text{image})}{p(\text{nozebra}|\text{image})} = \frac{p(\text{image}|\text{zebra})}{p(\text{image}|\text{nozebra})} \cdot \frac{p(\text{zebra})}{p(\text{nozebra})}$$

- Approches discriminatives : modélisation de l'a posteriori
- Approches génératives : modélisation de la vraisemblance et de l'a priori

La reconnaissance : Approches discriminatives

Modélisation directe de $\frac{p(\text{zebra}|\text{image})}{p(\text{nozebra}|\text{image})}$



La reconnaissance : Approches génératives

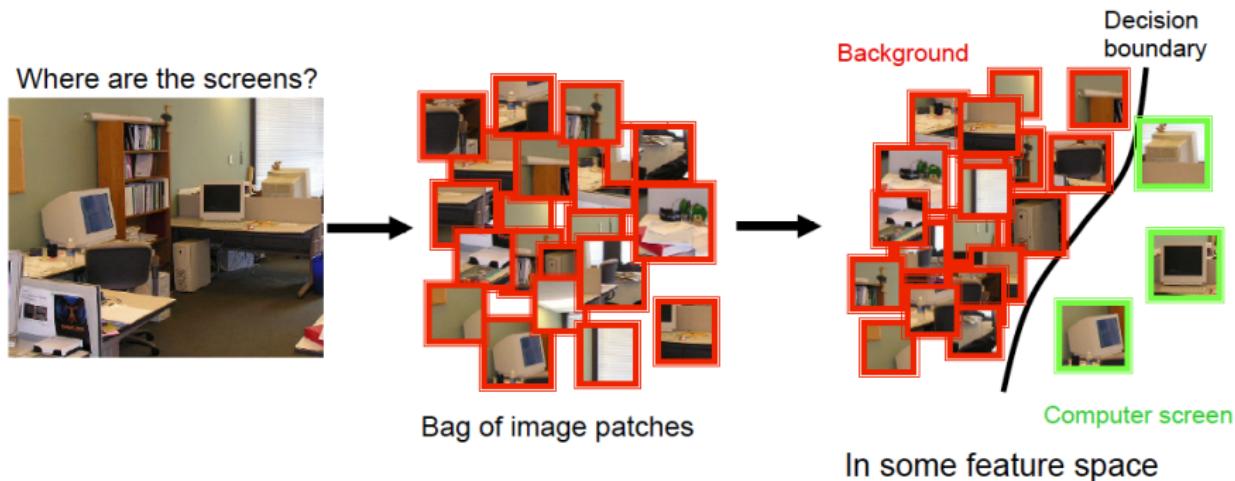
Modélisation de $p(\text{image} | \text{zebra})$ et $p(\text{image} | \text{no zebra})$



$p(\text{image} \text{zebra})$	$p(\text{image} \text{no zebra})$
Low	Middle
High	Middle → Low

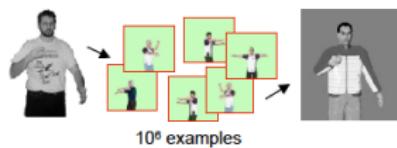
Approches discriminatives

La détection et la reconnaissance d'objets sont souvent formulées comme des problèmes de classification.



Approches discriminatives

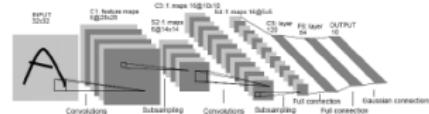
Nearest neighbor



Shakhnarovich, Viola, Darrell 2003
Berg, Berg, Malik 2005

...

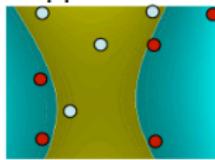
Neural networks



LeCun, Bottou, Bengio, Haffner 1998
Rowley, Baluja, Kanade 1998

...

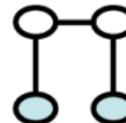
Support Vector Machines and Kernels



Guyon, Vapnik
Heisele, Serre, Poggio, 2001

...

Conditional Random Fields

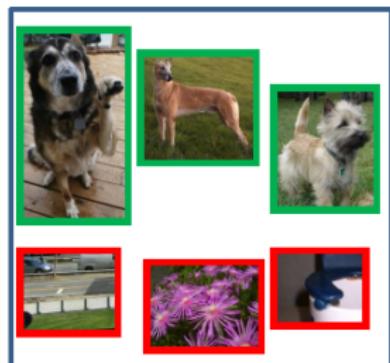


McCallum, Freitag, Pereira 2000
Kumar, Hebert 2003

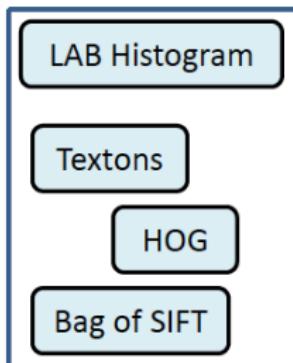
...

La reconnaissance d'objet : approches discriminatives

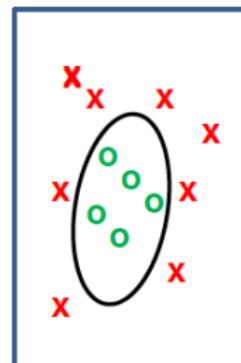
Un problème souvent traité comme un problème d'apprentissage supervisé.



Examples



+ Image Features



+ Classifier

= Category label

Reconnaissance : formulation

- Données (cas d'une classification binaire)

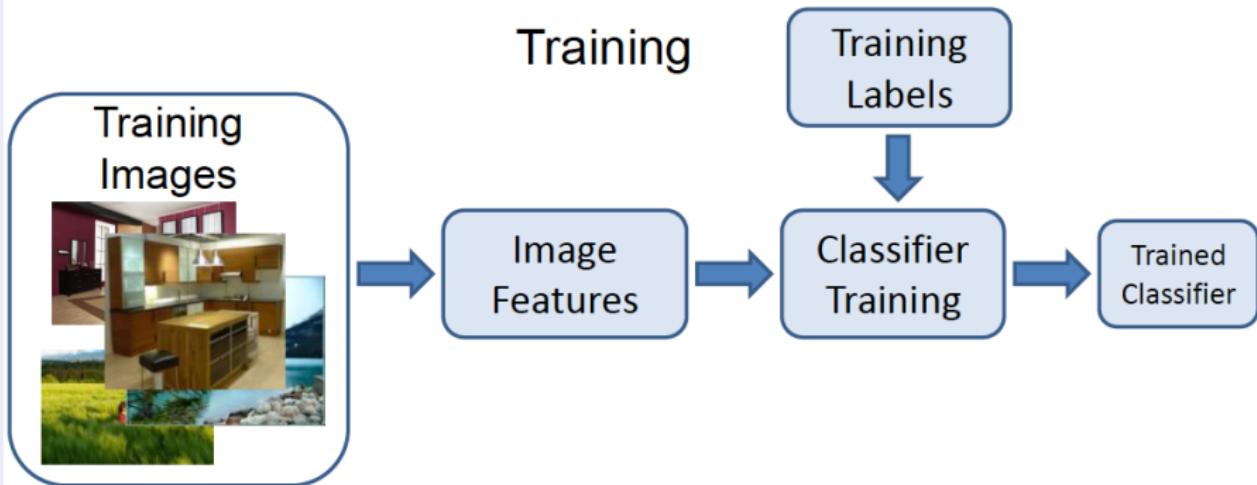
				...				
Features $x =$	X_1	X_2	X_3	...	X_N	X_{N+1}	X_{N+2}	$\dots X_{N+M}$
Labels $y =$	-1	+1	-1		-1	?	?	?

Training data: each image patch is labeled as containing the object or background Test data

- Fonction de décision : $y = F(x)$
- Objectif : minimiser le taux d'erreur de classification.

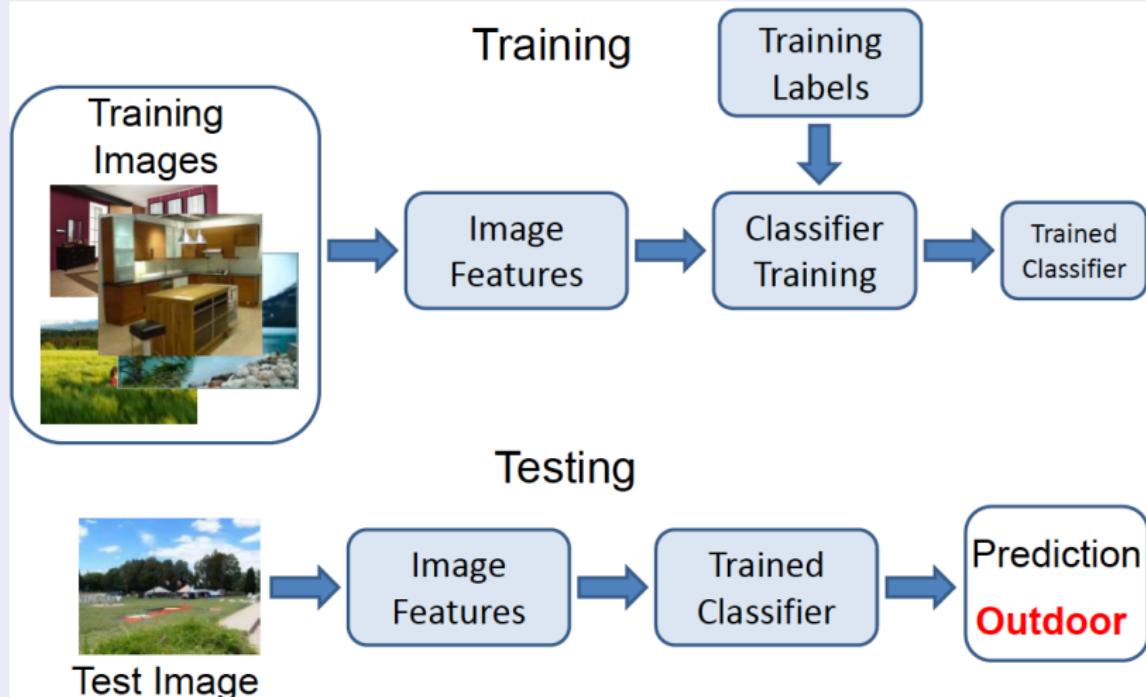
Reconnaissance : formulation comme un problème de classification

Etape d'apprentissage



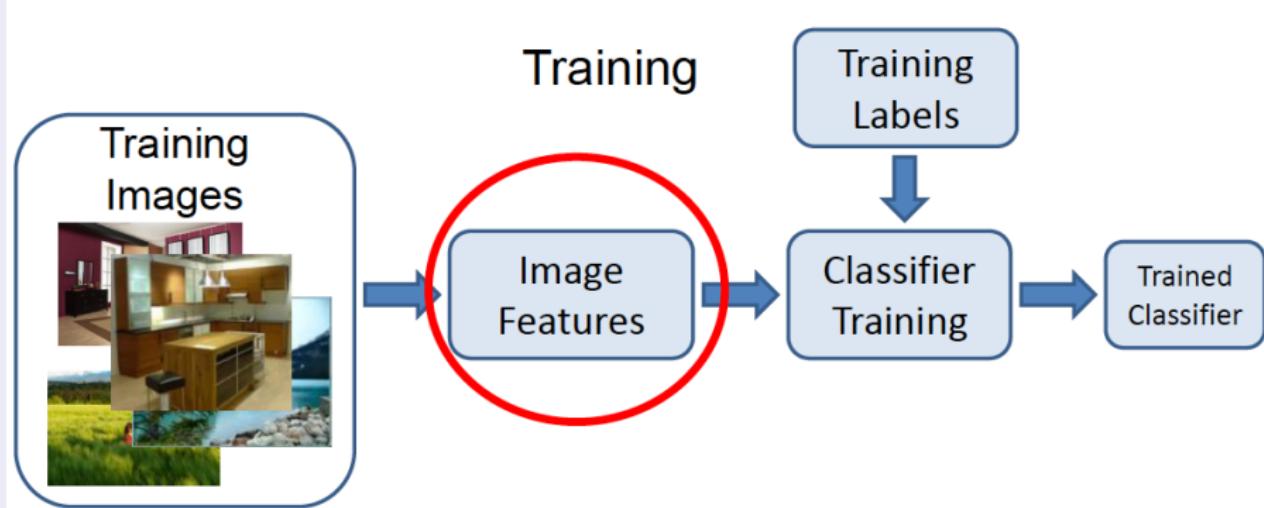
Reconnaissance : formulation comme un problème de classification

Etape de test



Reconnaissance : formulation comme un problème de classification

Quels descripteurs ?



Plan

1 Introduction

2 Description d'images

- Descripteurs basiques
- SIFT
- Gist

3 Petite parenthèse : problème du panorama

- Bow

4 Classification

5 Conclusion

Recap

Application : création de panorama



FIGURE – Source : Daria Frolova

- Détection de points caractéristiques dans chaque image.

Recap

Problème à résoudre 1

Déetecter le même point de manière indépendante dans les deux images.



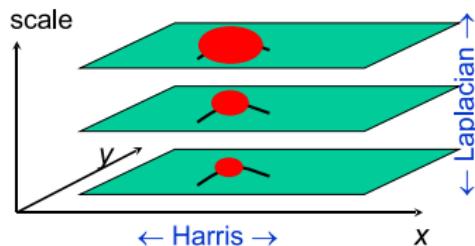
no chance to match!

FIGURE – Source : Daria Frolova

⇒ : besoin d'un détecteur fiable.

Recap : DéTECTEURS INVARIANTS EN ÉCHELLE

Harris-Laplace



K.Mikolajczyk, C.Schmid. Indexing Based on Scale Invariant Interest Points. ICCV 2001

Principe

Trouver le maximum local pour :

- Les coins de Harris dans l'espace.
- Le laplacien en échelle.

Recap

Application : création de panorama



FIGURE – Source : Daria Frolova

- Détection de points caractéristiques dans chaque image : **nous sommes capables de détecter des zones robustes (et informatives) de l'image.**
- Chercher les paires correspondantes.
- Utiliser les paires pour l'alignement.

Recap

Problème à résoudre 2

Pour chaque point trouver la bonne correspondance.



FIGURE – Source : Daria Frolova

⇒ : besoin d'un descripteur fiable et discriminant.

Descripteurs de points d'intérêts : descripteurs locaux

Comment représenter la région au voisinage du point d'intérêt ?



Vos idées ?

Descripteurs de points d'intérêts : descripteurs locaux

Qu'est ce qu'un descripteur ?

- Représentation de la zone d'intérêt.
- Sous la forme d'un vecteur.
- Appartenant à un espace muni d'une distance.

Qu'est ce qu'un bon descripteur ?

- Invariant.
- Discriminant.
- Compact.

Motivation : pourquoi calculer des descripteurs ?

A la base de nombreuses applications

- Recalage d'images (alignement).
- Reconstruction 3D.
- Suivi du mouvement.
- Reconnaissance d'objets et de scènes.
- Indexation et recherche d'images.
- Navigation (robotique)
- Classification
- ...

Les différents types de descripteurs

De nombreux types de descripteurs

- Descripteurs basiques.
- Descripteurs invariants à la luminosité et à la rotation : invariants différentiels.
- Descripteurs à base de moments couleurs affine-invariants.
- Descripteur à base d'histogrammes : SIFT, HOG, ...
- ...

Choix de descripteurs

Les « bons » descripteurs dépendent de notre objectif :

- Objet :
 - ▶ SIFT, HOG, color, Bow.
- Scène :
 - ▶ GIST, Bow, color.
- Propriétés sur les matériaux :
 - ▶ Color, texture
- Mouvement.
 - ▶ Flot optique, ...

Attention

Le descripteur extrait l'information du média brut (pixels).

→ il définit l'espace dans lequel le classifieur cherche à discriminer

Importance de l'espace de représentation.

Plan

1 Introduction

2 Description d'images

- Descripteurs basiques
- SIFT
- Gist

3 Petite parenthèse : problème du panorama

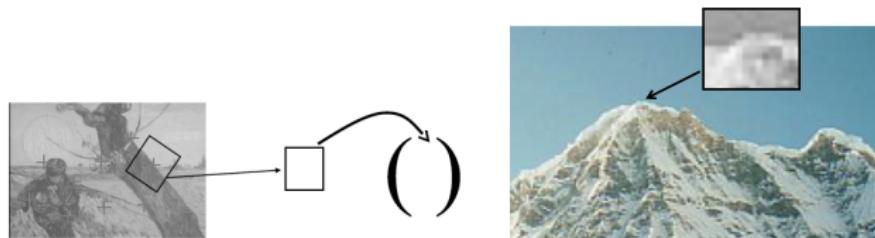
- Bow

4 Classification

5 Conclusion

Descripteur pixel

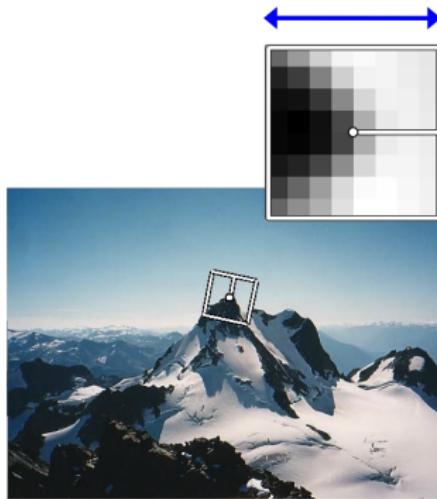
- On range les pixels au voisinage du point d'intérêt dans un vecteur (descripteur local)



- Calcul simple.
- Volumineux.
- Utilisé comme *baseline* en classification.

Est-ce un bon descripteur ?

Descripteur pixel



- Rotation de la fenêtre selon l'orientation du point d'intérêt.
- Prise en compte de l'échelle adaptée.

Mais reste encore très sensible aux petites translations ou aux changements d'intensité.

Invariants différentiels

- Descripteurs de points
- Basés sur des dérivées calculées en un point (x, y) .

$$v(x, y) = \begin{bmatrix} I(x, y) \\ I_x(x, y) \\ I_y(x, y) \\ I_{xx}(x, y) \\ I_{yy}(x, y) \\ \vdots \\ \vdots \end{bmatrix}$$

Plan

- 1 Introduction
- 2 Description d'images
 - Descripteurs basiques
 - SIFT
 - Gist
- 3 Petite parenthèse : problème du panorama
 - Bow
- 4 Classification
- 5 Conclusion

SIFT : Scale Invariant Feature Transform

[Lowe,04] Distinctive Image Features from Scale-Invariant Keypoints, IJCV
<http://www.cs.ubc.ca/~lowe/papers/ijcv04.pdf>

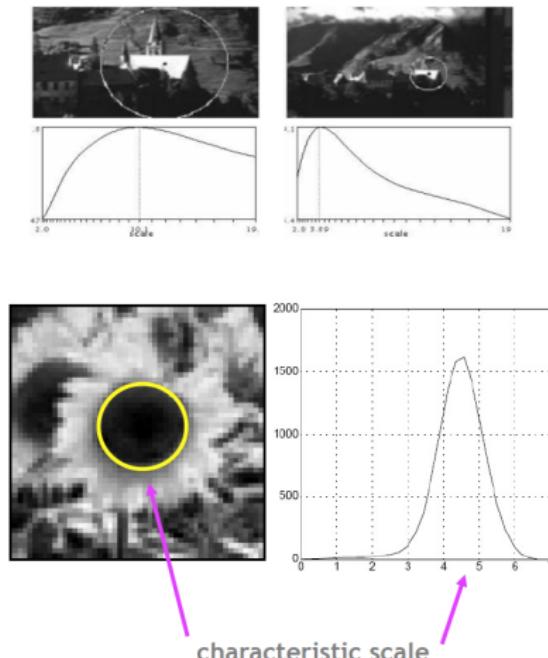
4 étapes

- ① Etape 1 : **Scale-space extrema Detection** : détection de points d'intérêts invariants en échelle et en orientation à l'aide des DoG.
- ② Etape 2 : **Keypoint Localization** : determination de la position et de l'échelle de chaque point d'intérêt candidat en se basant sur un critère de stabilité.
- ③ Etape 3 : **Orientation Estimation** : Utilisation des gradients locaux pour affecter une orientation à chaque point d'intérêt.
- ④ Etape 4 : **Descripteur de points d'intérêts** : Extraction des gradients locaux à l'échelle sélectionnée dans un voisinage du point d'intérêt et construction d'un descripteur invariant aux transformations locales et aux changements d'illuminations.

SIFT : Scale-space extrema detection (étape 1)

Rappel : sélection de l'échelle

- Objectif : selection pour invariance à l'échelle.
- L'échelle caractéristique est celle qui produit un pic dans la réponse du laplacien (extrema du laplacien).
- SIFT : approximation du laplacien par une différence de gaussiennes.

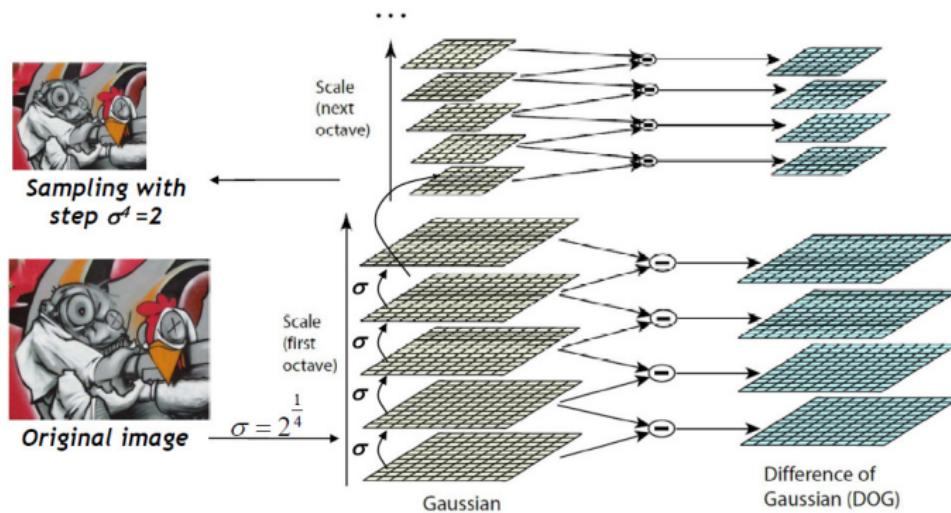


SIFT : Scale-space extrema detection (étape 1)

Points d'intérêts = extrema locaux de différences de gaussiennes à différentes échelles.

