Modélisation et Reconnaissance des formes
Représenter et comparer des formes

Marie-Odile Berger
http://members.loria.fr/moberger

20 septembre 2023

Première partie I

Quelques considérations générales

Problèmes de représentation

- la taille : Les données sont rarement de taille raisonnable (spectrogramme, images,. . .) $\rightarrow$ il faut adopter une représentation des données de taille raisonnable avec le moins possible de perte d'information
- l'invariance : les données peuvent être enregistrées dans des repères différents (ex orientation différente). Les données peuvent aussi être des mesures indirectes d'un même phénomène : les mesure ne sont donc pas directement semblables même si elles concernent un même phénomène
  - Utiliser des mesures invariantes (ou le plus possible) pour caractériser des formes.
  - dans les cas complexes, il n'y a pas de caractéristiques évidentes résumant au mieux les données et tenant compte des variations d'apparence. Celles ci **doivent être apprises**.

Type d'invariance visé

Il peut y avoir de simples mouvements de l'objet, des changements de points de vue, des changements d'illumination, des occultations....

Problème (**malédiction (*)**) de la dimension

* : terme inventé par Richard Bellman pour parle de la difficulté de travailler avec des données appartenant à des espaces de grande dimension.

- Représenter une forme par un vecteur de caractéristiques de **petite taille** permet de limiter la complexité des processus
- Un grand vecteur de caractéristiques peut avoir tendance à modéliser l'accessoire (le bruit) plutôt que l'essentiel des données.
- malédiction : il faut énormément de données pour obtenir une bonne estimation. Soient 100 observations d'un phénomène faites dans l'intervalle $[0, 1]$. Pour réaliser dans $[0, 1]^{10}$ une couverture équivalente à celle des 100 points il faudrait $100^{10} = 10^{20}$ observations, ce qui est la plupart du temps inenvisageable.

Représenter les données

Objectifs :

- représenter les formes de manière compacte et discriminante
- avoir des métriques pour évaluer leur similarité

Deux grandes classes de représentation

- Représentations explicites (handcrafted) créées à la main en fonction des connaissances sur le domaine
- Représentations et métriques de comparaison issues des techniques d'apprentissage

Difficulté invariance souhaitée vis à vis de l'espace de mesure (rotation, translation et changement d'échelle) et de changements des conditions d'observation des objets.

Deuxième partie II

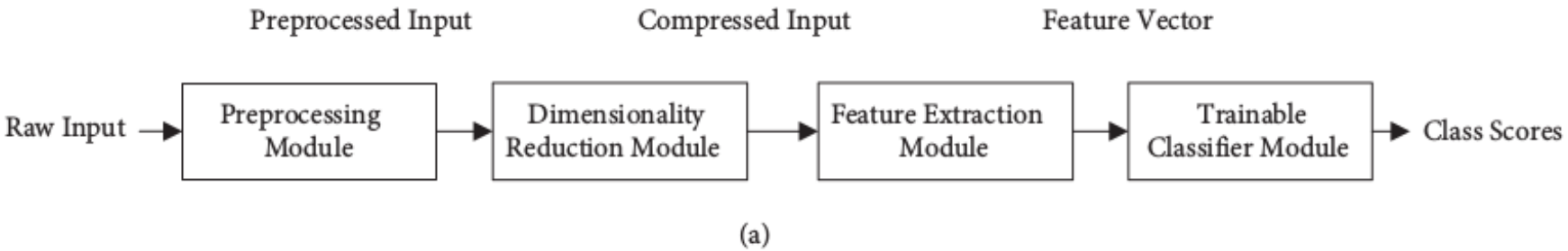
Les approches classiques (Handcrafted) de représentation des formes

Les approches classiques

- classique, explicite, historique...
- la représentation des formes est naturelle ou vient de connaissances a priori sur les formes considérées
- exemples :
  - vecteur de paramètres,
  - courbe, courbe paramétrée
  - histogramme
  - ensembles d'indices spécifiques (empreintes digitales, visages,...)