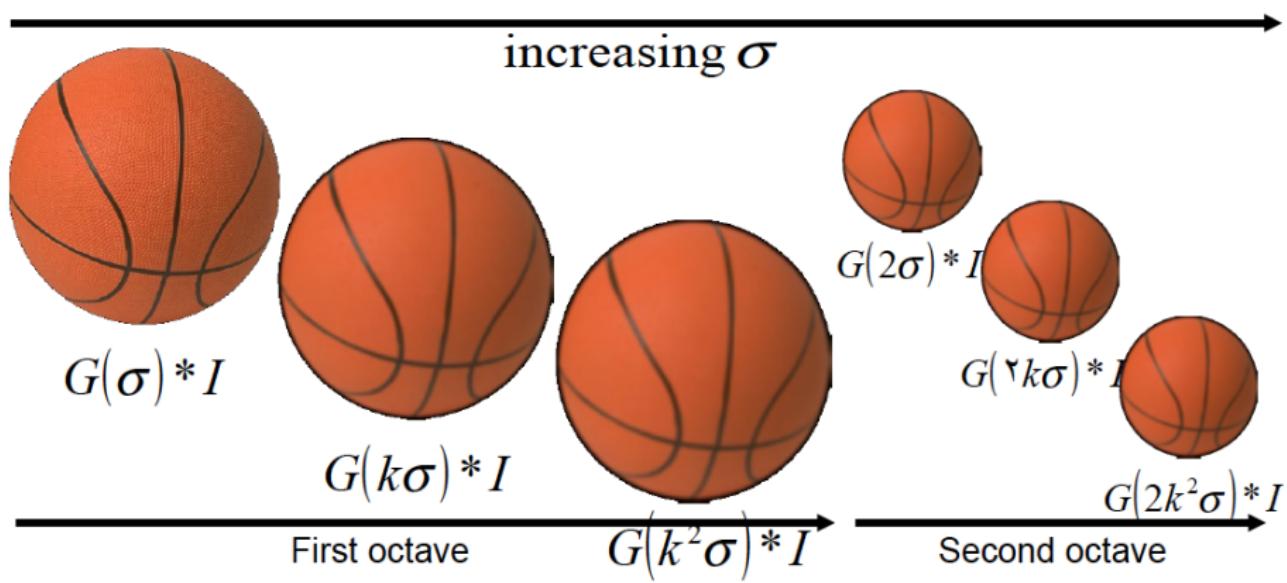
$$D(x, y, \sigma) = L(x, y, k\sigma) - L(x, y, \sigma)$$

avec $L(x, y, \sigma) = G(x, y, \sigma) * I(x, y)$ et $G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} \exp^{-\frac{(x^2+y^2)}{\sigma^2}}$
 (objets observables dans des facteurs d'échelle entre σ et $k\sigma$)



SIFT : Scale-space extrema detection (étape 1)

- Les images des DoG sont groupées par octave ($\sigma_0 \times 2$)
- Nombre fixé de niveaux par octave.

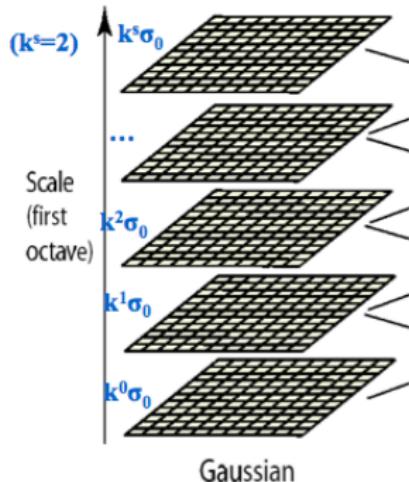


SIFT : Scale-space extrema detection (étape 1)

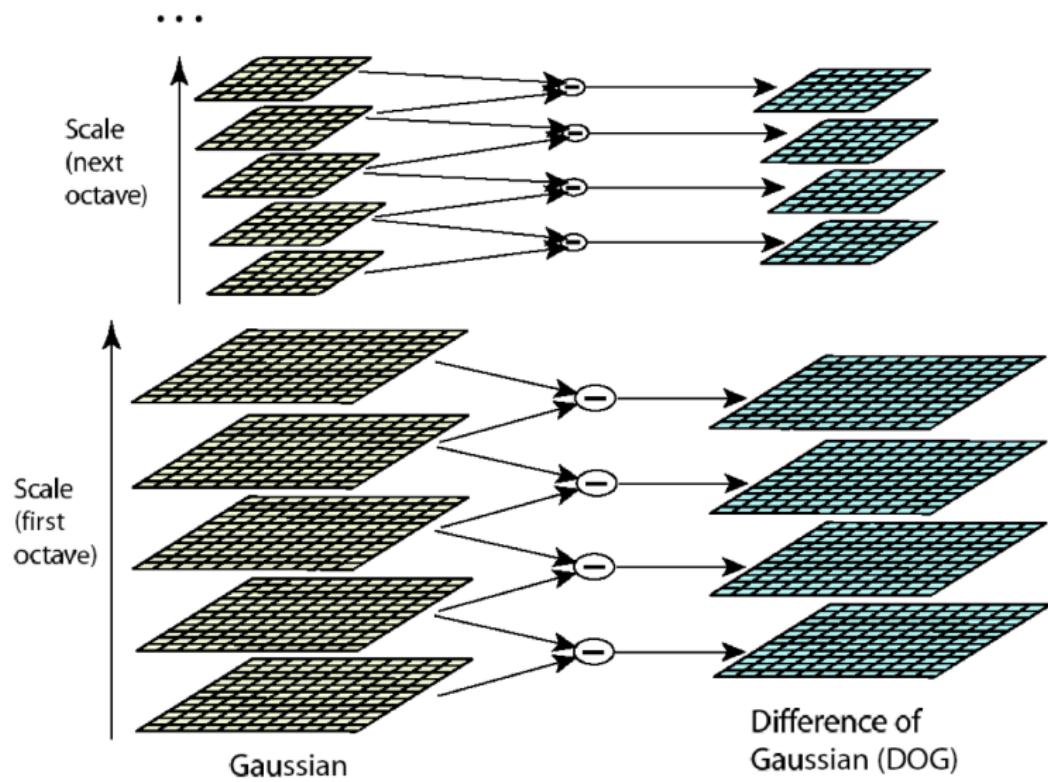
- Dans chaque octave, les images sont séparées par un facteur constant k .
- Si chaque octave est séparé en $s - 1$ intervalles :

$$k^s = 2 \text{ ou } k = 2^{\frac{1}{s}}$$

- Dans le papier de Lowe :
 - Nombre d'échelles par octave : 3
 - $\sigma_0 = 1.6$
- Après chaque octave, l'image est sous-échantillonnée d'un facteur de 2 : image de taille 4 fois plus petite



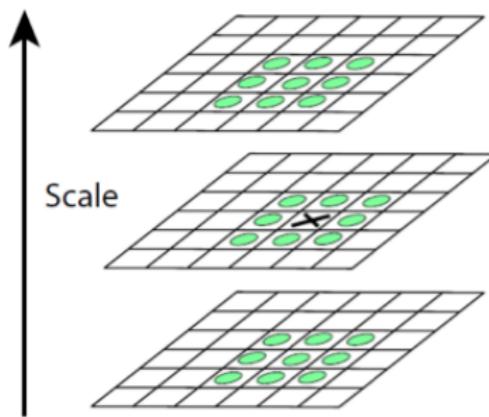
SIFT : Scale-space extrema detection (étape 1)



SIFT : Scale-space extrema detection (étape 1)

Détection des extrema locaux dans l'espace d'échelle.

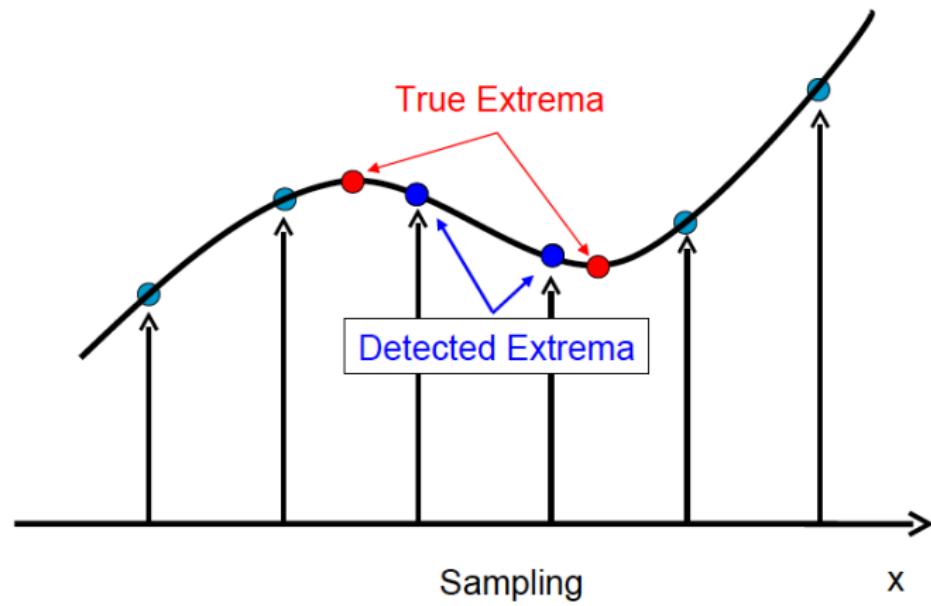
- X est conservé s'il est le plus grand ou le plus petit de tout le voisinage de 26 voisins.



SIFT : Localisation précise des points clés (étape 2)

Problème

Trop de points d'intérêts, certains sont instables.



SIFT : Localisation des points clés (étape 2)

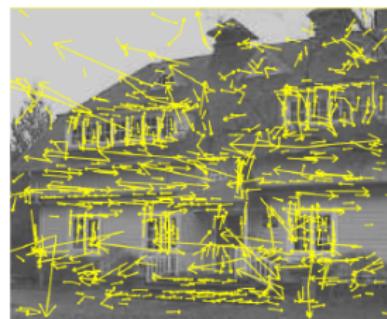
Problème

Trop de points d'intérêts, certains sont instables :

- Points avec peu de contraste (sensibilité au bruit)
- Points situés le long d'une arête.



(a) 233x189 image



(b) 832 DOG extrema

⇒ Elimination des points de faible contraste

SIFT : Localisation des points clés (étape 2)

Solution (pour le faible contraste)

Amélioration de la précision par interpolation des coordonnées.

- Ajuster le point d'intérêt x aux données voisines par une approximation quadratique (développement de Taylor de D , différence de gaussiennes.)

$$D(x) = D + \frac{\partial D^T}{\partial x}x + \frac{1}{2}x^T \frac{\partial^2 D^T}{\partial x^2}x$$

avec $x = (x, y, \sigma)^T$

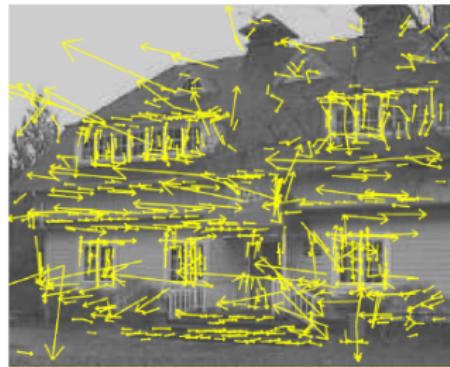
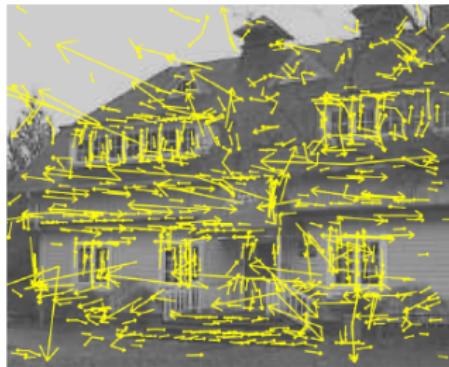
- Origine : coordonnées du point-clé candidat
- Position précise de l'extremum est déterminée en résolvant l'équation annulant la dérivée de cette fonction par rapport à x .

$$\hat{x} = -\frac{\partial^2 D^{-1}}{\partial x^2} \frac{\partial D}{\partial x}$$

- Eliminer un minimum local :

$$D(\hat{x}) < 0.03$$

SIFT : Localisation des points clés (étape 2)



729 de 832 demeurent après seuillage sur le contraste

SIFT : Localisation des points clés (étape 2)

Solution (pour éviter les arêtes)

- Un point d'intérêt sur une arête n'est pas bien localisé.
- Calculer la matrice hessienne :

$$H = \begin{bmatrix} D_{xx} & D_{xy} \\ D_{xy} & D_{yy} \end{bmatrix}$$

avec $\alpha = \lambda_{max}$, plus grande valeur propre et $\beta = \lambda_{min}$ plus petite valeur propre (relative à la courbure).

- $Trace(H) = D_{xx} + D_{yy} = \alpha + \beta$
- $det(H) = D_{xx}D_{yy} - D_{xy}^2 = \alpha\beta$
- $\frac{Trace(H)^2}{det(H)} = \frac{(\alpha+\beta)^2}{\alpha\beta} = \frac{(r+1)^2}{r}$ avec $r = \frac{\alpha}{\beta}$
- Rejeter le point si : $\frac{Trace(H)^2}{det(H)} < \frac{(r+1)^2}{r}$ (SIFT prend $r = 10$).

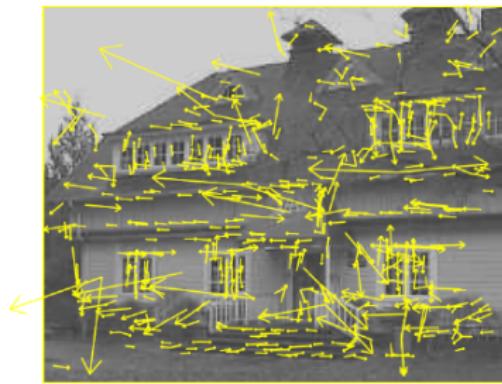
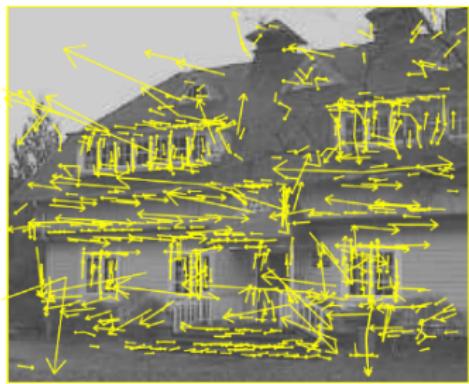
SIFT : Localisation des points clés (étape 2)

Solution (pour éviter les arêtes)

Explication

- Les points d'intérêts situés sur une arête sont sensibles au bruit et devraient être éliminés.
⇒ Vérifier si un point d'intérêt est sur une arête ou sur un coin
- Sur une arête, une des courbures principales est beaucoup plus grande que l'autre.
- Sur un coin, pas de courbure principale dominante.
- Les courbures principales sont proportionnelles aux valeurs propres de la matrice Hessienne.
- ...

SIFT : Localisation des points clés (étape 2)



536 de 729 demeurent après détection de coin/arête

SIFT : Detection de l'orientation (étape 3)

Ici, nous pouvons parler de DESCRIPTEUR.

On a un ensemble de points : **de bons points**. On choisit une région autour de chaque point.

- Objectif : supprimer les effets d'échelles et de rotation



SIFT : Detection de l'orientation (étape 3)

- L'orientation du point d'intérêt dépend des propriétés locales de l'image autour du point d'intérêt.
- Choisir le niveau σ sur la pyramide correspondant à l'échelle du point d'intérêt.

$$L(x, y) = G(x, y, \sigma) * I(x, y)$$

- Calculer la magnitude et l'orientation du gradient en utilisant le calcul des gradients par un masque horizontal $[-1 \ 0 \ 1]$ ($\rightarrow H$) et un masque vertical $[1 \ 0 \ -1]^T$ ($\rightarrow V$)

► Amplitude du gradient : $\sqrt{H^2 + V^2}$

$$m(x, y) = \sqrt{(L(x+1, y) - L(x-1, y))^2 + (L(x, y+1) - L(x, y-1))^2}$$

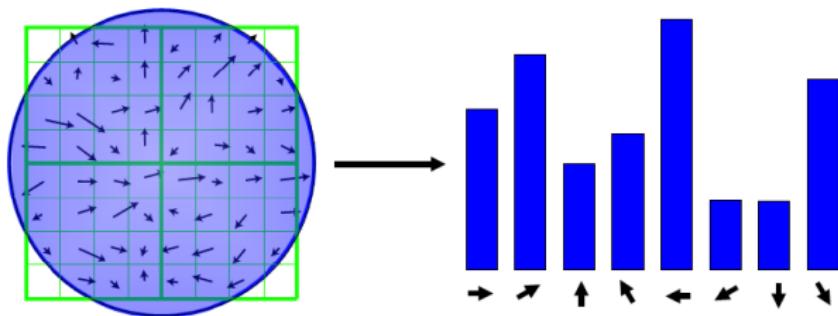
► Direction du gradient : $\arctan(\frac{V}{H})$.

$$\theta(x, y) = \arctan\left(\frac{(L(x, y+1) - L(x, y-1))}{(L(x+1, y) - L(x-1, y))}\right)$$

- Pour chaque (x, y) au voisinage d'un point d'intérêt.

SIFT : Detection de l'orientation (étape 3)

- Construire un histogramme pondéré (36 bins = 10 deg/bin) (par l'amplitude et gaussienne à 1,5 fois l'échelle du point d'intérêt) des directions des gradients locaux calculés à l'échelle du point d'intérêt et dans le voisinage du point d'intérêt.
- Déterminer l'orientation canonique (dominante) = pic de l'histogramme pondéré.
- Le point d'intérêt a plusieurs orientations lorsque l'histogramme présente plusieurs pics.



A l'issue de cette étape, un point clé est donc défini par 4 paramètres (x, y, σ, θ).

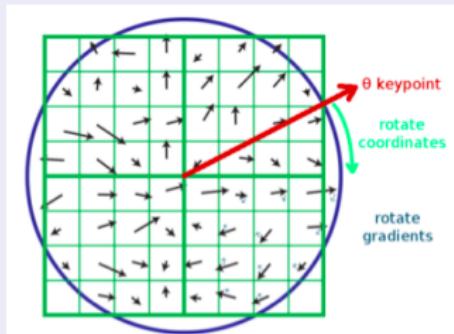
SIFT : Description des points clés (étape 4)

- Une localisation, une échelle et une orientation ont été attribuées à chaque point d'intérêt.
- Il reste maintenant à définir un descripteur local invariant aux variations restantes :
 - ▶ Illumination.
 - ▶ Angle de vue 3D.

SIFT : Description des points clés (étape 4)

Première étape

- Modification du système de coordonnées locales pour garantir l'invariance à la rotation, avec une rotation d'angle égal à l'orientation du point-clé, mais de sens opposé.



- Lissage de l'image avec le paramètre de facteur d'échelle le plus proche de celui du point-clé considéré

SIFT : Description des points clés (étape 4)

Descripteur local

- Chaque point d'intérêt est décrit par ses coordonnées sur l'image, son échelle et son orientation.
- On considère une région de 16×16 pixels, subdivisée en 4×4 zones de 4×4 pixels chacune.
- Sur chaque zone, on calcule un histogramme des orientations comportant 8 intervalles.
- En chaque point de la zone, l'orientation et l'amplitude du gradient sont calculés.
- Descripteur = matrice 4×4 d'histogrammes d'orientations (8 bins) entourant le point d'intérêt.
- La matrice est normalisée pour atténuer les effets de l'éclairage.

SIFT : Description des points clés (étape 4)

Histogrammes pondérés (par l'amplitude et la fenêtre gaussienne) des orientations du gradient.

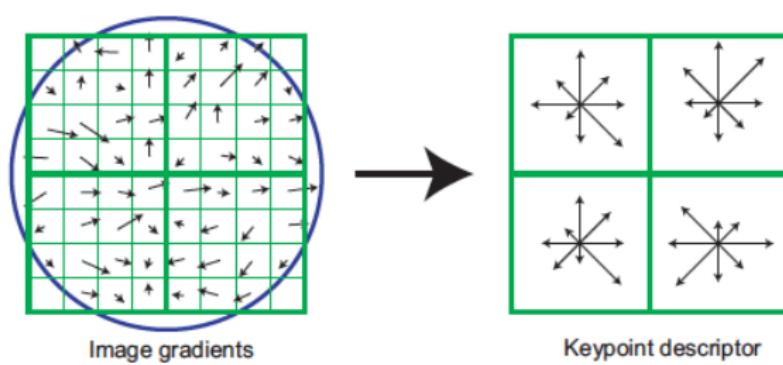
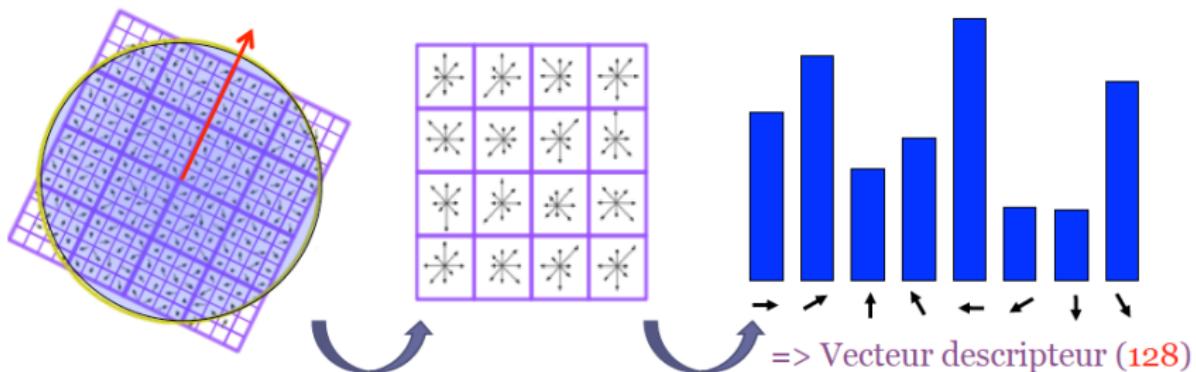


Illustration 8×8 pour un descripteur 2×2 mais dans le papier 16×16 pour un descripteur 4×4

SIFT : Description des points clés (étape 4)

Assembler les 16 histogramme de gradients (8 bins chacun) :

- Valeurs pondérés par l'amplitude du gradient *et* la fenêtre gaussienne.
- Les valeurs de l'histogramme et du gradient sont interpolées et filtrées.