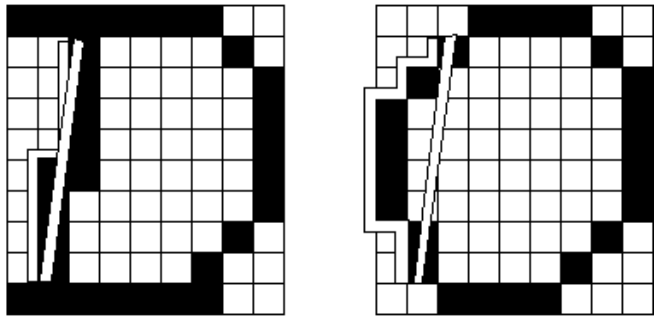
Un système classique de reconnaissance des formes

l'approche conventionnelle :



- Les caractéristiques des données sont extraites (de manière statique) **indépendamment** du processus de classification

Exemples de représentations :



caractéristiques possibles : aire, périmètre, compacité, histogramme . . .

Exemples de représentation

Empreinte : extraire des données caractéristiques comme les birfurcations et les points terminaux.

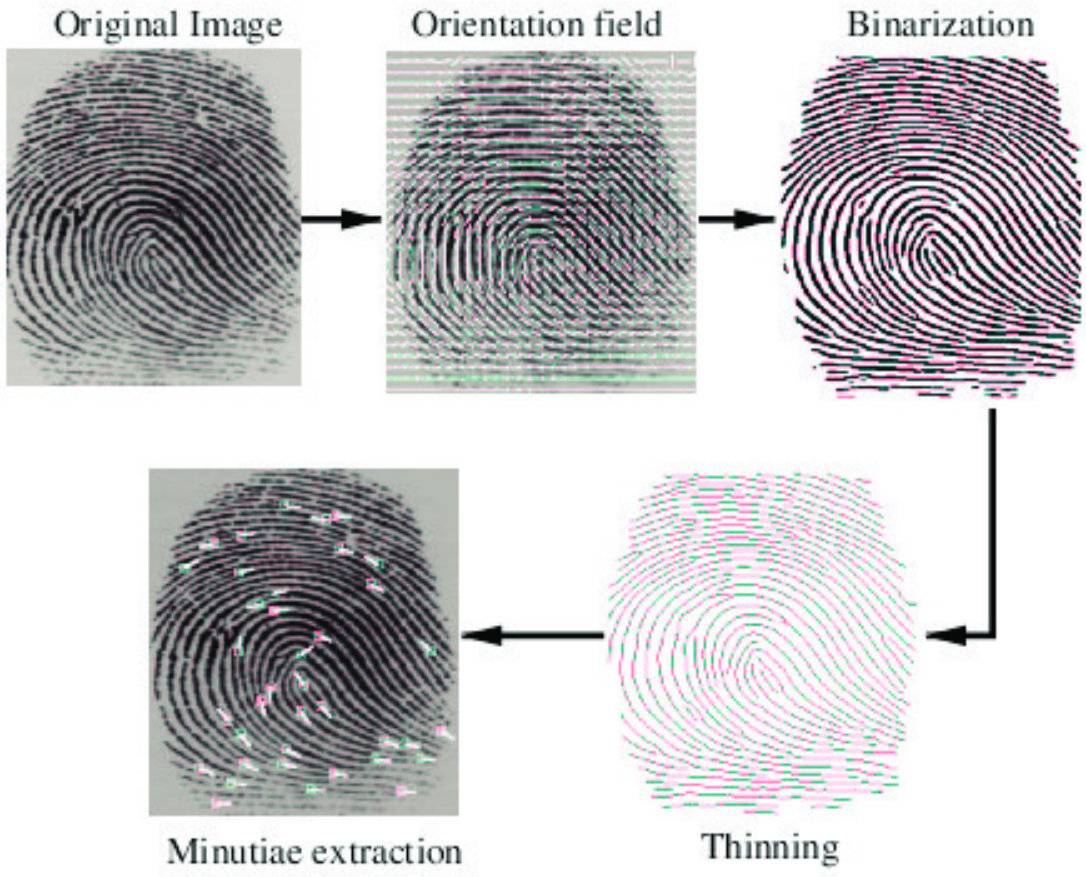


Figure 3: Various stages in a typical minutiae extraction algorithm [1].