SIFT : Quelques extensions

Couleur

- Color SIFT : Performance Evaluation of Local Colour Invariants G. J. Burghouts, J. M. Geusebroek. *In Computer Vision and Image Understanding*, 2009.
- Hue and Opponent Histograms : Joost van de Weijer, Cordelia Schmid Coloring Local Feature Extraction, *Proc. ECCV06*, Graz, Austria, 2006.
- 11 color names : J. van de Weijer, C. Schmid, Applying Color Names to Image Description , *Proc. ICIP2007*, San Antonio, USA, 2007.

SIFT : Quelques extensions

Compacité du(des) descripteur

- PCA-SIFT : Yan Ke and Rahul Sukthankar. 2004. PCA-SIFT : a more distinctive representation for local image descriptors. *In Proceedings of the 2004 IEEE computer society conference on Computer vision and pattern recognition (CVPR'04)*. IEEE Computer Society, Washington, DC, USA, 506-513.

SIFT : Quelques extensions

Amélioration du temps de calcul

- SURF : Herbert Bay, Andreas Ess, Tinne Tuytelaars, Luc Van Gool, "SURF : Speeded Up Robust Features", *Computer Vision and Image Understanding (CVIU)*, Vol. 110, No. 3, pp. 346–359, 2008
- Approx SIFT : Michael Grabner, Helmut Grabner, and Horst Bischof. 2006. Fast approximated SIFT. In *Proceedings of the 7th Asian conference on Computer Vision - Volume Part I (ACCV'06)*, P. J. Narayanan, Shree K. Nayar, and Heung-Yeung Shum (Eds.), Vol. Part I. Springer-Verlag, Berlin, Heidelberg, 918-927
- GPU implementation : Sudipta N Sinha, Jan-Michael Frahm, Marc Pollefeys and Yakup Genc, "GPU-Based Video Feature Tracking and Matching ", EDGE 2006, *workshop on Edge Computing Using New Commodity Architectures*, Chapel Hill, May 2006

SIFT : Implémentation

- OpenCV : https://docs.opencv.org/4.5.1/da/df5/tutorial_py_sift_intro.html
https://docs.opencv.org/4.5.1/d0/d13/classcv_1_1Feature2D.html
- VL-FEAT : <http://www.vlfeat.org/> : la bibliothèque de référence concernant les détecteurs et descripteurs locaux (en C , interface matlab, python <https://pypi.org/project/pyvlfeat/>).
- ...

SIFT : Implémentation



Descripteurs

De nombreux autres descripteurs

- HOG = Histograms of oriented gradients
 - ▶ Find robust feature set that allows object form to be discriminated.
 - ▶ HoG is usually used to describe larger image regions. SIFT is used for key point.
- DAISY
- Dense SIFT
- MESR : Maximally Stable Extremal Regions
- ...

Plan

1 Introduction

2 Description d'images

- Descripteurs basiques
- SIFT
- Gist

3 Petite parenthèse : problème du panorama

- Bow

4 Classification

5 Conclusion

Relation Objet /Scène



Relation Objet /Scène



Relation Objet /Scène



A propos de l'interprétation d'images

Mary Potter (1975,1976) : Rapid Image Understanding

Les travaux de Mary Potter ont montré que lors d'une présentation rapide d'une séquence d'images (100 ms par image), une nouvelle image est instantanément interprétée et nous pouvons comprendre beaucoup de l'information visuelle



A propos de l'interprétation d'images

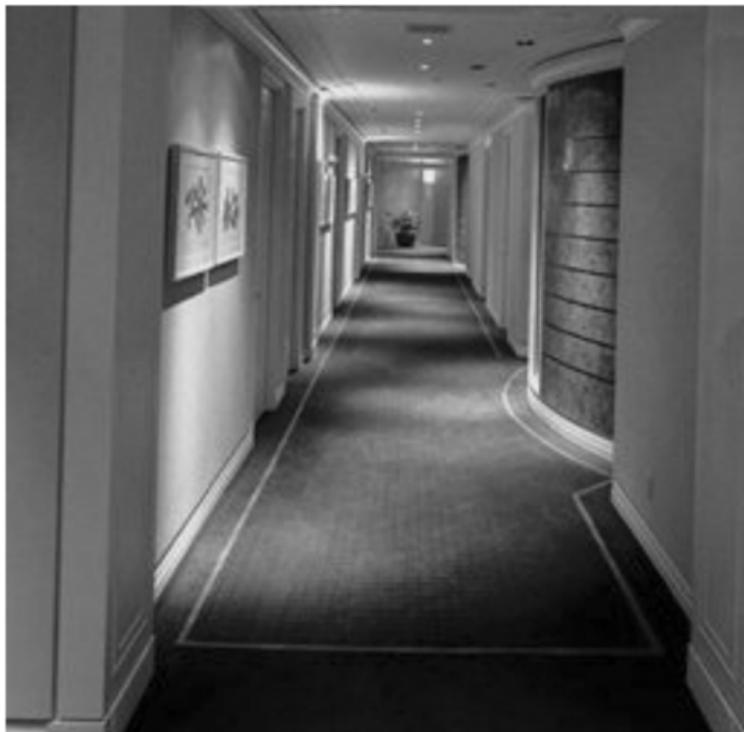
Démonstration

Par Aude Oliva

Vous allez voir rapidement 9 images (0.5s). Vous devez les mémoriser.



















A propos de l'interprétation d'images

Test de mémoire

Parmi les prochaines images, quelles sont les images que vous avez-vu ?



Have you seen this picture ?





Have you seen this picture ?





Have you seen this picture ?





Have you seen this picture ?





Have you seen this picture ?





Have you seen this picture ?



You have seen these pictures



You were tested with these pictures



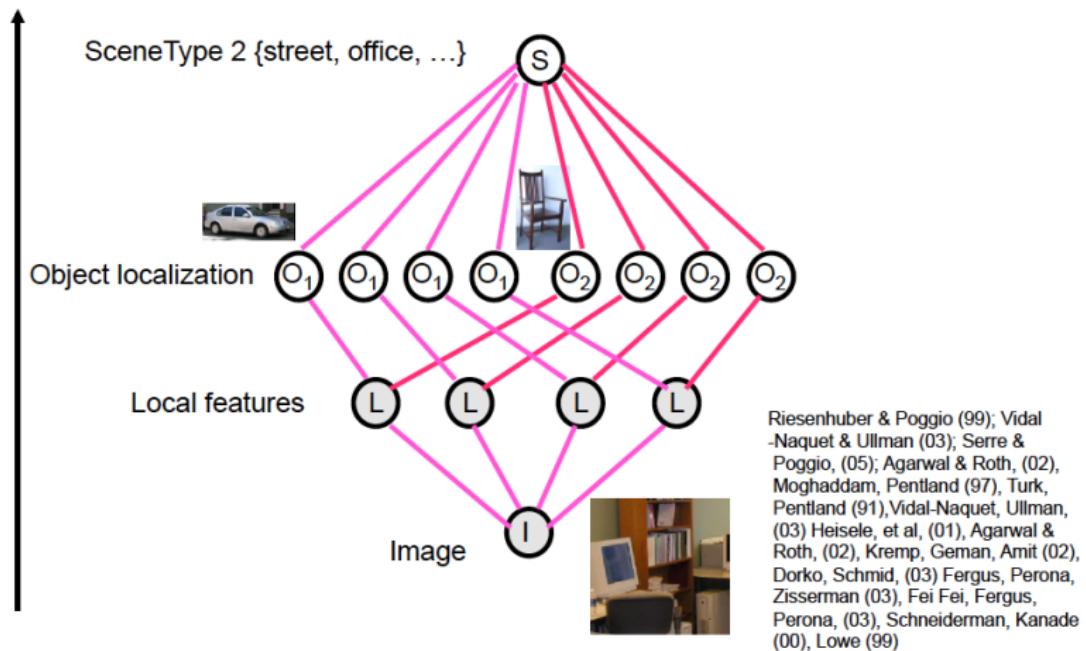
Gist

On se rappelle la signification (la sémantique d'une image) et son apparence globale mais plusieurs objets et détails sont oubliés.



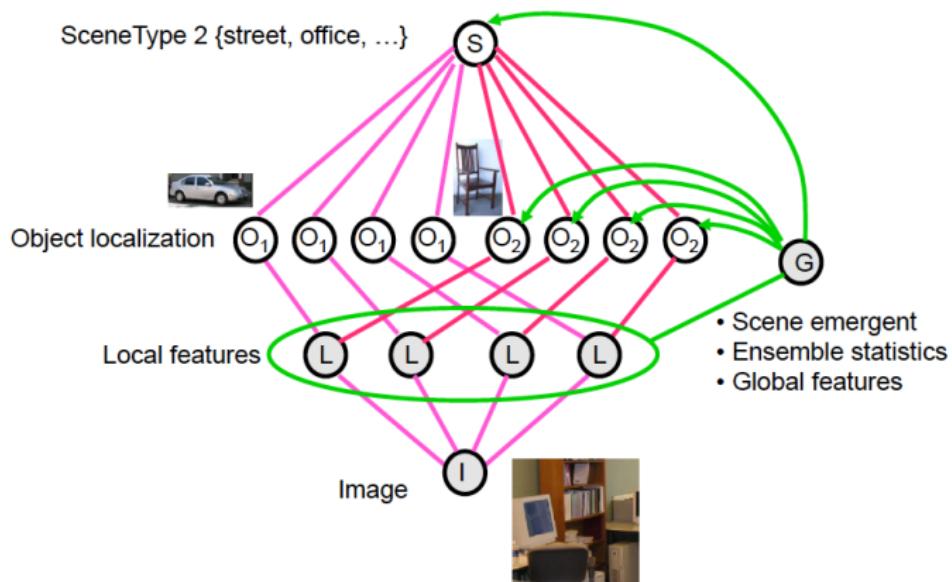
Comment décrire une scène ?

Des objets à la scène



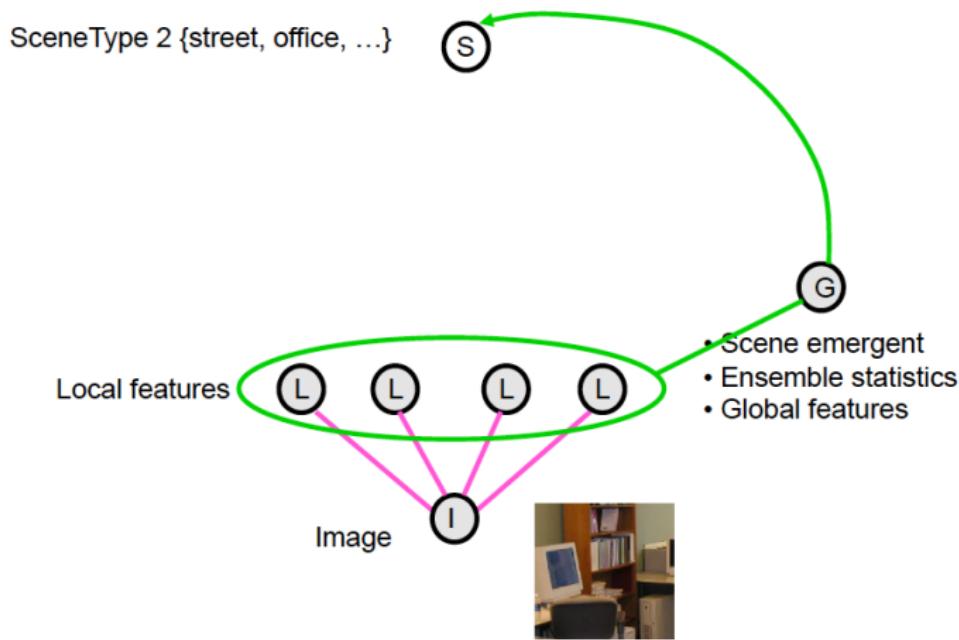
Comment décrire une scène ?

S'abstraire autant que possible des objets à l'aide de caractéristiques émergentes de la scène (suggestion de la scène) et de statistiques (nous avons une représentation plus claire de la moyenne que de l'individu)

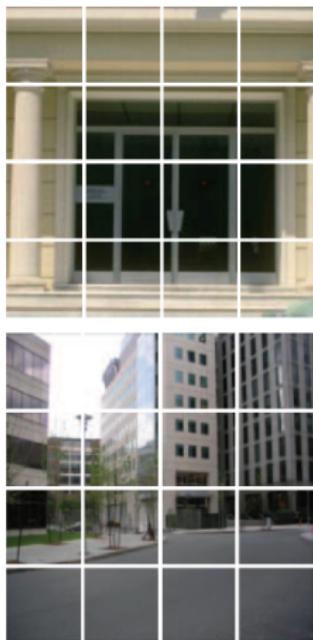


Comment décrire une scène ?

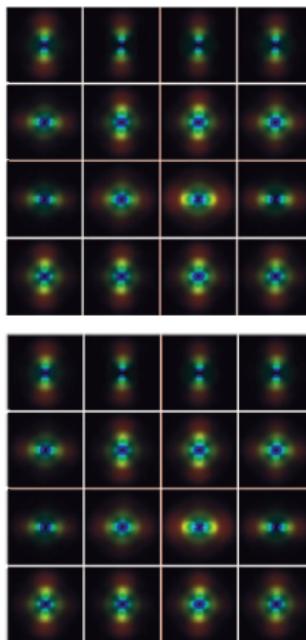
Et sans les objets ?



Descripteur Gist



Oliva and Torralba, 2001



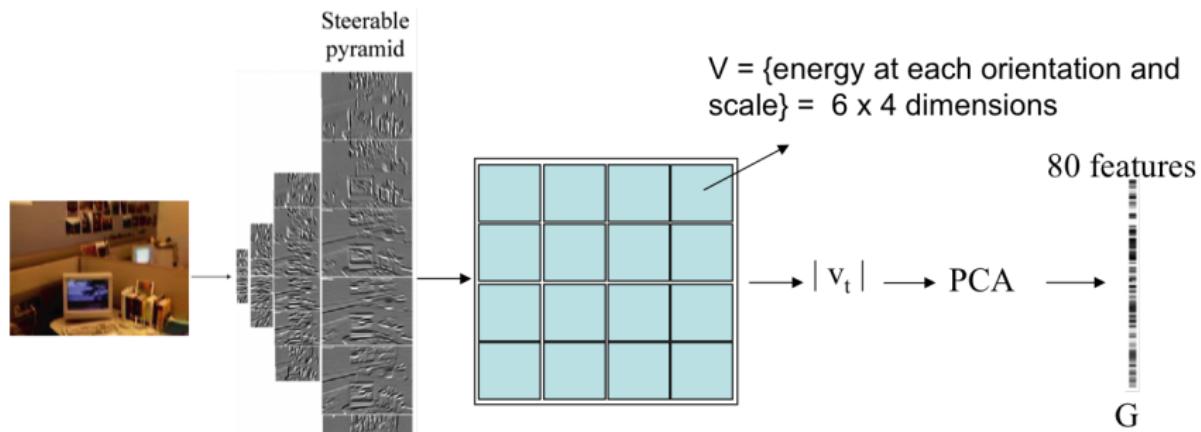
- Apply oriented Gabor filters over different scales
- Average filter energy in each bin

8 orientations
 4 scales
 \underline{x} 16 bins
 512 dimensions

Similar to SIFT (Lowe 1999) applied to the entire image

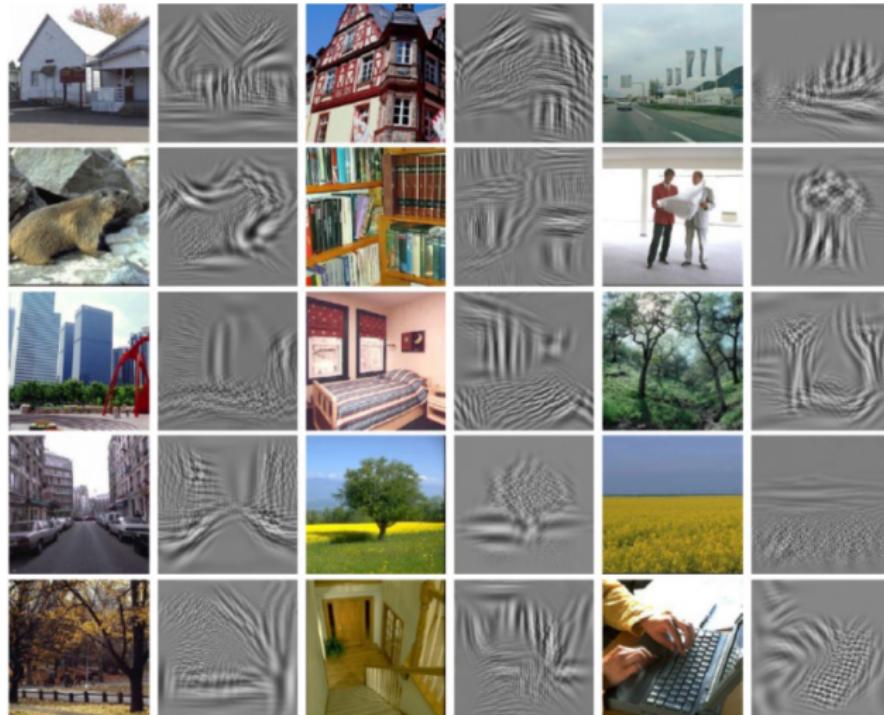
M. Gorkani, R. Picard, ICPR 1994; Walker, Malik. Vision Research 2004; Vogel et al. 2004;
 Fei-Fei and Perona, CVPR 2005; S. Lazebnik, et al, CVPR 2006; ...

Descripteur Gist



Oliva, Torralba. IJCV 2001

Descripteur Gist



Global features (I) \sim global features (I')

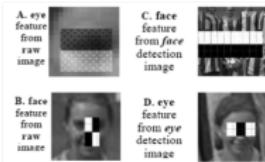
Oliva & Torralba (2001)

Importance du contexte en vision par ordinateur

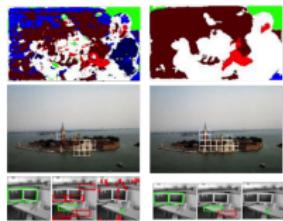
Torralba, Sinha (2001)



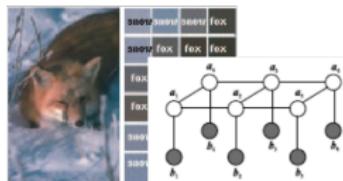
Fink & Perona (2003)



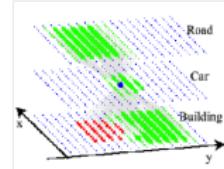
Kumar, Hebert (2005)



Carbonetto, de Freitas & Barnard (2004)



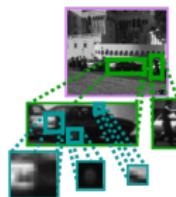
Torralba Murphy Freeman (2004)



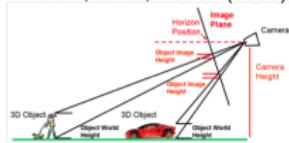
Rabinovich et al (2007)



Sudderth, Torralba, Wilsky, Freeman (2005)



Hoiem, Efros, Hebert (2005)



Desai, Ramantan, and Fowlkes (2009)



Bilan : descripteurs

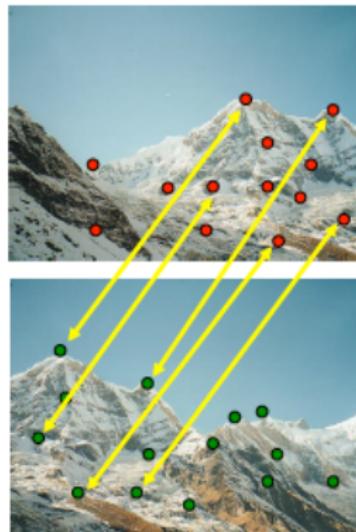
Les bons descripteurs dépendent de l'objectif visé.

- Objet :
 - ▶ SIFT, HOG, color, Bow.
- Scène :
 - ▶ GIST, Bow, color.
- Propriétés sur les matériaux :
 - ▶ Color, texture
- Mouvement.
 - ▶ Flot optique, ...

Création d'un panorama

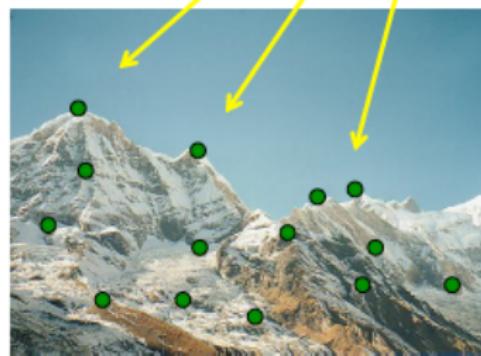
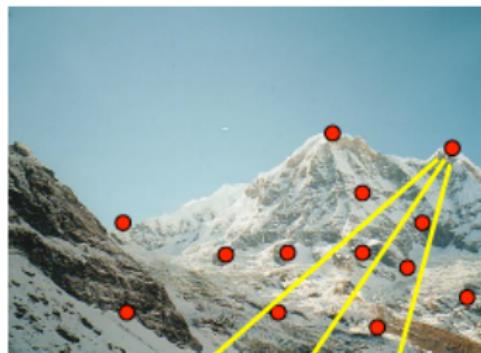
Retour au problème de création d'un panorama.

- 1) Feature Detection:
Identify image features
- 2) Feature Description:
Extract feature descriptor
for each feature
- 3) **Feature Matching:**
Find candidate matches
between features
- 4) Feature Correspondence:
Find consistent set of
(inlier) correspondences
between features



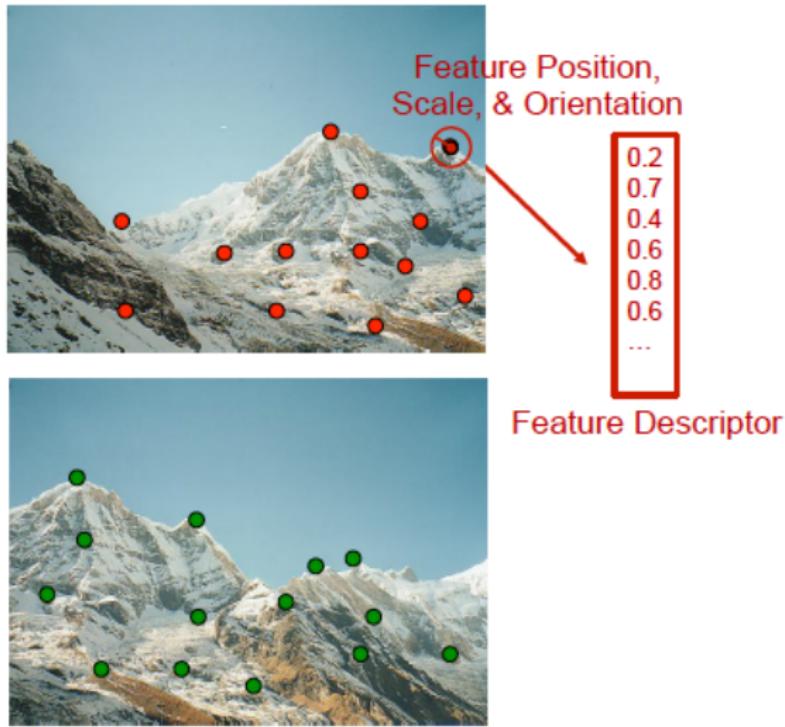
Création d'un panorama

Comment ?



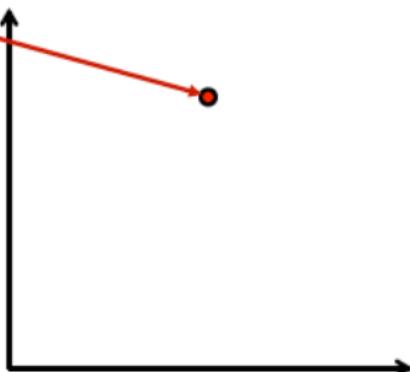
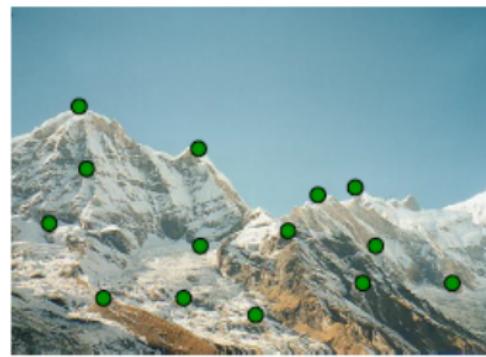
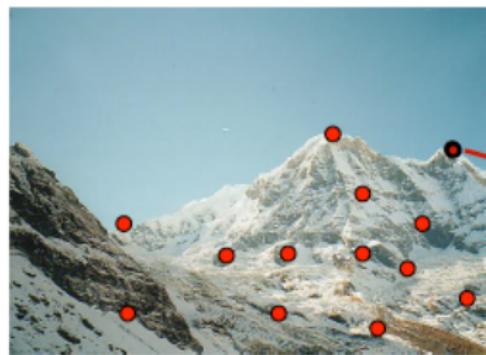
Création d'un panorama

Comment ?



Création d'un panorama

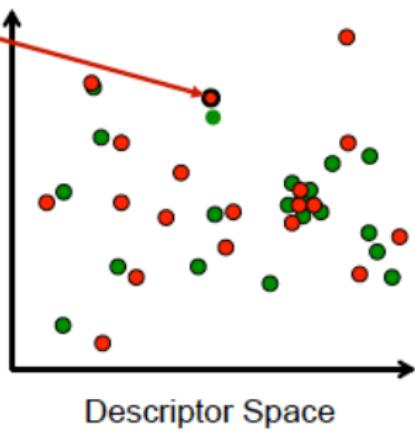
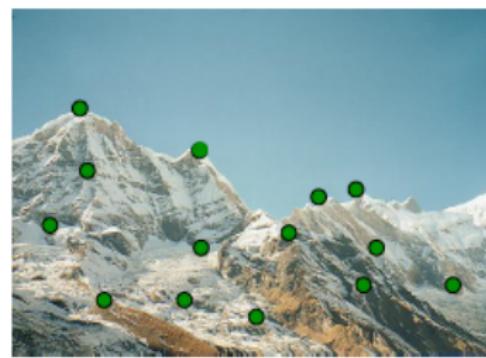
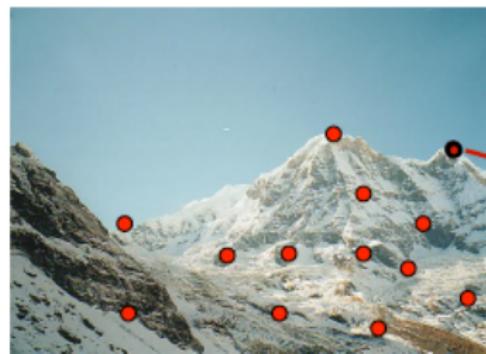
Comment ?



Descriptor Space

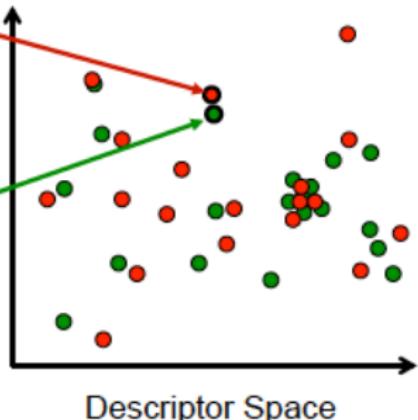
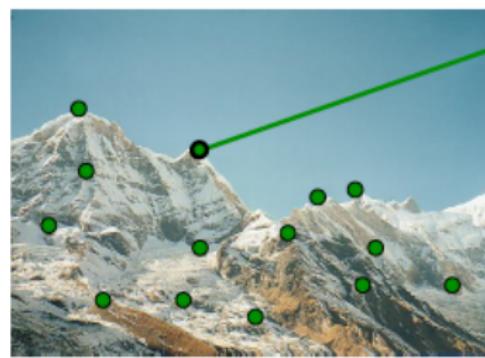
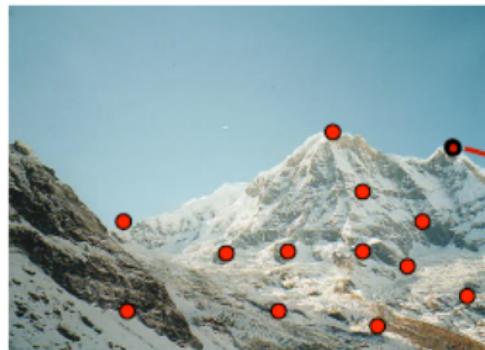
Création d'un panoramae

Comment ?



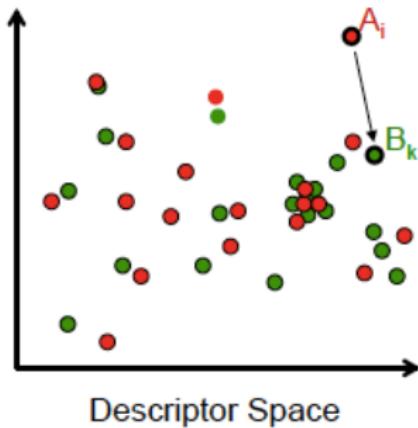
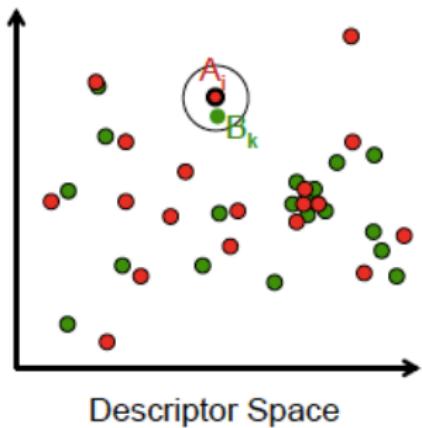
Création d'un panorama

Comment ? Quelles stratégies pour la recherche des candidats ?



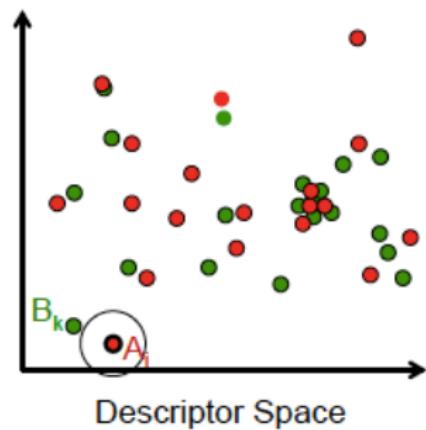
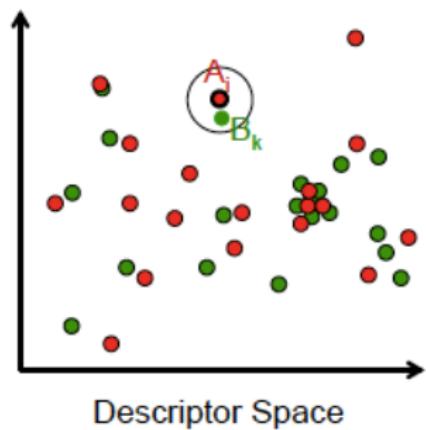
Création d'un panorama

Stratégie 1 : on prend le point le plus proche selon une distance définie dans l'espace des descripteurs.



Création d'un panorama

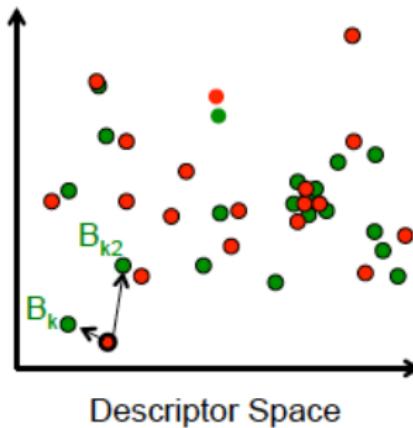
Stratégie 2 : on prend le point dont la distance avec le point cible est inférieure à un seuil.



Création d'un panorama

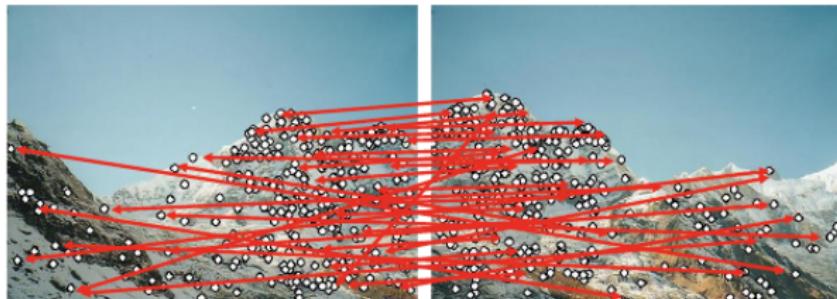
Stratégie 3 : Méthode du ratio

On prend en compte le deuxième voisin le plus proche et on accepte la correspondance si et seulement si $\frac{d(A_i, B_k)}{d(A_i, B_{k2})} < \text{thresh}$.

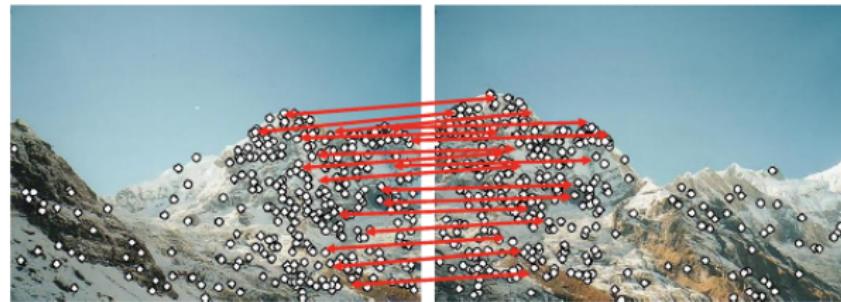


Création d'un panorama

Des candidats aux correspondances.



Candidate matches



Correspondences

Création d'un panorama

Etape suivante : **Image warping**.

- Il s'agit de rechercher une homographie H (transformation géométrique qui a 8 paramètres libres et que l'on représente avec une matrice 3×3) qui permet de déformer une image pour qu'elle se superpose avec l'autre image.
- 4 paires de correspondance sont nécessaires pour calculer l'homographie mais pour que l'estimation de l'homographie soit bonne il vaut mieux sélectionner plus de 4 points de correspondance.
- Algorithme **RANSAC** (RANdom SAmple Consensus) (Fischler & Bolles 1981)³

3. <http://www.cs.ait.ac.th/~mdailey/cvreadings/Fischler-RANSAC.pdf>   

Création d'un panorama

RANSAC

- Objectif : éviter l'impact des outliers dans le calcul du modèle en cherchant des *inliers* et en utilisant uniquement ces points.

RANSACloop:

Repeat for k iterations:

1. Randomly select a *seed group* of points on which to perform a model estimate (e.g., a group of edge points)
2. Compute model parameters from seed group
3. Find *inliers* to this model
4. If the number of inliers is sufficiently large, re-compute least-squares estimate of model on all of the inliers
 - Keep the model with the largest number of inliers

Création d'un panorama

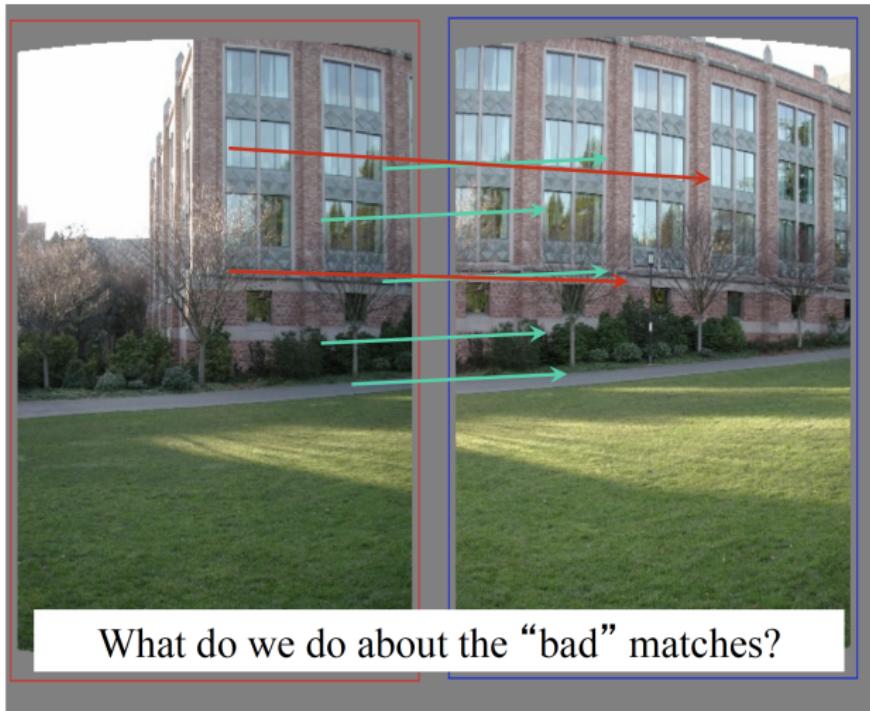
Pourquoi RANSAC est nécessaire ?



Points jaunes : inliers - Points bleus : outliers - Points rouges : points rejetés par la sélection 2-NN

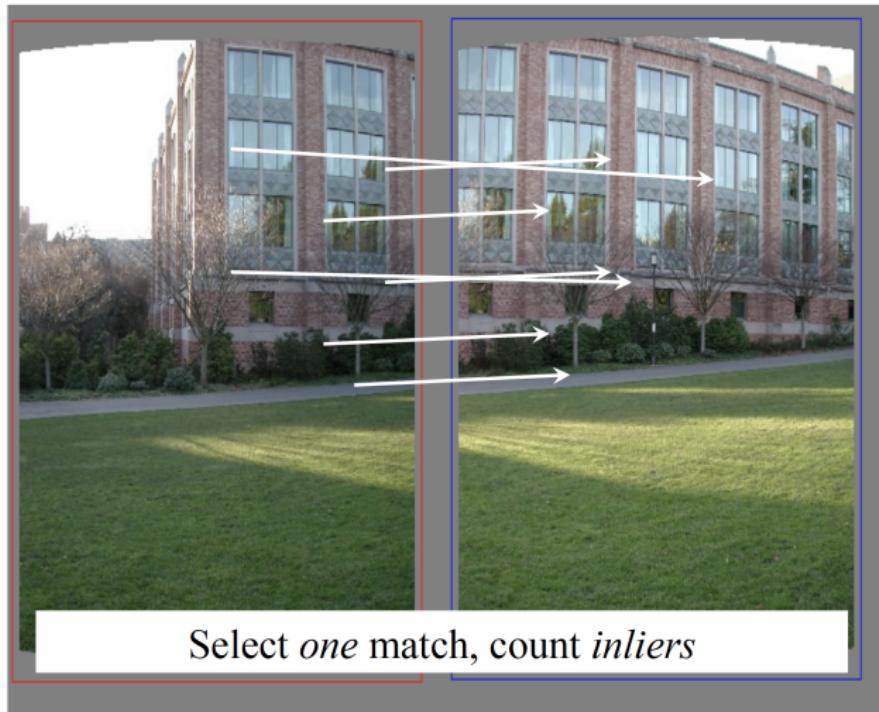
Création d'un panorama

Pourquoi RANSAC est nécessaire ?



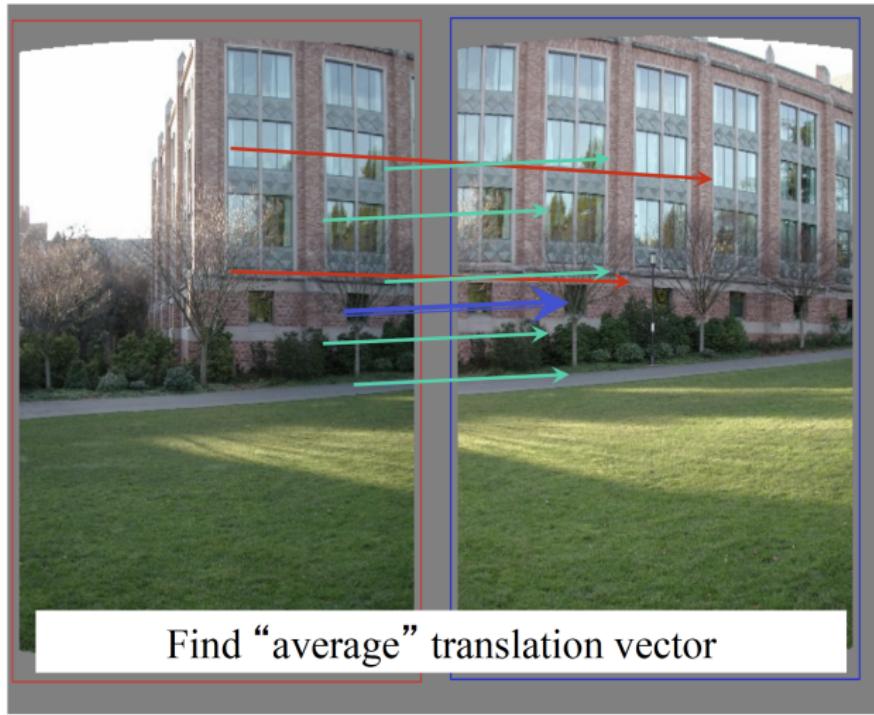
Création d'un panorama

Pourquoi RANSAC est nécessaire ?



Création d'un panorama

Pourquoi RANSAC est nécessaire ?



Création d'un panorama

Pourquoi RANSAC est nécessaire ?

RANSAC loop:

- 
1. Select four feature pairs (at random)
 2. Compute homography H (exact)
 3. Compute *inliers* where $SSD(p_i', H p_i) < \varepsilon$
 4. Keep largest set of inliers
 5. Re-compute least-squares H estimate on all of the inliers

Plan

- 1 Introduction
- 2 Description d'images
 - Descripteurs basiques
 - SIFT
 - Gist
- 3 Petite parenthèse : problème du panorama
 - Bow
- 4 Classification
- 5 Conclusion

La reconnaissance : famille d'approches

Les différents modèles

Bag of words models



Csurka, Dance, Fan, Willamowski, and Bray 2004
Sivic, Russell, Freeman, Zisserman, ICCV 2005

Voting models



Viola and Jones, ICCV 2001
Heisele, Poggio, et. al., NIPS 01
Schniederma, Kanade 2004
Vidal-Naquet, Ullman 2003

Shape matching Deformable models



Berg, Berg, Malik, 2005
Cootes, Edwards, Taylor, 2001

Constellation models



Fischler and Elschlager, 1973
Burl, Leung, and Perona, 1995
Weber, Welling, and Perona, 2000
Fergus, Perona, & Zisserman, CVPR 2003

Rigid template models



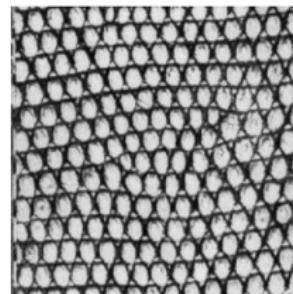
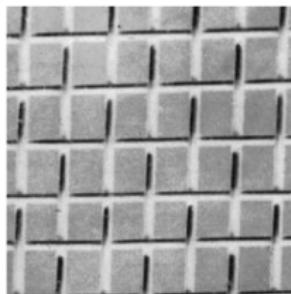
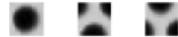
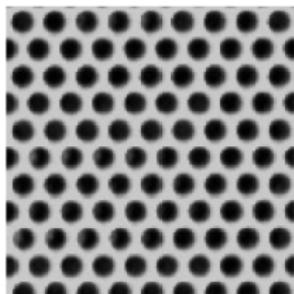
Sirovich and Kirby 1987
Turk, Pentland, 1991
Dalal & Triggs, 2006

Bow : Bag of words



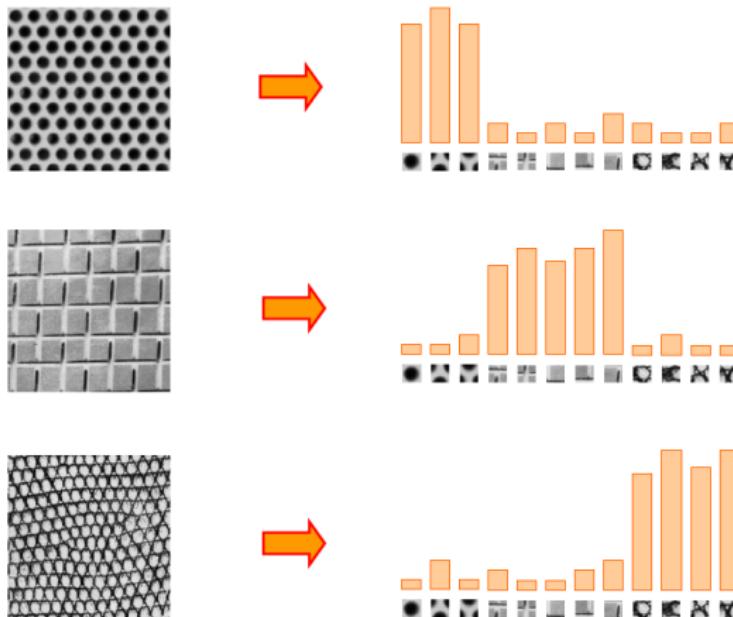
Origine 1 : reconnaissance de texture

- Une texture est caractérisée par la répétition d'éléments ou primitives de base : les *textons*.
- Pour les textures stochastiques (analyse statistique de la manière dont les niveaux de gris s'organisent les uns par rapport aux autres dans l'image), la nature des primitives (ou textons) compte plus que leur arrangement spatial.