Quelle distance pour comparer les formes ?

Si la forme est un ensemble de points, sans structure spécifique :

- pour les représentations par un vecteur dans $\mathbb{R}^k$, on utilise les normes classiques, par exemple
  - norme L_2,

$$||x||_2 = \sqrt{\sum_1^k x_i^2}$$

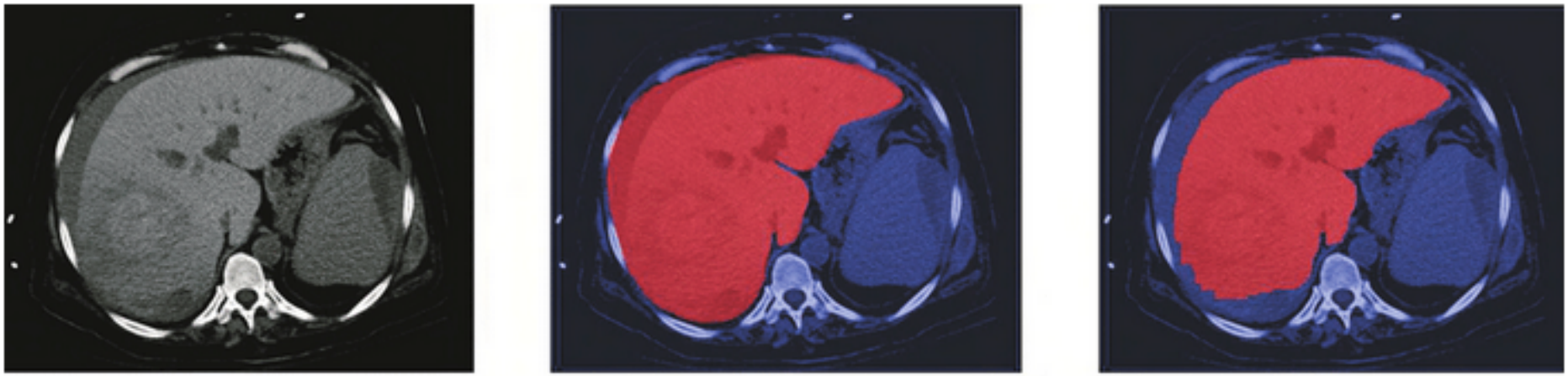
  - norme L_1,

$$||x||_1 = \sum |x_i|$$

Voir le cours sur l'estimation robuste pour comprendre le comportement de ces normes vis à vis de données aberrantes. L_1 est mieux adaptée en cas de données aberrantes mais elle n'est pas dérivable.

Et si on compare des objets complexes ?

exemple : valider une segmentation. Il faut comparer la forme obtenue à une vérité terrain



(a) image (b) vérité terrain (c) résultat d'un algorithme de segmentation

Comparer des formes binaires pour une segmentation

On s'intéresse au recouvrement des deux surfaces (Intersection Over Union)

- comparaisons ensemblistes : métrique de DICE, de jaccard

$$jaccard(A, B) = IoU(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

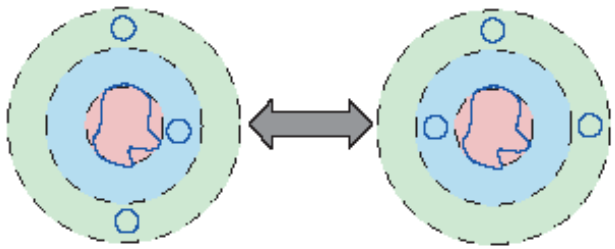
Différence entre distance et similarité

Mathématiquement, Une **distance** est une fonction à valeur positive qui vérifie les propriétés

1. symmétrie : $d(x, y) = d(y, x)$
2. $d(x, y) = 0 \rightarrow x = y$
3. l'inégalité triangulaire : $d(x, y) < d(x, z) + d(z, y)$

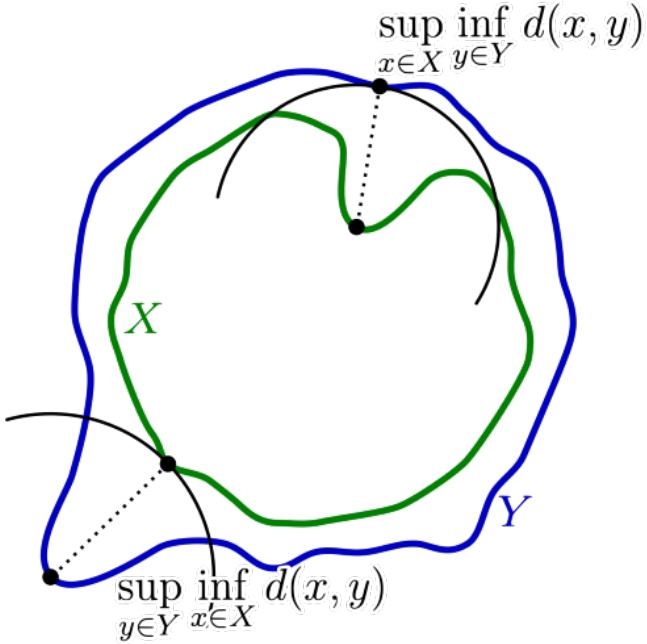
La propriété (2) est rarement vérifiée par les mesures de similarité.

- ex 1 : deux zones d'une image peuvent avoir le même histogramme sans être identiques
- utiliser l'angle du contour par rapport au rayon issu du centre de gravité et passant par ce point. Inconvénient : deux formes différentes peuvent partager la même signature



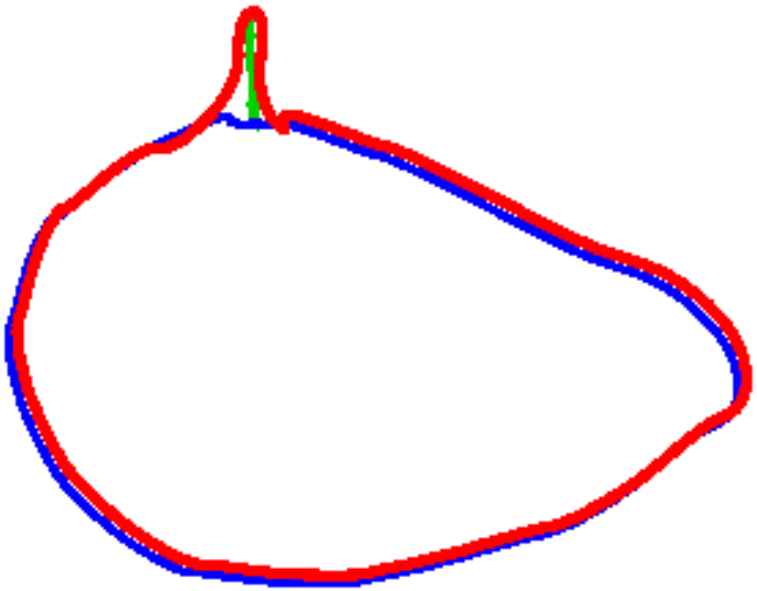
Comparaison entre formes : la distance de Hausdorff

- le recouvrement n'a pas de sens pour qualifier la distance entre contours. Recours à des distances ensemblistes
- La distance de Hausdorff (qui ne vérifie pas l'inégalité triangulaire)
  - distance entre un point et un ensemble :$d(x, Y) = min_{y \in Y} d(x, y)$
  - distance entre deux ensembles : $d_H(X, Y) = max_{x \in X} d(x, Y)$... mais ce n'est pas symmétrique...
  - distance de Hausdorff
    $d_H(X, Y) = max\{max_{x \in X} d(x, Y), max_{y \in Y} d(y, X)\}$



Robustesse de Haussdorf ?