[Julesz,81] ; [Cula and Dana, 01] ; [Leung and Malik,01] ; [Mori, Belongie and Malik,01] ;
[Schmid, 01] ; [Varma and Zisserman, 02,03] ; [Lazebnik, Schmidt and Ponce, 03]

Origine 1 : reconnaissance de texture



[Julesz,81] ; [Cula and Dana, 01] ; [Leung and Malik,01] ; [Mori, Belongie and Malik,01] ;
[Schmid, 01] ; [Varma and Zisserman, 02,03] ; [Lazebnik, Schmidt and Ponce, 03]

Origine 2 : recherche d'information - modèle sac de mots

Représentation d'un document par un sac de mots = fréquence des mots d'un dictionnaire dans le document [Salton and McGill, 83]

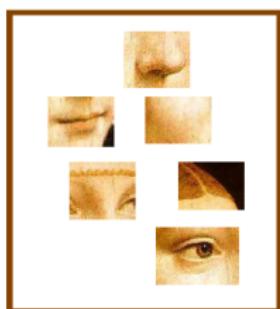


Approche sac de mots pour la description d'image

Description globale de l'image.

Principe

- ① Extraire des caractéristiques (locales)



Approche sac de mots pour la description d'image

Principe

- ① Extraire des caractéristiques (locales)
- ② Apprendre un vocabulaire visuel à partir des caractéristiques extraites.



Approche sac de mots pour la description d'image

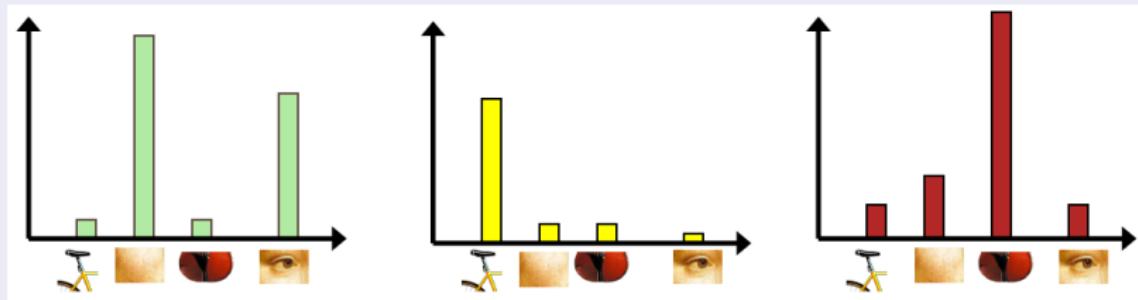
Principe

- ① Extraire des caractéristiques (locales)
- ② Apprendre un vocabulaire visuel à partir des caractéristiques extraites.
- ③ Quantifier les caractéristiques à l'aide du vocabulaire visuel.

Approche sac de mots pour la description d'image

Principe

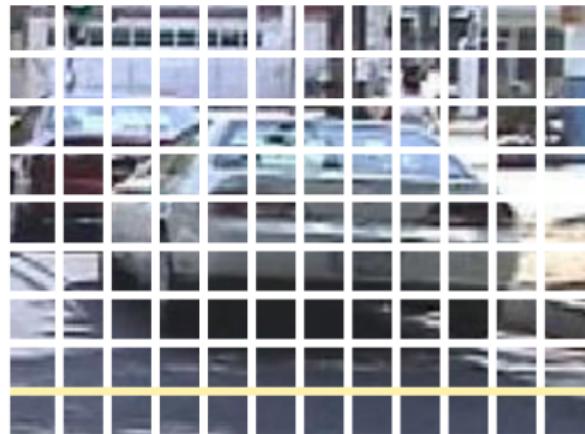
- ① Extraire des caractéristiques (locales)
- ② Apprendre un vocabulaire visuel à partir des caractéristiques extraites.
- ③ Quantifier les caractéristiques à l'aide du vocabulaire visuel.
- ④ Représenter les images par la fréquence des différents mots visuels : histogrammes de mots visuels.



Approche sac de mots pour la description d'image

Etape 1 : Quelles caractéristiques ?

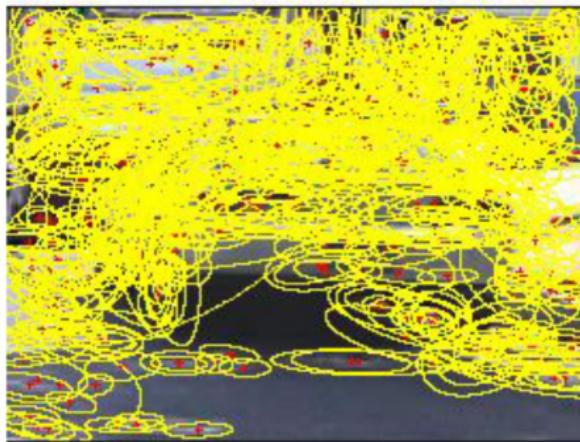
On découpe l'image en une grille régulière



Approche sac de mots pour la description d'image

Etape 1 : Quelles caractéristiques ?

Points d'intérêts



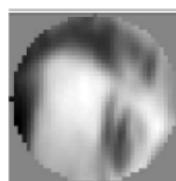
Approche sac de mots pour la description d'image

Etape 1 : Comment décrire les caractéristiques ?

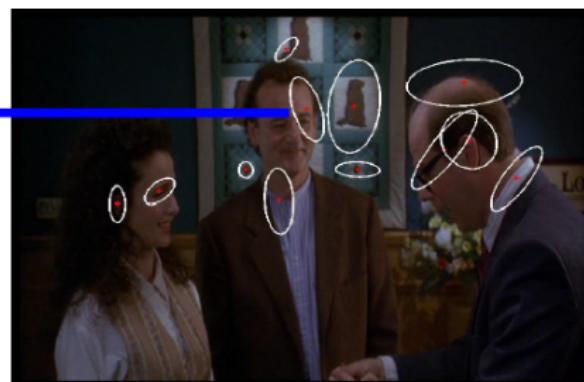


Compute SIFT
descriptor

[Lowe'99]



Normalize patch



Detect patches

[Mikojaczyk and Schmid '02]

[Mata, Chum, Urban & Pajdla, '02]

[Sivic & Zisserman, '03]

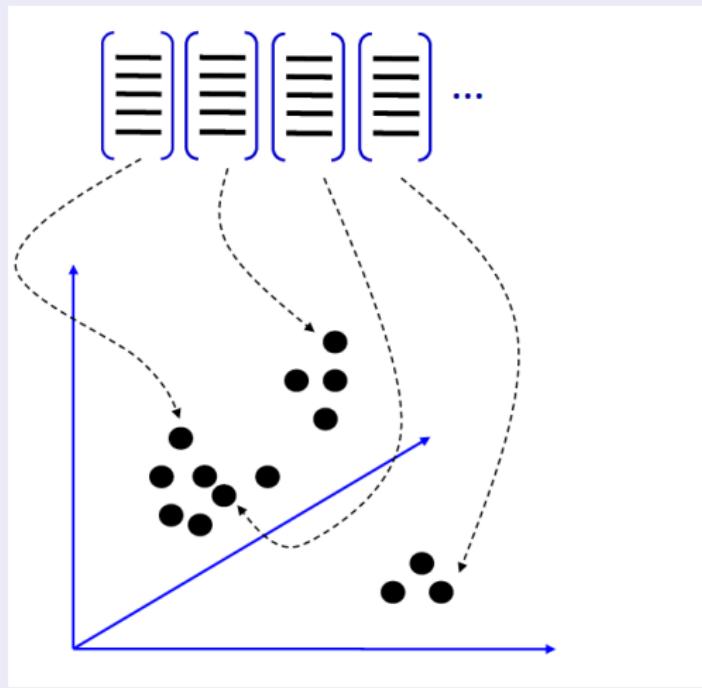
Approche sac de mots pour la description d'image

Etape 1 : Comment décrire les caractéristiques ?



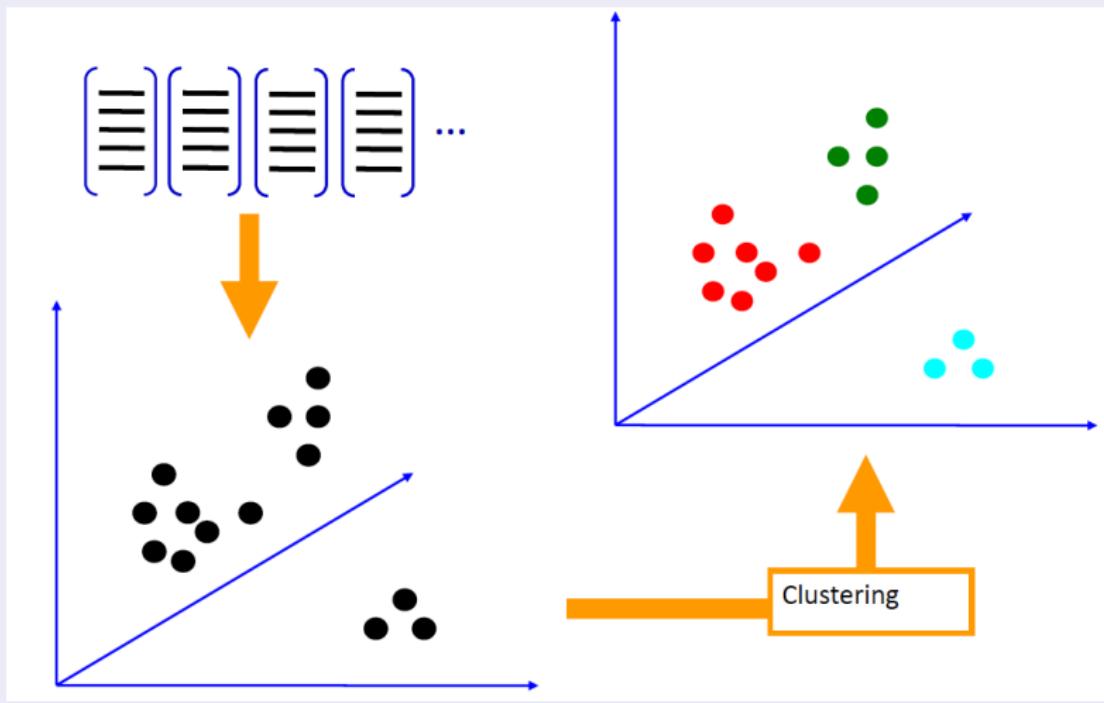
Approche sac de mots pour la description d'image

Etape 2 : Apprendre un vocabulaire de mots visuels ?



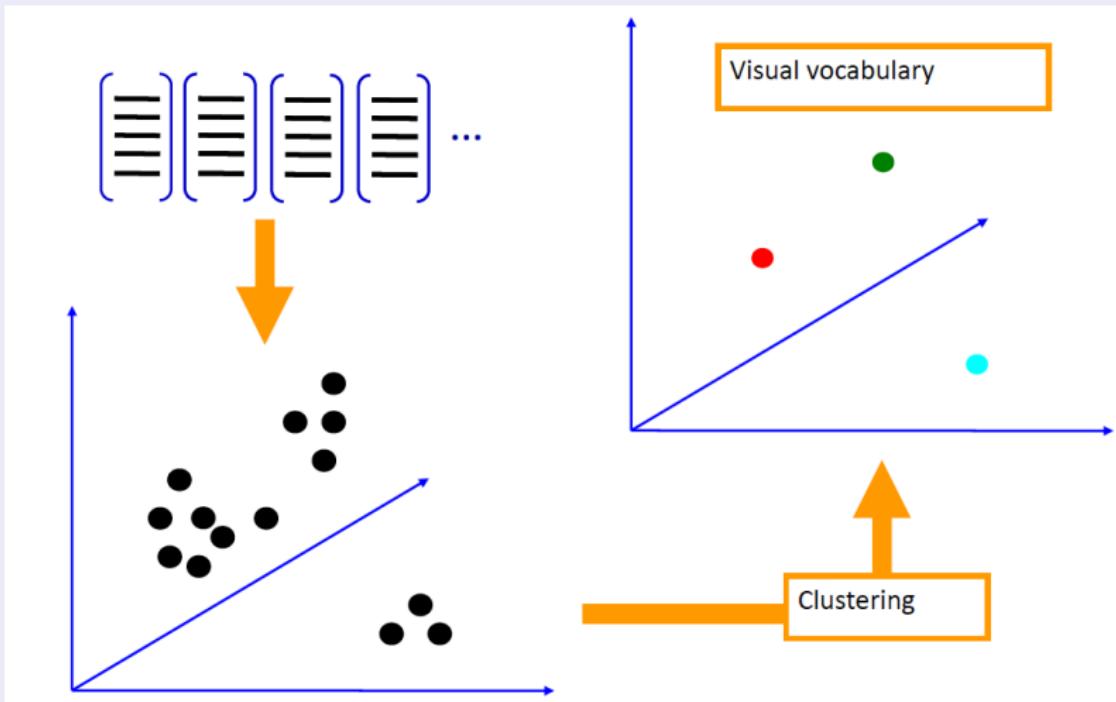
Approche sac de mots pour la description d'image

Etape 2 : Apprendre un vocabulaire de mots visuels ?



Approche sac de mots pour la description d'image

Etape 2 : Apprendre un vocabulaire de mots visuels ?



Approche sac de mots pour la description d'image

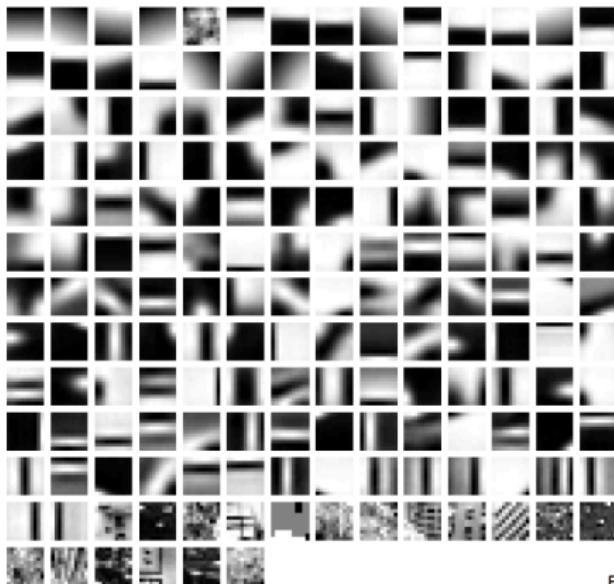
Etape 2 : Apprendre un vocabulaire de mots visuels ?

- Le clustering est une méthode classique pour apprendre un vocabulaire de mots visuels (ou codebook).
 - ▶ Processus non supervisé (mais choix de k pour fixer la taille du vocabulaire)
 - ▶ Chaque centre de cluster produit devient un mot visuel.
- Le vocabulaire de mots visuels est utilisé pour quantifier les caractéristiques.
 - ▶ Un vecteur de caractéristiques est pris en entrée et mis en correspondance avec les mots visuels du dictionnaire le plus proche.
 - ▶ Codebook = vocabulaire visuel.
 - ▶ Codevector = mot visuel.
 - ▶ Chaque centre de cluster produit devient un mot visuel.

Approche sac de mots pour la description d'image

Etape 2 : Apprendre un vocabulaire de mots visuels ?

Exemple de dictionnaire de mots visuels.



Fei-Fei et al. 2005

Approche sac de mots pour la description d'image

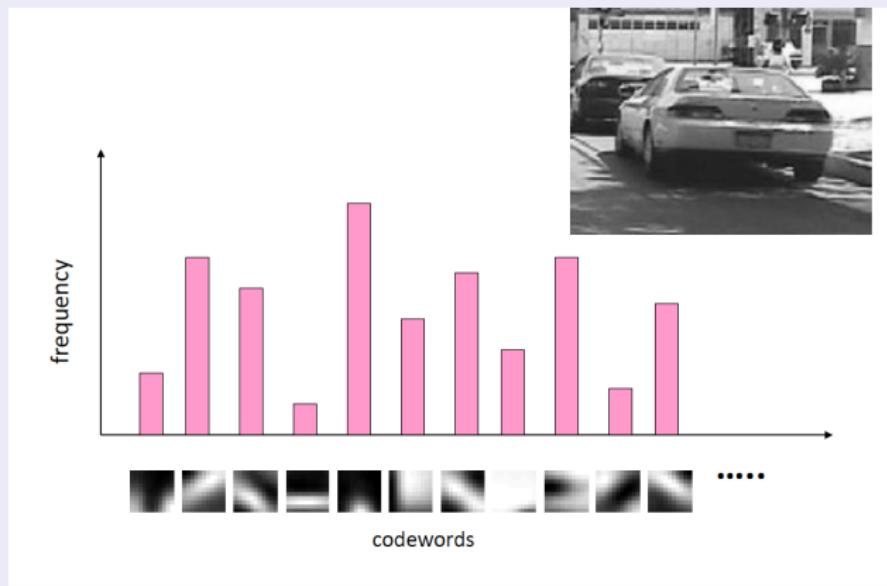
Etape 2 : Apprendre un vocabulaire de mots visuels ?

Les principaux problèmes

- Comment choisir la taille du vocabulaire ?
 - ▶ Trop petit : non représentatif de l'ensemble des patchs.
 - ▶ Trop grand : génère des erreurs de quantifications.
- Le calculer de manière efficace : utilisation d'arbres [Nister and Stewenius, 06]

Approche sac de mots pour la description d'image

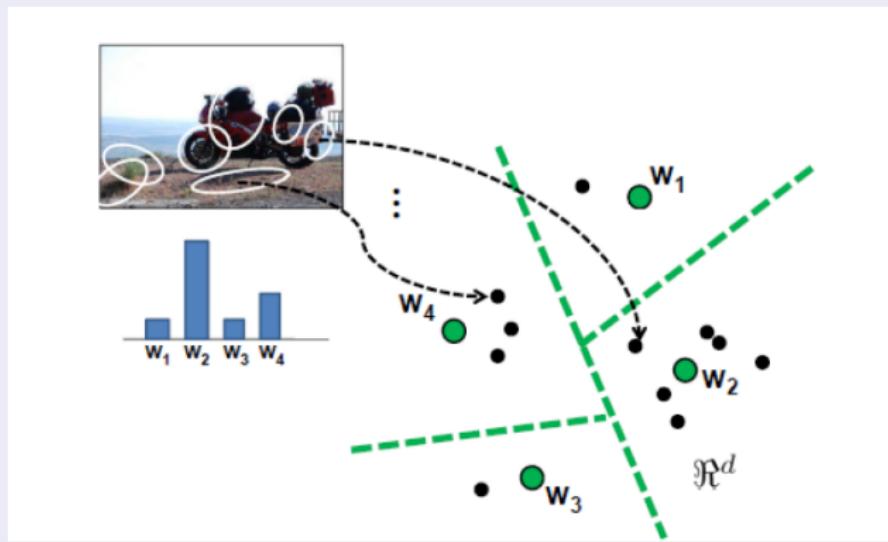
Etape 3 : Representation de l'image ?



Approche sac de mots pour la description d'image

Etape 3 : Représentation de l'image ?

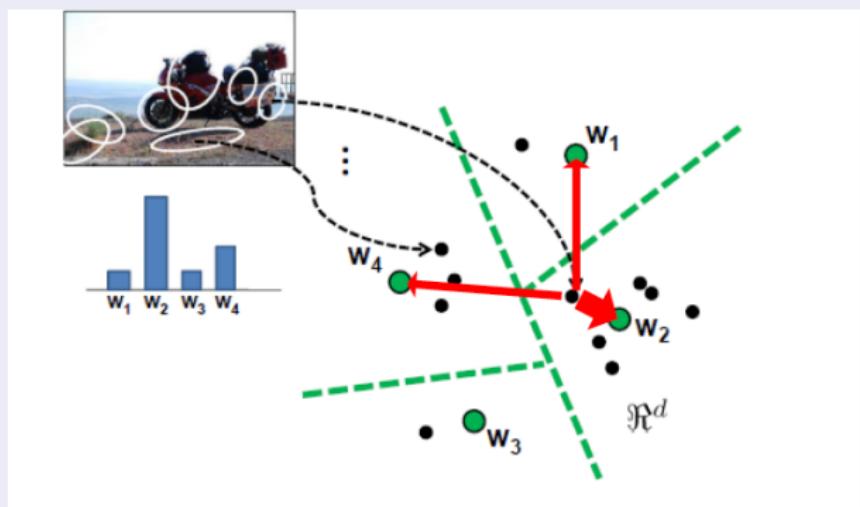
Codage simple : on code avec le mot visuel le plus proche.



Approche sac de mots pour la description d'image

Etape 3 : Representation de l'image ?

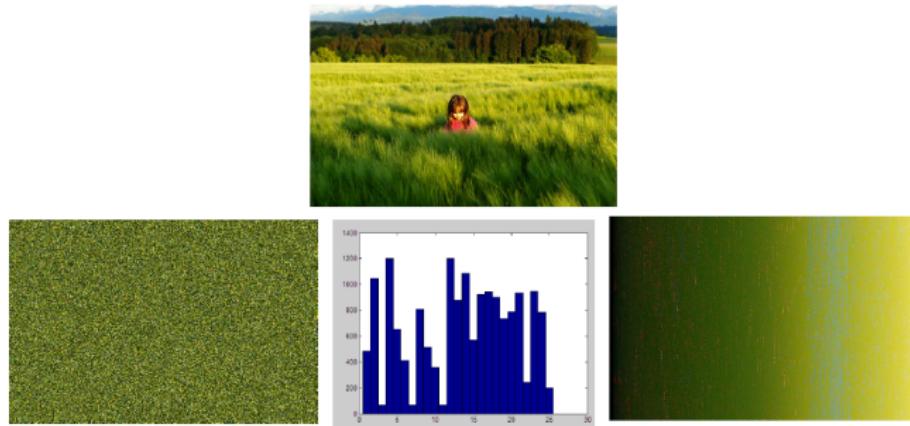
Codage avec prise en compte de contraintes locales : chaque caractéristique est décrite comme une combinaison linéaire de mots visuels.