La mesure de Hausdorff n'est pas robuste à de petites variation locales de formes

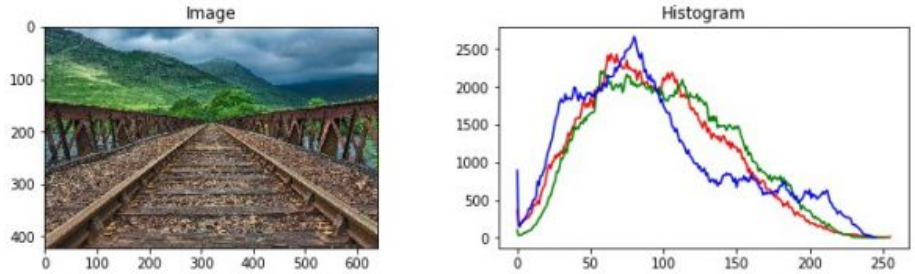


- différence seulement locale qui impacte la distance de Hausdorff
- divers moyens d'amélioration : $d(X,Y) = \frac{1}{|X|} \sum_{x \in X} min_{y \in Y}\{d(x,y\}$ ou utiliser des méthodes robustes

Troisième partie III

L'histogramme comme descripteur

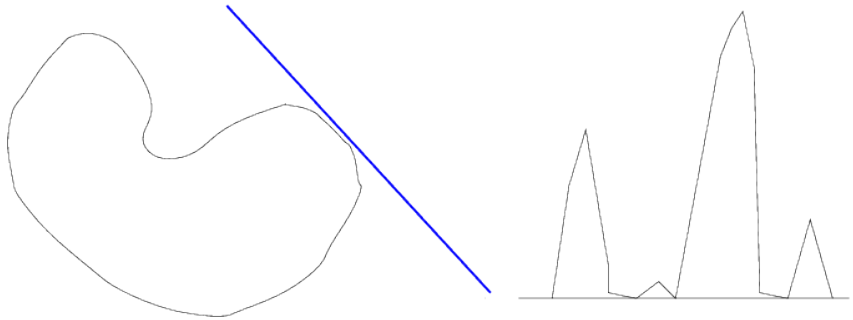
Utilisation des histogrammes comme descripteurs d'une forme ou d'une image



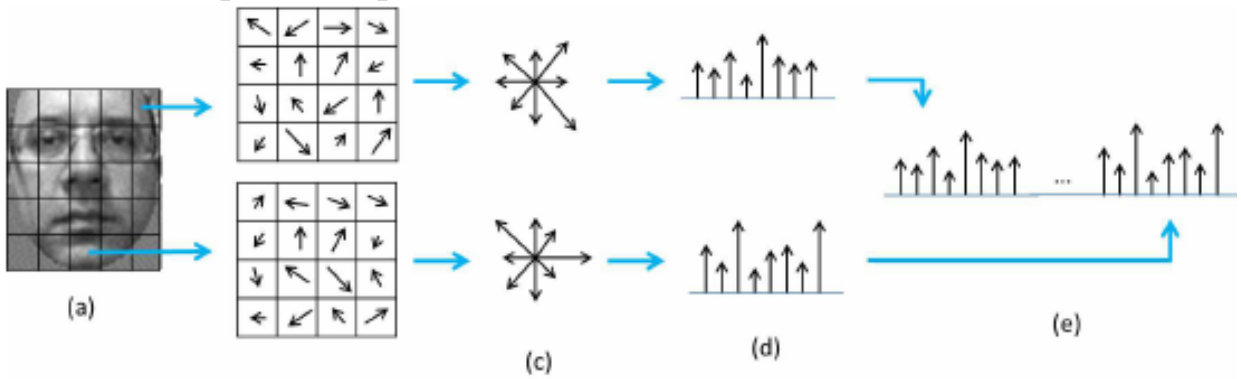
- un histogramme représente la répartition d'une variable continue à l'intérieur d'intervalles de valeurs appelés *bin* ou *baquets*.
- on peut représenter l'effectif d'un baquet ou représenter sa fréquence : l'histogramme est alors une représentation de la proba
- il permet une représentation compacte d'un objet
- représentation généralement non injective : plusieurs formes différentes partagent le même histogramme. Perte de l'information de spatialité.
- souvent utilisé pour constituer rapidement une short liste des formes candidates. Une étude plus fine est ensuite faite sur la short liste pour juger de la ressemblance. Voir par exemple [MBM01]

Exemples de représentation avec des histogrammes

- **codage de la pente** Balayer un contour. Construire la distribution $f(x)$, ou l'histogramme de l'angle polaire de la tangente à la courbe.