Spatial Pyramid Representation

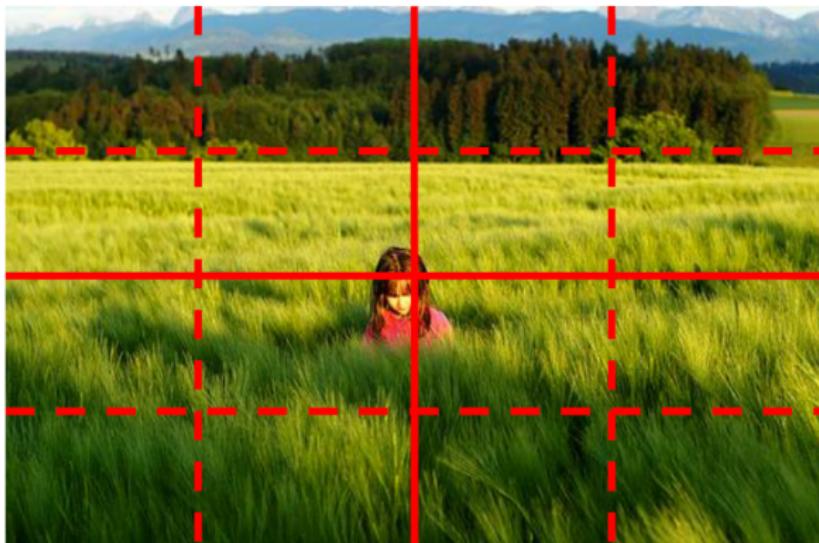
Objectif = prise en compte de l'information spatiale :

- Calculer les sacs de mots visuels sur des sous-images de l'image.



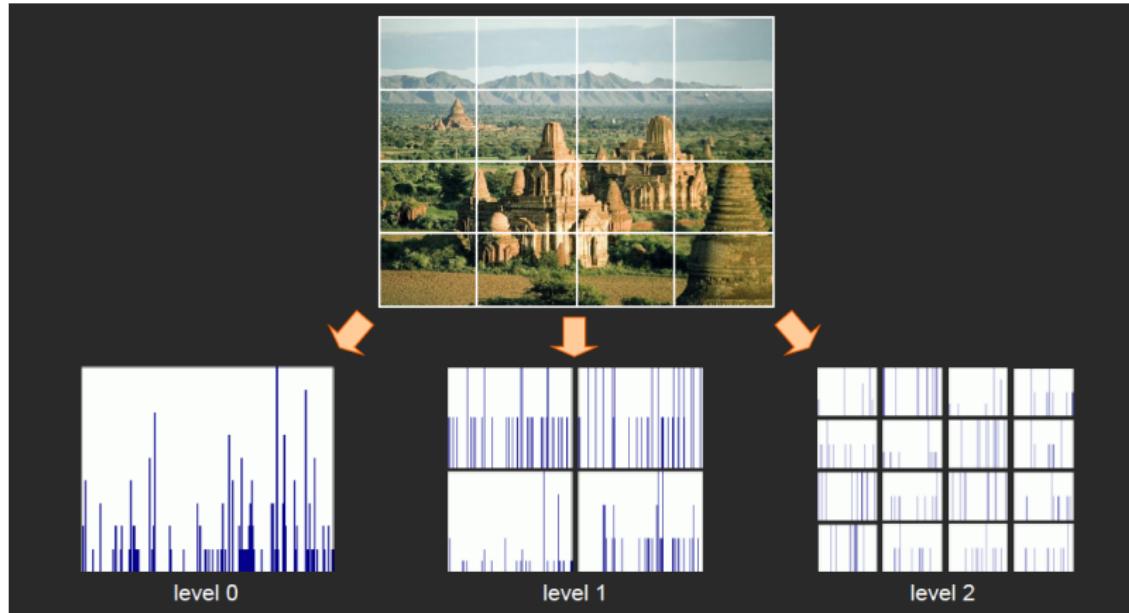
All of these images have the same color histogram

Spatial Pyramid Representation



Compute histogram in each spatial bin

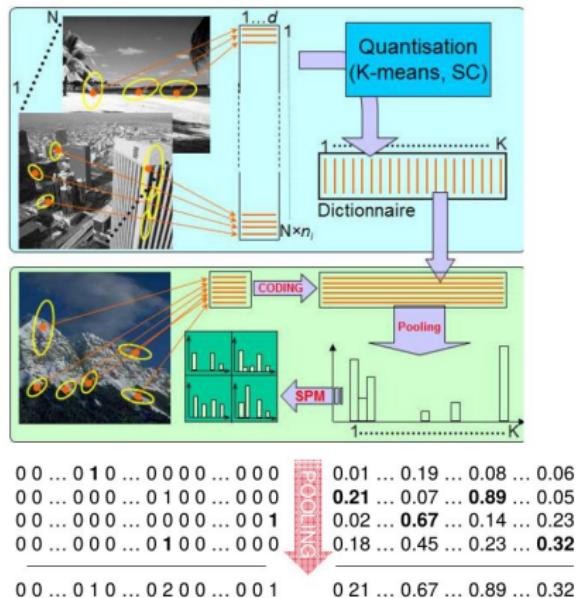
Spatial Pyramid Representation



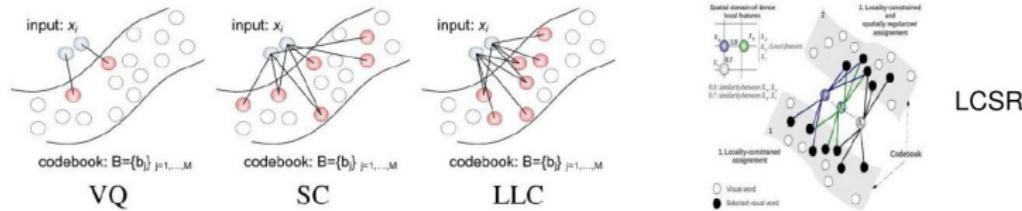
Bow : point de vue codage

Une fois un dictionnaire appris, trois étapes importantes

- extraction de descripteurs locaux
- codage des descripteurs *relativement au dictionnaire*
 - ▶ simple occurrence = hard coding
 - ▶ (local) soft coding
- agrégation (pooling) : somme ou max
- (raffinement) SPM



Bow : point de vue codage



Nombreux codage possibles

- hard coding [Csurka et al., 2004] = vector quantization
- sparse coding [Yang et al., 2009] = impose des codes creux (avec 0)
- locality constrained linear coding [Wang et al., 2010]= approx. linéaire de la variété définie par le codebook
- Local et régularisé spatialement [Shabou & Le Borgne, 2012] : LLC + prise en compte de la proximité spatiale locale

La reconnaissance : famille d'approches

Les différents modèles

Bag of words models



Csurka, Dance, Fan, Willamowski, and Bray 2004
Sivic, Russell, Freeman, Zisserman, ICCV 2005

Voting models



Viola and Jones, ICCV 2001
Heisele, Poggio, et. al., NIPS 01
Schniederma, Kanade 2004
Vidal-Naquet, Ullman 2003

Shape matching Deformable models



Berg, Berg, Malik, 2005
Cootes, Edwards, Taylor, 2001

Constellation models



Fischler and Elschlager, 1973
Burl, Leung, and Perona, 1995
Weber, Welling, and Perona, 2000
Fergus, Perona, & Zisserman, CVPR 2003

Rigid template models

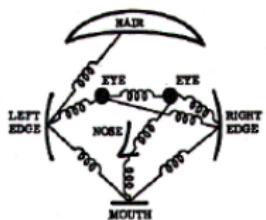


Sirovich and Kirby 1987
Turk, Pentland, 1991
Dalal & Triggs, 2006

La reconnaissance : famille d'approches

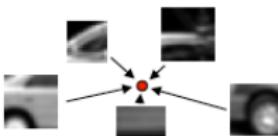
Modèles avec structure

Part Models



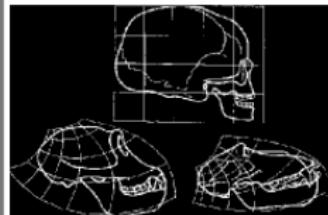
- Few parts (~6)

Voting models



- Many patches (>100)

Deformable Models

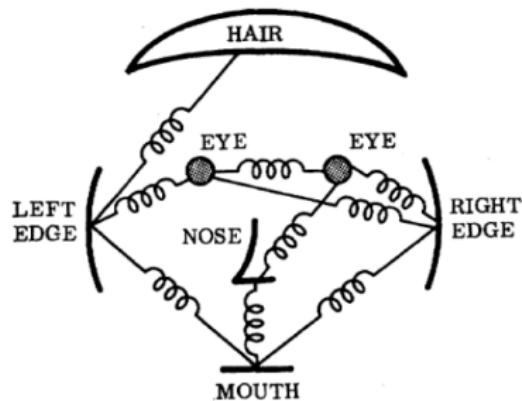


- No parts

La reconnaissance : famille d'approches

Modèles de parties

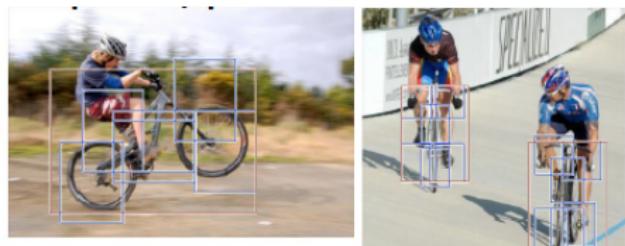
- L'objet est une configuration de plusieurs parties.
- Chaque partie est détectable.



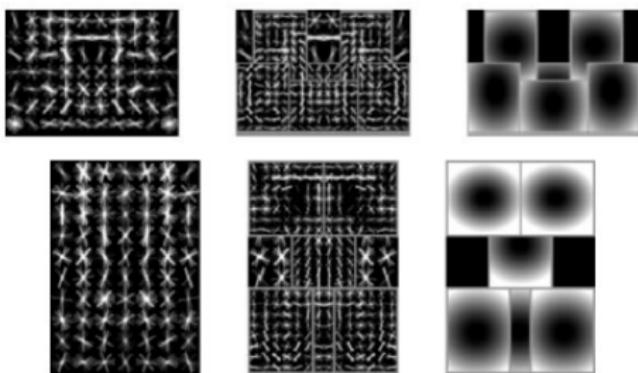
La reconnaissance : famille d'approches

Modèles de parties

Detections



Template Visualization



root filters
coarse resolution

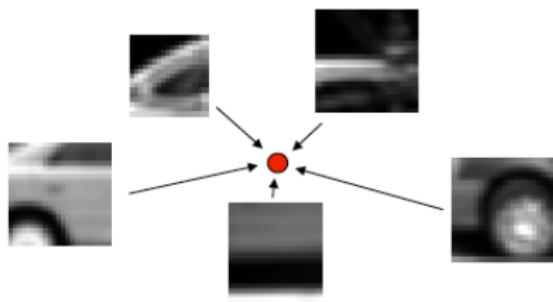
part filters
finer resolution

deformation
models

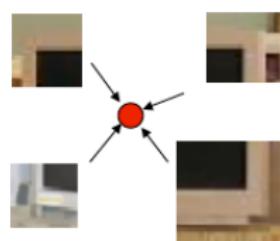
La reconnaissance : famille d'approches

Modèles de votes

Idée principale : création de détecteurs faibles à partir des parties et procédure de vote pour détecter l'objet (par exemple son centre).



Car model

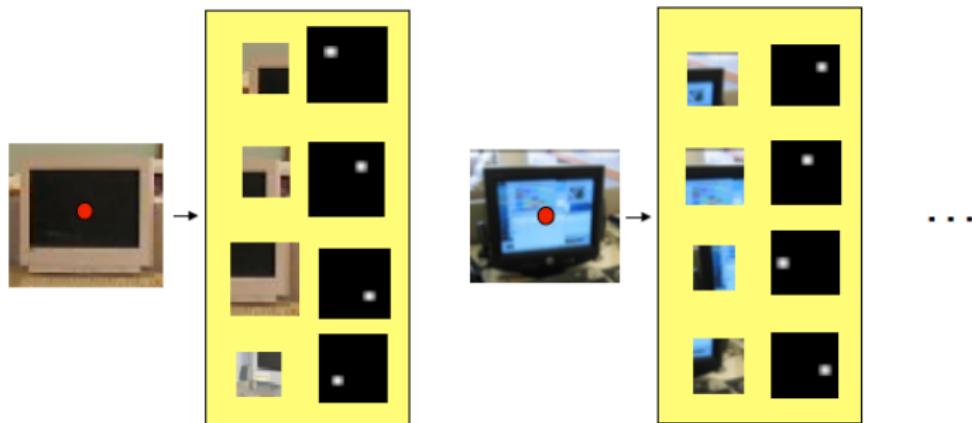


Screen model

La reconnaissance : famille d'approches

Modèles de votes

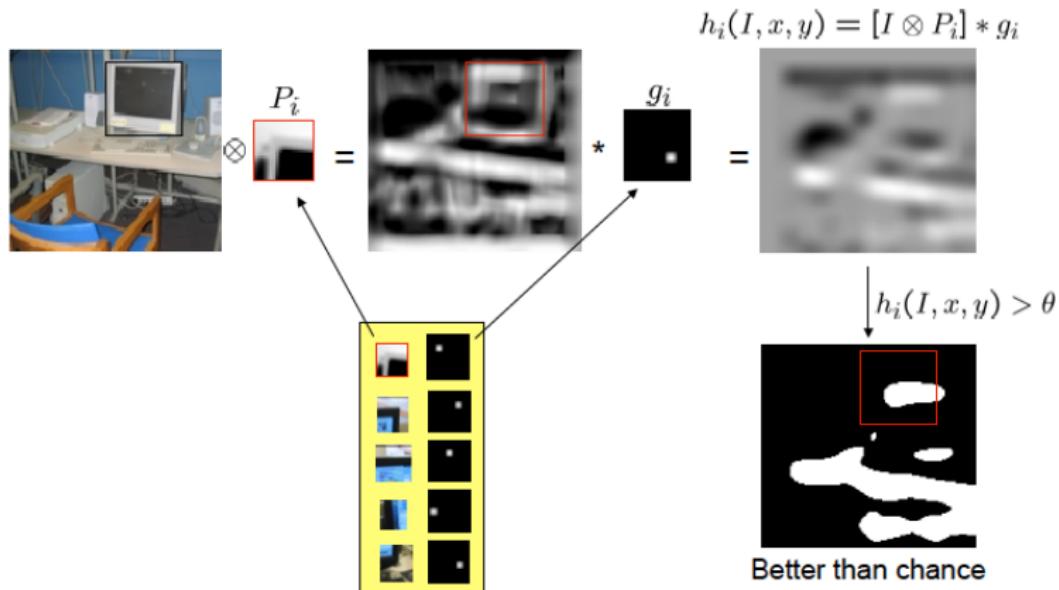
On apprend un ensemble de modèles de parties à partir d'une base d'apprentissage.



La reconnaissance : famille d'approches

Modèles de votes

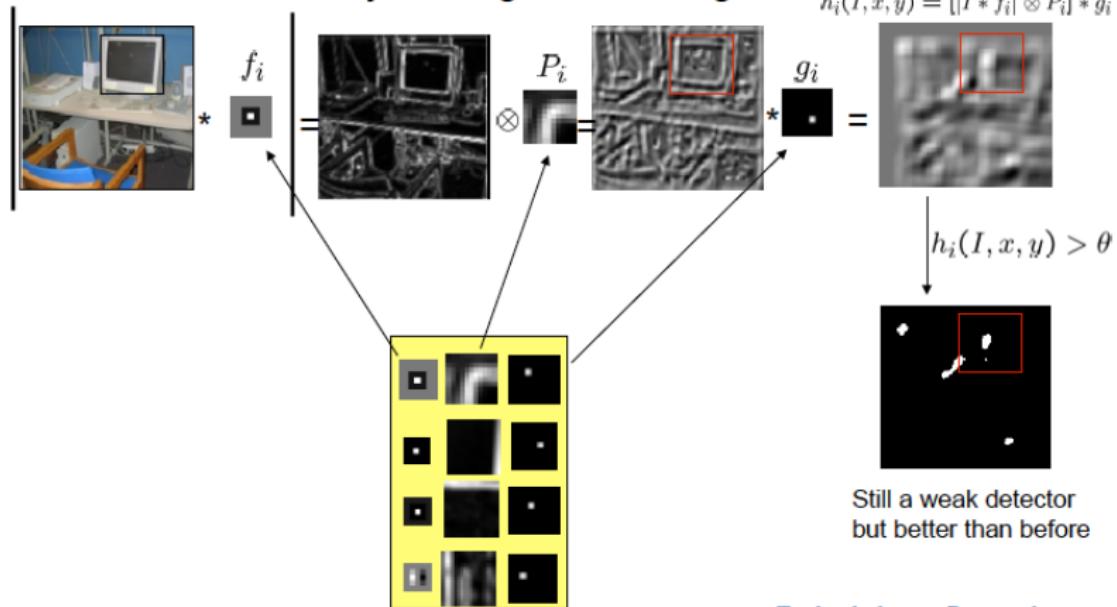
On en construit des détecteurs faibles.



La reconnaissance : famille d'approches

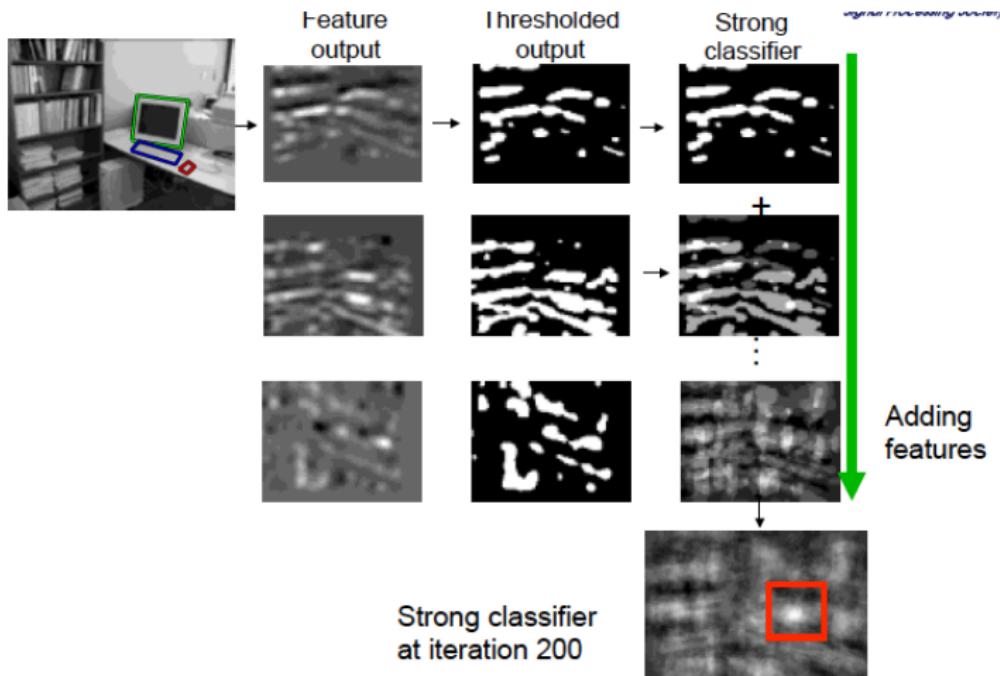
Modèles de votes

We can do a better job using filtered images



La reconnaissance : famille d'approches

Modèles de votes

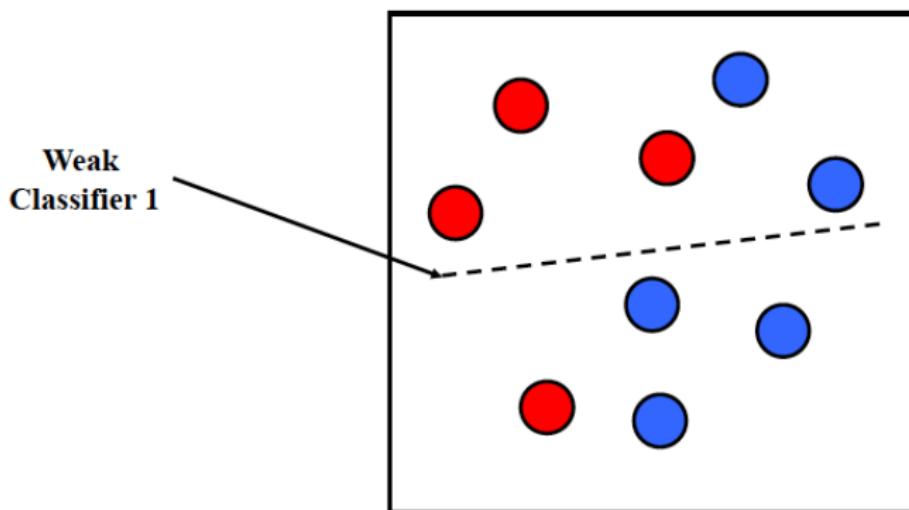


L'idée du boosting en apprentissage statistique

La reconnaissance : famille d'approches

Modèles de votes

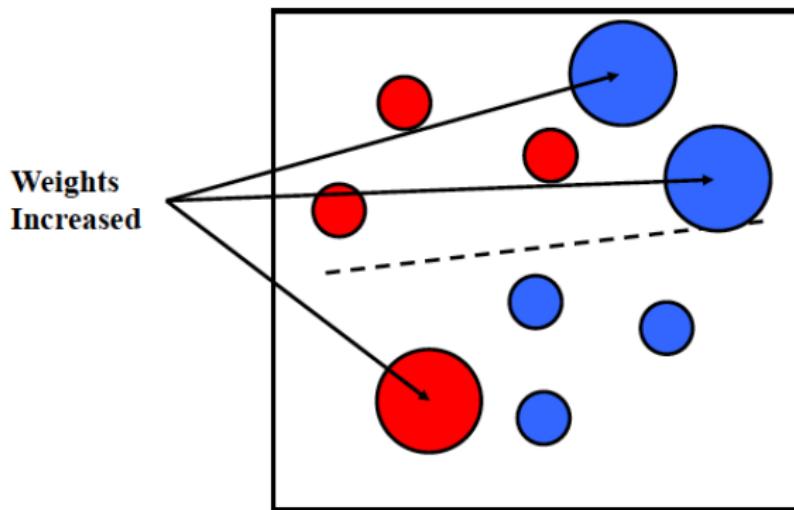
Parenthèse sur le boosting



La reconnaissance : famille d'approches

Modèles de votes

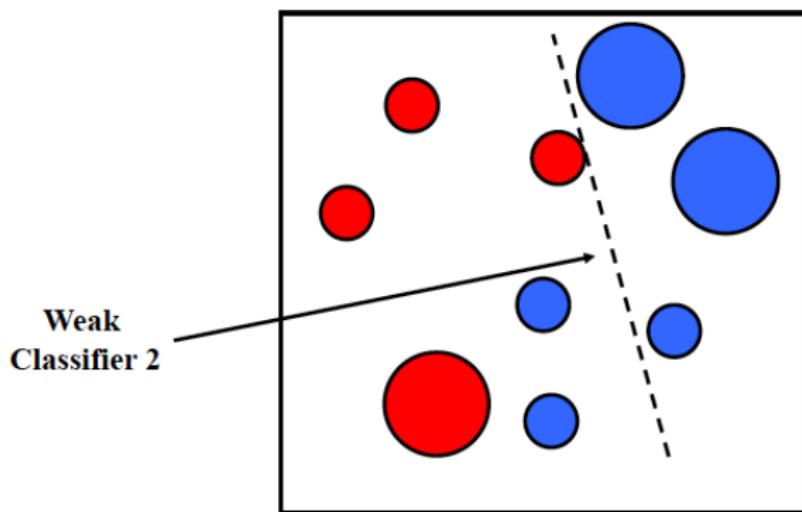
Parenthèse sur le boosting



La reconnaissance : famille d'approches

Modèles de votes

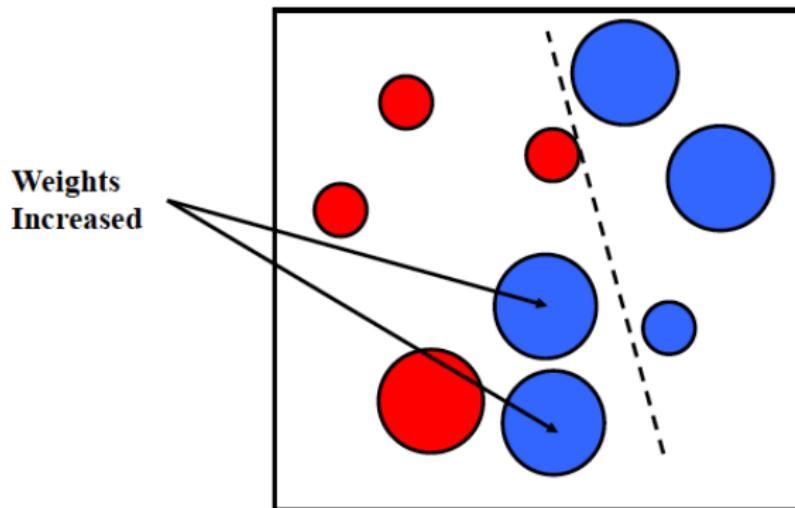
Parenthèse sur le boosting



La reconnaissance : famille d'approches

Modèles de votes

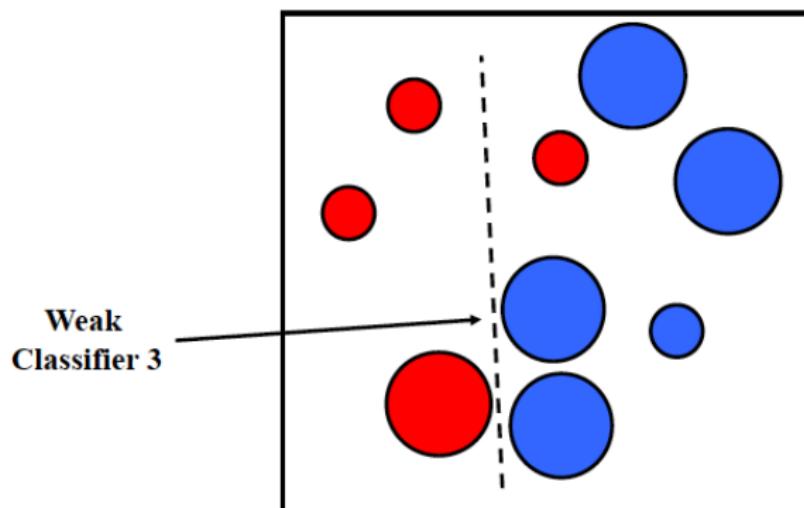
Parenthèse sur le boosting



La reconnaissance : famille d'approches

Modèles de votes

Parenthèse sur le boosting

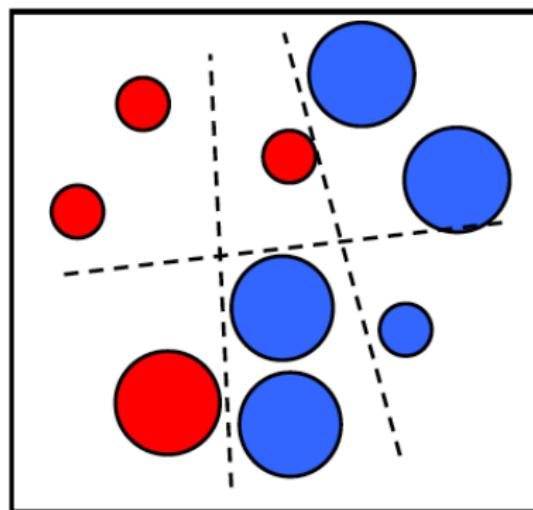


La reconnaissance : famille d'approches

Modèles de votes

Parenthèse sur le boosting

**Final classifier is
a combination of weak
classifiers**



La reconnaissance : famille d'approches

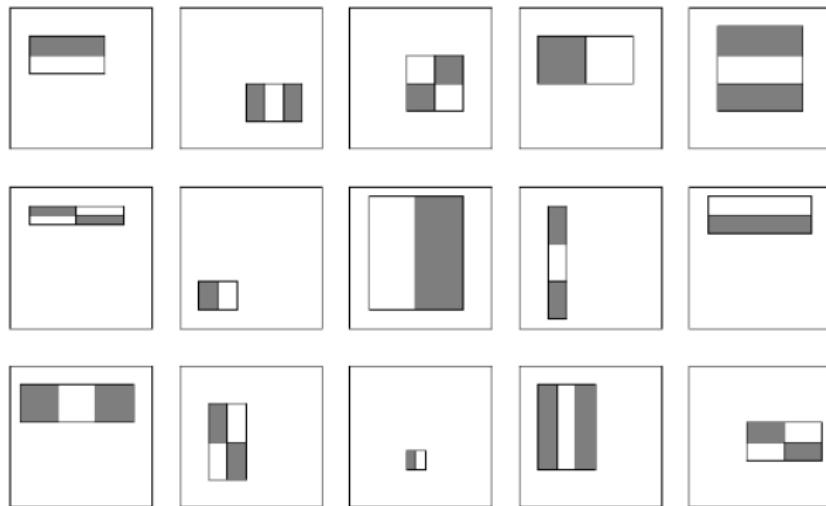
Modèles de votes

Un cas d'utilisation très connu !

La reconnaissance : famille d'approches

Modèles de votes

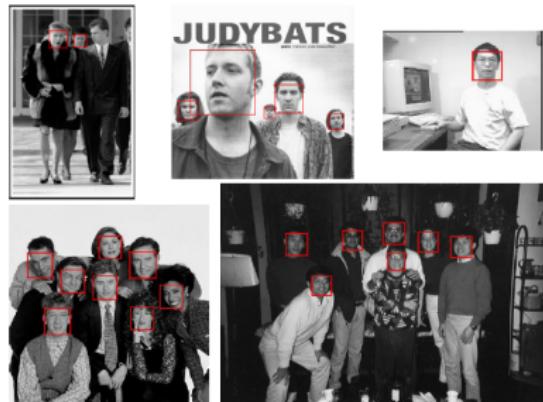
Un cas d'utilisation très connu



La reconnaissance : famille d'approches

Modèles de votes

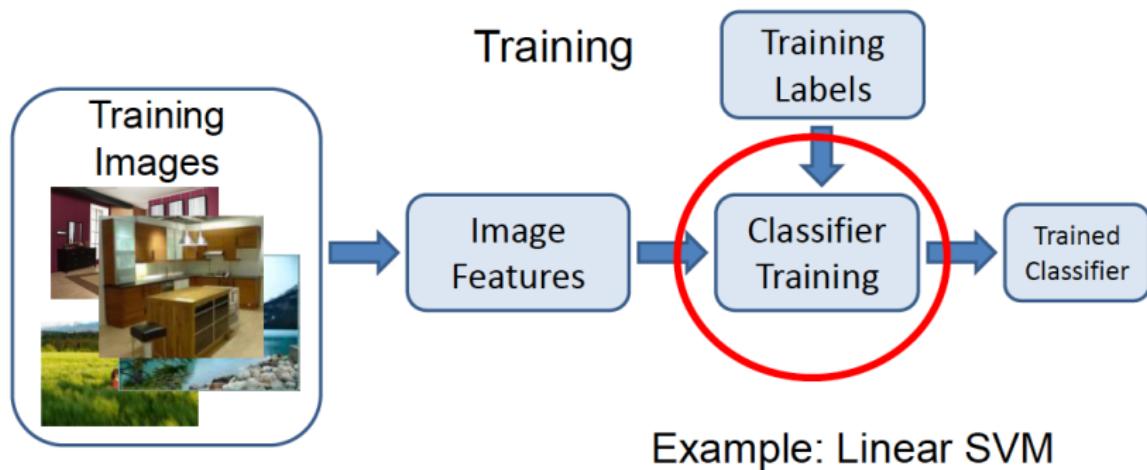
Approche de Viola Jones pour la détection de visages.



- (CVPR,01) Paul Viola et Michael Jones, Rapid Object Detection using a Boosted Cascade of Simple Features , IEEE CVPR, ? 2001
- (IJCV,04) Paul Viola et Michael Jones, Robust Real-time Face Detection , IJCV, ? 2004

Paul Viola dirige le projet PrimeAir chez Amazon (<https://www.amazon.com/Amazon-Prime-Air/b?ie=UTF8&node=8037720011>)

Quelle est l'étape suivante ?



Plan

1 Introduction

2 Description d'images

- Descripteurs basiques
- SIFT
- Gist

3 Petite parenthèse : problème du panorama

- Bow

4 Classification

5 Conclusion

Reconnaissance - Classification

Problème

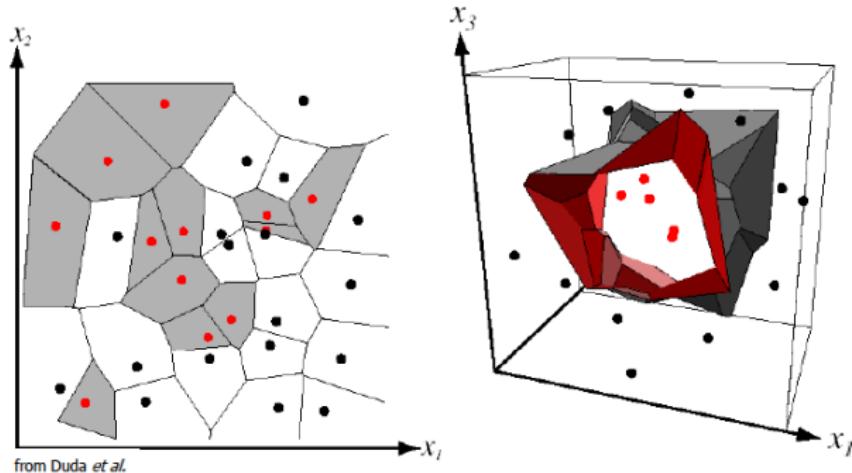
Etant donnée une représentation d'images provenant de différentes classes, comment peut-on apprendre un modèle pour les différencier ?

Deux approches

- Approches discriminatives.
- Approches génératives.

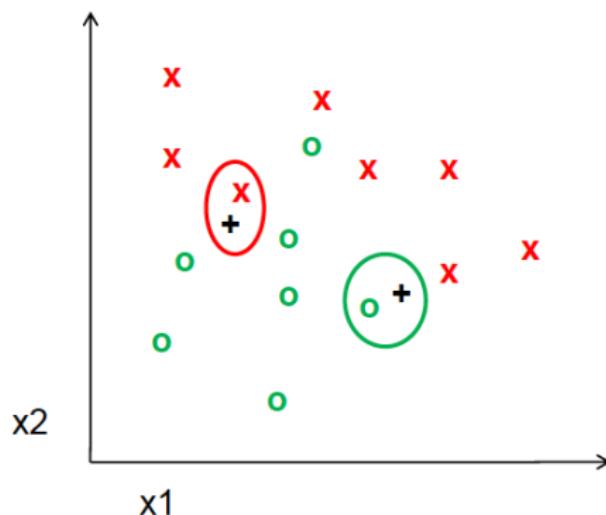
Classifieur des plus proches voisins (k-NN)

Assigner la classe du (des) point(s) le(s) plus proche(s) dans l'ensemble d'apprentissage de l'exemple à classer.

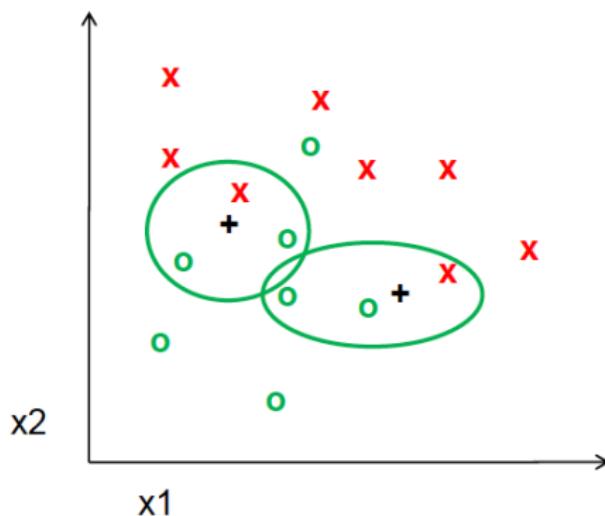


Voronoi partitioning of feature space
for two-category 2D and 3D data

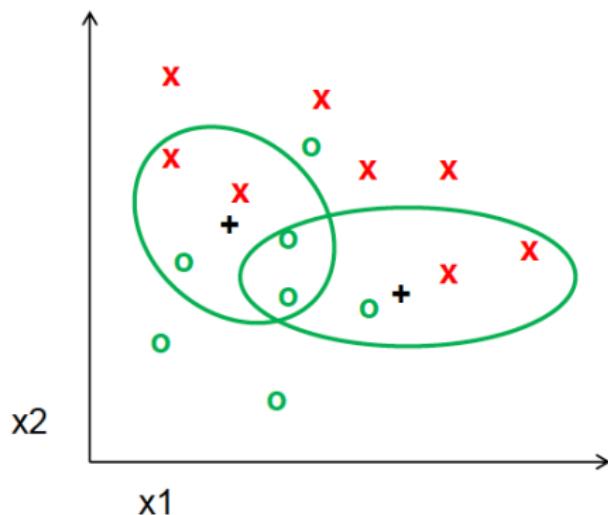
Classifieur des plus proches voisins (1-NN)



Classifieur des plus proches voisins (3-NN)



Classifieur des plus proches voisins (5-NN)

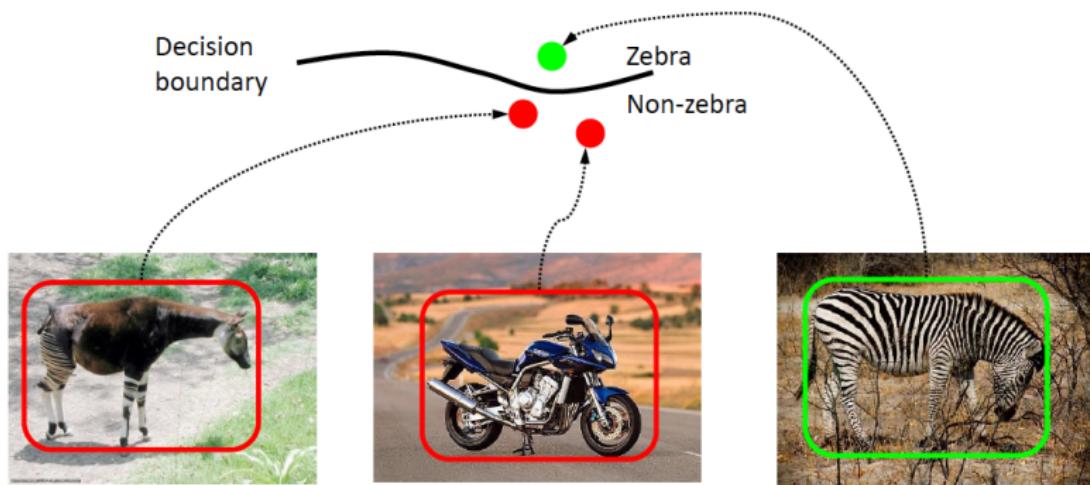


Classifieur des plus proches voisins (k-NN)

- Une approche non paramétrique très simple, que l'on peut toujours essayer en premier.
- Choix de k
- Pas de temps d'apprentissage.
- possibilité de pondérer par la distance/similarité

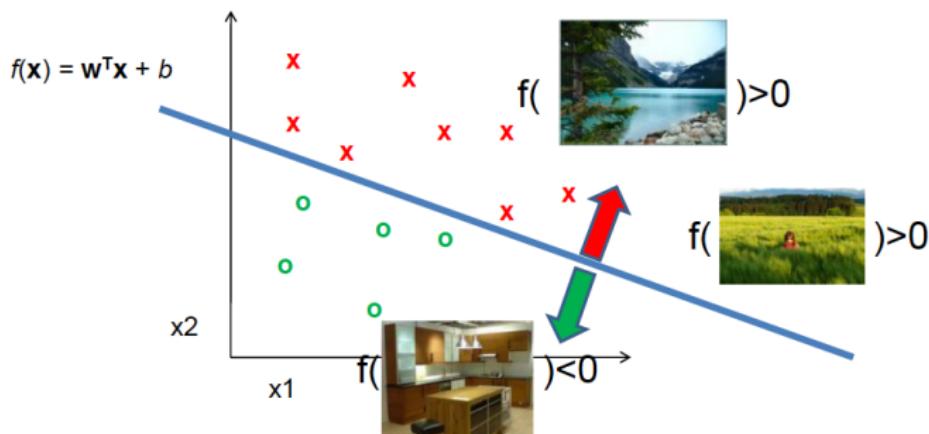
Approches discriminatives

Apprendre une règle de décision (classifieur) qui associe à une représentation (par exemple sac de mots), une ou plusieurs classes.



Exemple : classifieur linéaire

Trouver l'hyperplan séparant les exemples des différents classes (ou catégories) dans l'espace de description



Séparateur linéaire binaire

But

- Données d'apprentissage $D_n = \{(x_i, y_i)\}_{1 \leq i \leq n}, x_i \in \mathcal{X}, y_i \in \{-1, +1\}$, ensemble de points étiquetés.
- On cherche à construire à partir de D_n une fonction de décision $f : \mathcal{X} \rightarrow \{-1, 1\}$ ou $f : \mathcal{X} \rightarrow \mathbb{R}$ qui permet de prédire la classe -1 ou 1 d'un point $x \in \mathcal{X}$.

Fonction de décision

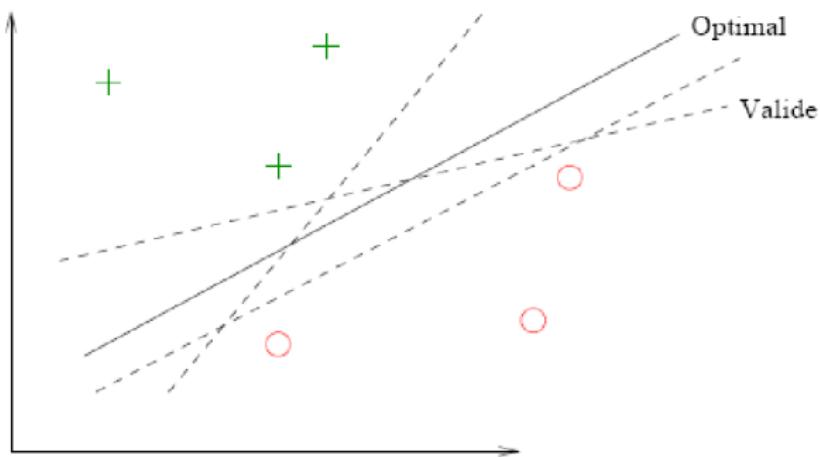
- $\mathcal{X} = \mathbb{R}^d$ et $x = (x^{(1)}, x^{(2)}, \dots, x^{(d)})^T$
- Fonction de décision $f : \mathbb{R}^d \rightarrow \mathbb{R}$ telle que x soit affecté à la classe -1 si $f(x) < 0$ et à la classe $+1$ sinon.
- Fonction de décision linéaire :

$$f(x) = \sum_{j=1}^d w_j x^{(j)} + b = \mathbf{w}^T \mathbf{x} + b, \quad \mathbf{w} \in \mathbb{R}^d, b \in \mathbb{R}$$

$\langle \mathbf{w}, \mathbf{x} \rangle + b$ est l'équation d'un hyperplan qui sépare \mathcal{X} en deux demi-espaces correspondant aux deux classes.

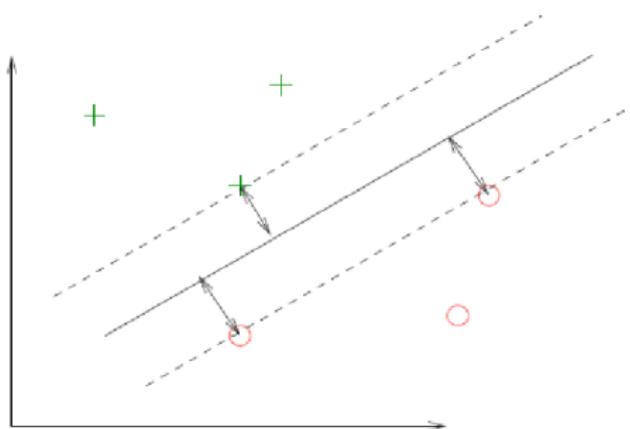
Séparateur linéaire : exemple dans \mathbb{R}^2

Supposons qu'on puisse séparer les deux classes par un classifier linéaire :
Plusieurs séparateurs sont possibles



Séparateur linéaire : exemple dans \mathbb{R}^2

- Hyperplan qui classifie correctement les données et qui se trouve *le plus loin possible de tous les exemples.*
- Hyperplan de marge maximale ($\frac{1}{2}$ marge = distance minimale entre un exemple et la surface de séparation)



Formulation du problème de maximisation de la marge

Séparateur à vaste marge : formulation (primale)

- $\mathcal{D} = \{(x_i, y_i)\}_{i=1..n}$: ensemble de points linéairement séparables.
- Objectif : trouver un hyperplan qui maximise la marge et discrimine correctement les points de \mathcal{D}

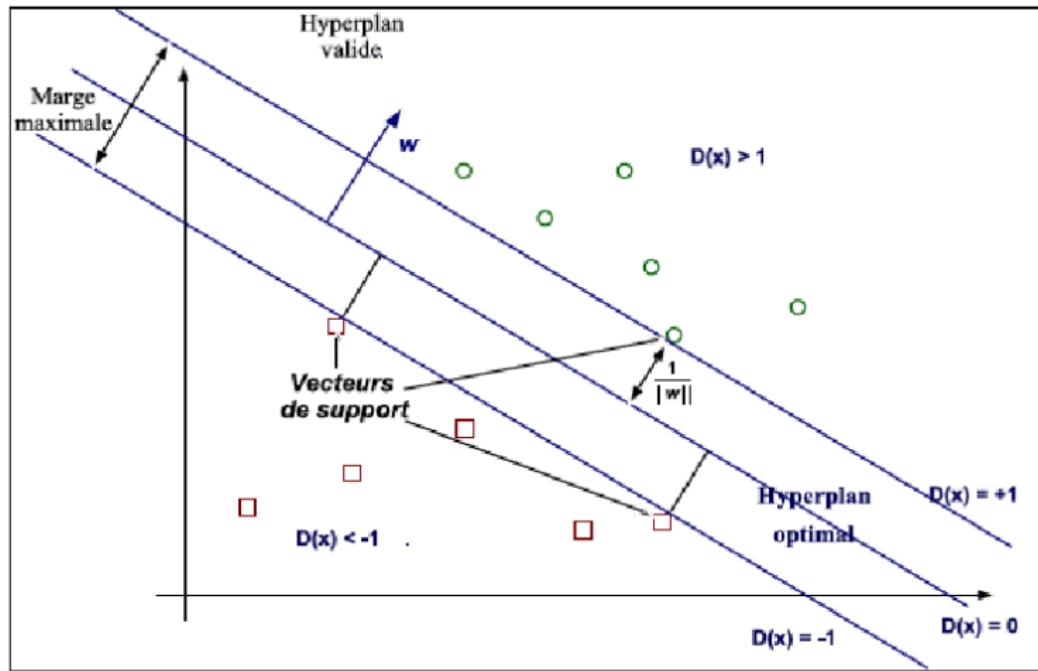
$$\min_{w,b} \frac{1}{2} \|w\|^2 \text{ maximisation de la marge}$$

s.c. $y_i(w^T x_i + b) \geq 1 \quad \forall i = 1, \dots, n$ tous les points bien classés

Problème de minimisation sous contraintes qui peut être résolu par des approches numériques comme la programmation quadratique (minimiser le carré de la norme). On obtient le vecteur solution w^* et b^* définissant la fonction de décision :

$$f(x) = \operatorname{sign}\{w^{*T} x + b^*\}$$

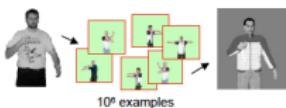
Vecteurs supports



Approches discriminatives

De nombreuses approches.

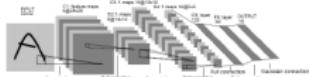
Nearest neighbor



Shakhnarovich, Viola, Darrell 2003
Berg, Berg, Malik 2005

...

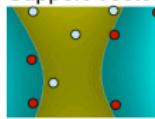
Neural networks



LeCun, Bottou, Bengio, Haffner 1998
Rowley, Baluja, Kanade 1998

...

Support Vector Machines and Kernels



Guyon, Vapnik
Heisele, Serre, Poggio, 2001

...

Conditional Random Fields

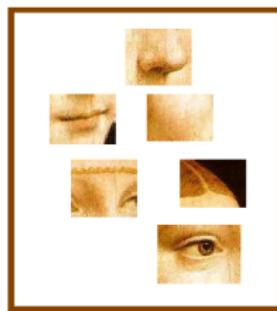


McCallum, Freitag, Pereira 2000
Kumar, Hebert 2003

...

Approches génératives

Modéliser la probabilité d'un ensemble de descripteurs étant donnée une classe



Souvent inspirées de méthodes développées dans le contexte du texte et de la recherche d'information.

Exemple 1 : classifieur de bayes naïf

- On fait l'hypothèse que les caractéristiques sont indépendantes.

$$p(f_1, \dots, f_n | c) = \prod_{i=1}^N p(f_i | c) = \prod_{j=1}^M p(w_j | c)^{n(w_j)}$$

- f_i : i ème caractéristique.
- N : nombre de caractéristiques
- w_j : j ème mot visuel du vocabulaire.
- M : taille du vocabulaire visuel.
- $n(w_j)$: nombre de caractéristiques de type w_j dans l'image.
- $p(w_j | c) = \frac{\text{Nb features } w_j \text{ in training images of classe } c+1}{\text{Nb features in training images of classe } c+M}$

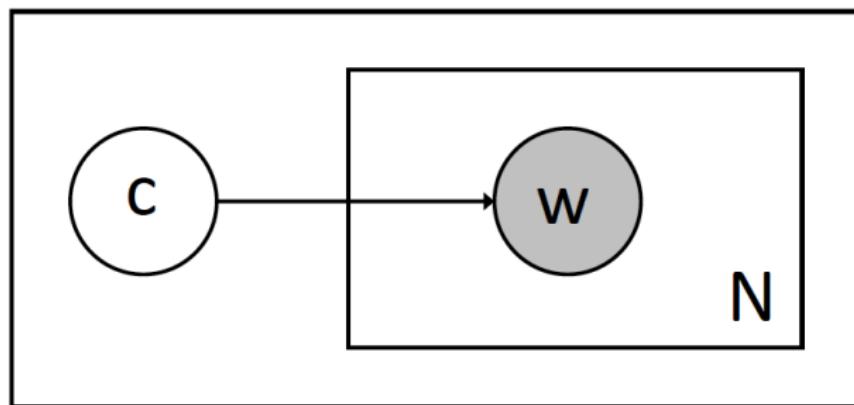
Exemple 1 : classifieur de bayes naïf

- Décision selon le maximum a posteriori :

$$c^* = \operatorname{argmax}_c p(c) \prod_{j=1}^M p(w_j|c)^{n(w_j)}$$

Exemple 1 : classifieur de bayes naïf

Modèle graphique correspondant



Exemple 2 : analyse en sémantique latente probabiliste



Image

$$= p_1$$



zebra

$$+ p_2$$



grass

$$+ p_3$$



tree



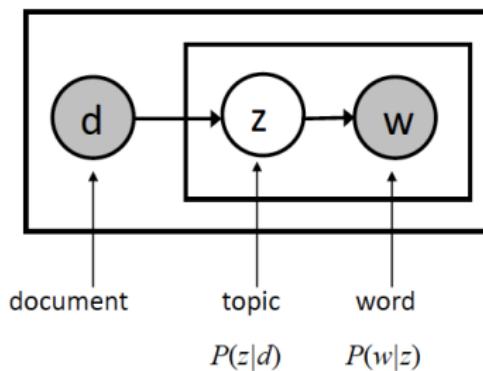
“visual topics”



T. Hofmann, [Probabilistic Latent Semantic Analysis](#), UAI 1999

Exemple 2 : analyse en sémantique latente probabiliste

Modèle graphique correspondant



Modèle génératif à deux niveaux = un document est un mélange de sujets (topics) et chaque sujet a sa propre distribution de mots visuels.

$$p(w_i|d_j) = \sum_{k=1}^K p(w_i|z_k)p(z_k|d_j)$$

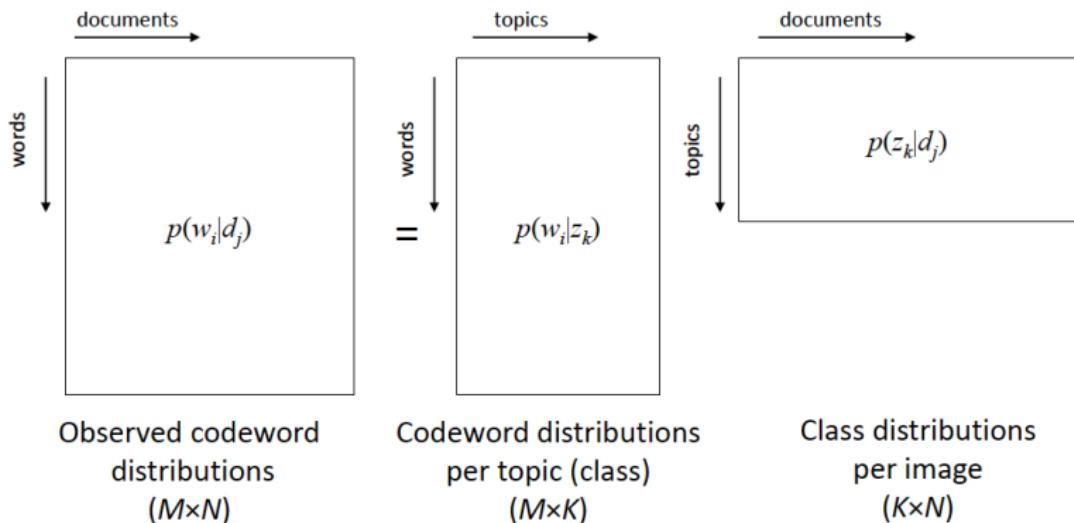
Exemple 2 : analyse en sémantique latente probabiliste

$$p(w_i | d_j) = \sum_{k=1}^K p(w_i | z_k) p(z_k | d_j)$$

Probability of word i in document j (known) Probability of word i given topic k (unknown) Probability of topic k given document j (unknown)

Exemple 2 : analyse en sémantique latente probabiliste

$$p(w_i | d_j) = \sum_{k=1}^K p(w_i | z_k) p(z_k | d_j)$$



Exemple 2 : analyse en sémantique latente probabiliste

Apprentissage des paramètres par maximisation de la vraisemblance des données

$$L = \prod_{i=1}^M \prod_{j=1}^N P(w_i|d_j)^{n(w_i, d_j)}$$

Observed counts of
word i in document j

M ... number of codewords

N ... number of images

$\sum_{k=1}^K P(z_k|d_j)P(w_i|z_k)$

Exemple 2 : analyse en sémantique latente probabiliste

Inférence

- Trouver le sujet le plus vraisemblable pour l'image :

$$z^* = \operatorname{argmax}_z p(z|d)$$

- Trouver le sujet le plus vraisemblable pour un mot visuel dans une image donné :

$$z^* = \operatorname{argmax}_z p(z|w, d) = \operatorname{argmax}_z \frac{p(w|z)p(z|d)}{\sum_{z'} p(w|z')p(z'|d)}$$

Approches génératives

Bilan

- Clasifieur bayesien naïf :
 - ▶ Modèle unigram en analyse de documents.
 - ▶ Hypothèse forte de l'indépendance des mots étant donnée une classe.
 - ▶ Estimation simple des paramètres : comptage d'occurrences.
- Analyse sémantique latente probabiliste
 - ▶ Chaque document est un mélange de sujets (mélange de classes)
 - ▶ Problème de décomposition de matrices.
 - ▶ Estimation des paramètres : algorithme EM.

Plan

1 Introduction

2 Description d'images

- Descripteurs basiques
- SIFT
- Gist

3 Petite parenthèse : problème du panorama

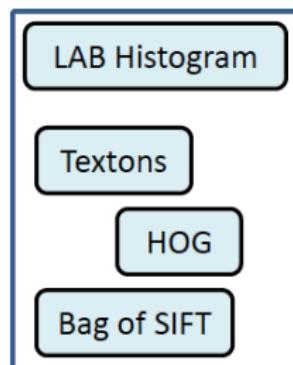
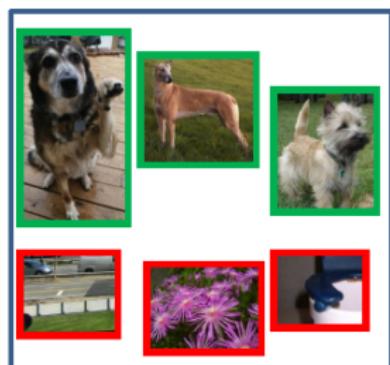
- Bow

4 Classification

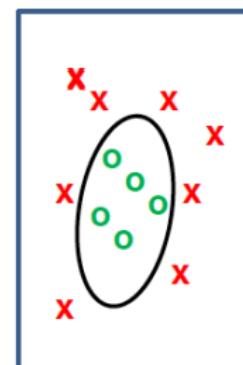
5 Conclusion

Bilan

Une chaîne simple de reconnaissance



+ Image Features



+ Classifier

Bilan

OpenCV

- Détection et description de caractéristiques

http://opencv-python-tutroals.readthedocs.org/en/latest/py_tutorials/py_feature2d/py_table_of_contents_feature2d/py_table_of_contents_feature2d.html

- Machine learning

http://opencv-python-tutroals.readthedocs.org/en/latest/py_tutorials/py_ml/py_table_of_contents_ml/py_table_of_contents_ml.html

Synthèse

- Problèmes de reconnaissance (catégorisation, détection, segmentation...)
- Difficultés de la reconnaissance
- Niveaux de catégorisation
- Principe des méthodes discriminatives (classif. supervisée)
chaîne de reconnaissance : train/test description/classification
- descripteur SIFT
- descripteur BoW
- descripteur GIST
- classifieur kNN
- classifieur SVM (et one-vs-all)

Et toutes les approches à bases de réseaux de neurones sous forme d'architectures profondes

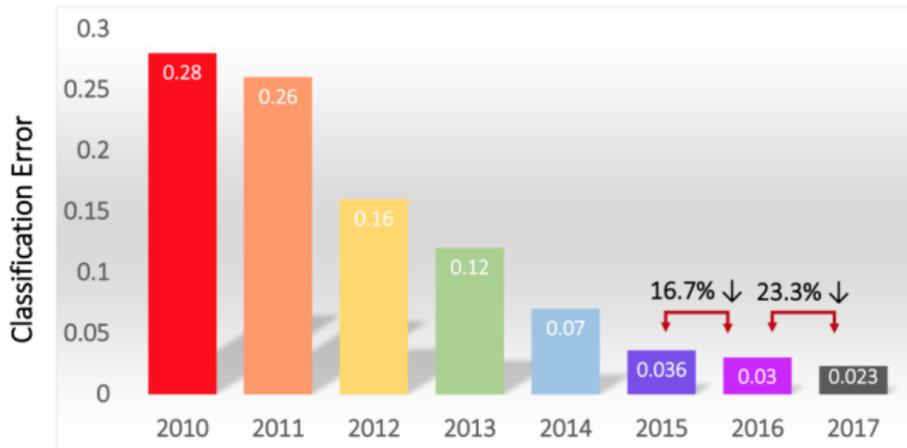
Pour aller plus loin

- Chapitre 14 du livre de Szeliski : Computer Vision : Algorithms and applications.
- Tutoriaux de Torralba, Li et Fergus : Recognizing and Learning Object Categories
<http://people.csail.mit.edu/torralba/shortCourseRLOC/index.html>
- Visual recognition and learning course, Université de Washington
<http://courses.cs.washington.edu/courses/cse599v1/13wi/>
- ...

La reconnaissance : un problème encore ouvert !

Résultat pour la tâche de classification pour challenge ILSVRC (1000 categories, 1000 images par classe pour l'entraînement, 100k images pour le test)

Classification Results (CLS)



S.o.t.a : (Hu et al, 2018) Squeeze-and-Excitation Networks, CVPR 2018

<https://arxiv.org/abs/1709.01507>

Omniprésence de l'apprentissage de représentation sur l'approche **hand-crafted features**.

Pourquoi l'approche hand-crafted est encore intéressante ?

Pour l'apprentissage auto-supervisé (manque de données d'annotation)

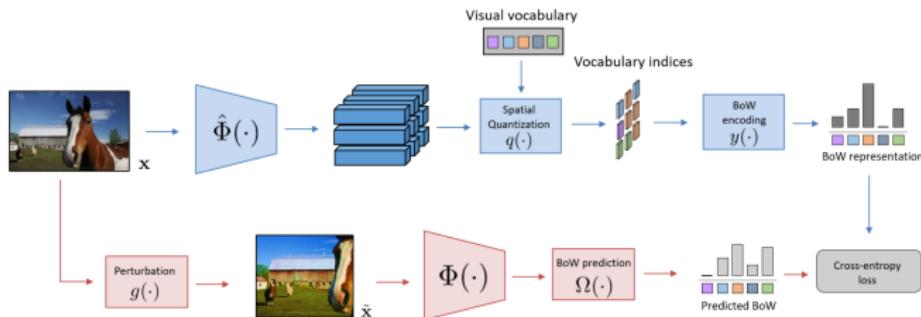


Figure 1: **Learning representations through prediction of Bags of Visual Words.** We first train a feature extractor $\hat{\Phi}(\cdot)$ for a self-supervised task, e.g. rotation prediction. Then we compute a visual vocabulary from feature vectors computed from $\hat{\Phi}$ feature maps and compute the corresponding image level BoW vectors. These BoW vectors will serve as ground truth for the next stage. In the second stage we perturb images with $g(\cdot)$ and send them as input to a second network $\Phi(\cdot)$. The BoW prediction module $\Omega(\cdot)$ processes $\Phi(\cdot)$ feature maps to predict BoW vectors corresponding to the original non-perturbed images. Both $\Phi(\cdot)$ and $\Omega(\cdot)$ are trained jointly with cross-entropy loss. The feature extractor $\Phi(\cdot)$ is further used for downstream tasks.

(Gidaris et al, CVPR 2020) Learning Representations by Predicting Bags of Visual Words.⁴

4. <https://arxiv.org/pdf/2002.12247.pdf>

Pourquoi l'approche hand-crafted est encore intéressante ?

Pour des tâches de mise en correspondance de descripteurs locaux.

		# Images	# Registered	# Sparse Points	# Observations	Track Length	Reproj. Error	# Inlier Pairs	# Inlier Matches	# Dense Points	Pose Error	Dense Error
Fountain	SIFT	11	11	10,004	44K	4.49	0.30px	49	76K	2.970K	0.002m (0.002m)	0.77 (0.90)
	SIFT-PCA	11		14,608	70K	4.80	0.39px	55	124K	3.021K	0.002m (0.002m)	0.77 (0.90)
	DSP-SIFT	11		14,785	71K	4.80	0.41px	54	129K	2.999K	0.002m (0.002m)	0.77 (0.90)
	ComOpt	11		14,179	67K	4.75	0.37px	55	114K	2.999K	0.002m (0.002m)	0.77 (0.90)
	DeepDesc	11		13,519	61K	4.55	0.35px	55	93K	2.972K	0.002m (0.002m)	0.77 (0.90)
	TFeat	11		13,696	64K	4.68	0.35px	54	103K	2.969K	0.002m (0.002m)	0.77 (0.90)
	LIFT	11		10,172	46K	4.55	0.59px	55	83K	3.019K	0.002m (0.002m)	0.77 (0.90)
Herzjesu	SIFT	8	8	4,916	19K	4.00	0.32px	27	28K	2.373K	0.004m (0.004m)	0.57 (0.73)
	SIFT-PCA	8		7,433	31K	4.19	0.42px	28	47K	2.374K	0.004m (0.004m)	0.57 (0.73)
	DSP-SIFT	8		7,760	32K	4.19	0.45px	28	50K	2.374K	0.004m (0.004m)	0.57 (0.73)
	ComOpt	8		6,939	28K	4.13	0.40px	28	42K	2.375K	0.004m (0.004m)	0.57 (0.73)
	DeepDesc	8		6,418	25K	3.92	0.38px	28	34K	2.380K	0.004m (0.004m)	0.57 (0.73)
	TFeat	8		6,606	27K	4.09	0.38px	28	38K	2.377K	0.004m (0.004m)	0.57 (0.73)
	LIFT	8		7,834	30K	3.95	0.63px	28	46K	2.375K	0.004m (0.004m)	0.57 (0.73)
South Building	SIFT	128	128	62,780	353K	5.64	0.42px	1K	1,003K	1.972K	—	—
	SIFT-PCA	128		107,674	650K	6.04	0.54px	3K	2,019K	1.993K	—	—
	DSP-SIFT	128		110,394	664K	6.02	0.57px	3K	2,079K	1.994K	—	—
	ComOpt	128		103,602	617K	5.96	0.51px	4K	1,856K	2,007K	—	—
	DeepDesc	128		101,154	558K	5.53	0.48px	6K	1,463K	2,002K	—	—
	TFeat	128		94,589	566K	5.99	0.49px	3K	1,567K	1,960K	—	—
	LIFT	128		74,607	399K	5.35	0.78px	3K	1,168K	1.975K	—	—
Madrid Metropolis	SIFT	1,344	440	62,729	416K	6.64	0.53px	14K	1,740K	435K	—	—
	SIFT-PCA		465	119,244	702K	5.89	0.57px	27K	3,597K	537K	—	—
	DSP-SIFT		470	107,329	661K	5.26	0.64px	21K	3,155K	579K	—	—
	ComOpt		485	115,134	634K	5.51	0.57px	29K	3,148K	561K	—	—
	DeepDesc		377	68,110	348K	5.11	0.53px	19K	1,570K	516K	—	—
	TFeat		439	90,274	512K	5.68	0.54px	18K	2,135K	522K	—	—
	LIFT		430	52,755	337K	6.40	0.76px	13K	1,498K	450K	—	—

(Schonberger et al, CVPR 2017) Comparative Evaluation of Hand-Crafted and Learned Local Features.⁵

5. https://openaccess.thecvf.com/content_cvpr_2017/papers/Schonberger_Comparative_Evaluation_of_CVPR_2017_paper.pdf

Pourquoi l'approche hand-crafted est encore intéressante ?

Neural Activation Constellations : rencontre entre les approches à base d'apprentissage profond et les modèles par parties (constellation models)

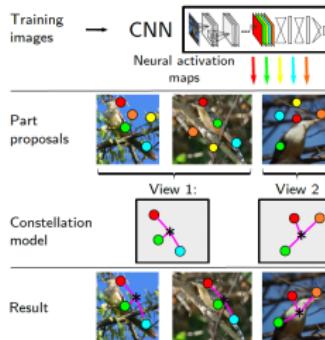


Figure 1. Overview of our approach. Deep neural activation maps are used to exploit the channels of a CNN as a part detector. We estimate a part model from completely unsupervised data by selecting part detectors that fire at similar relative locations. The created part models are then used to extract features at object parts for weakly-supervised classification.

(Simon et al, ICCV 2015) Neural Activation Constellations : Unsupervised Part Model Discovery with Convolutional Networks.⁶

6. <https://pub.inf-cv.uni-jena.de/pdf/Simon15:NAC.pdf>