- caractérisation de régions d'une image : HOG histogramme de gradient orienté [DT05] :



puis classification dans un SVM

Pour comparer des histogrammes

deux méthodologies : comparer les vecteurs h ou comparer les distributions de probabilité associées :

approche par comparaison de vecteurs :

- distance euclienne L_2 ou L_1 utilisables
- distance du Chi 2 :

$$\chi^2(h_1, h_2) = \sum_{i=1}^{i=N} \frac{(h_1(i) - h_2(i))^2}{h_1(i) + h_2(i)}$$

- Les histogrammes sont souvent normalisés $\rightarrow$ (diviser h par $\sum_1^N h(i)$)

  .

  cosinus distance : $1 - \sum h_1(i) * h_2(i)$ (deux vecteurs identiques normalisés ont un produit scalaire égal à 1)
- intersection des histogrammes $=$
  $1 - \sum_{i=1}^{i=N} min(h_1(i), h_2(i)) / \sum_1^N h_1(i)$

Pour comparer des histogrammes

approche par comparaison de vecteurs :

- distance euclienne L_2 ou L_1 utilisables
- distance du Chi 2 :

$$\chi^2(h_1, h_2) = \sum_{i=1}^{i=N} \frac{(h_1(i) - h_2(i))^2}{h_1(i) + h_2(i)}$$

  l'erreur est pondérée par la hauteur du bin : un même écart est jugée moin important si h(i) est grand

- intersection des histogrammes = $1 - \sum_{i=1}^{i=N} min(h_1(i), h_2(i)) / \sum_1^N h_1(i)$ les endroits où $h_1(i) = 0$ ne sont pas considérés

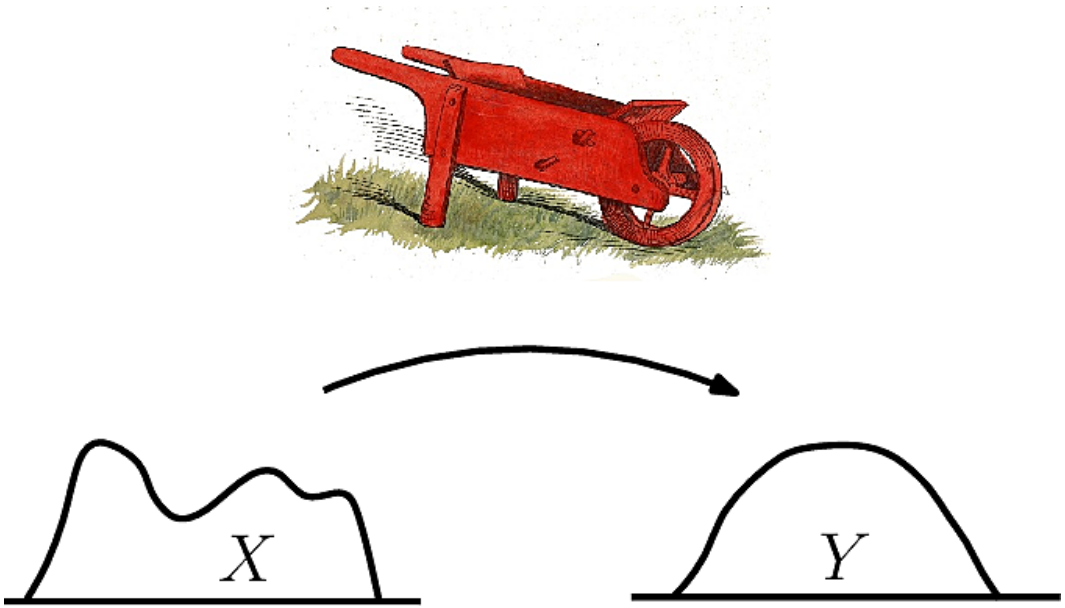
Comparer des histogrammes via les probabilités

On considère des histogrammes normalisés, interprétés comme une distribution de probabilité :

- divergence de Kullback-Leibler entre distributions de probabilité

$$KL(p,q) = \sum_i p(i) log \frac{p(i)}{q(i)}$$

- distance du terrassier [RTG00](métaphore du travail minimum qu'un cantonnier doit fournir pour transformer un tas de terre en un autre)

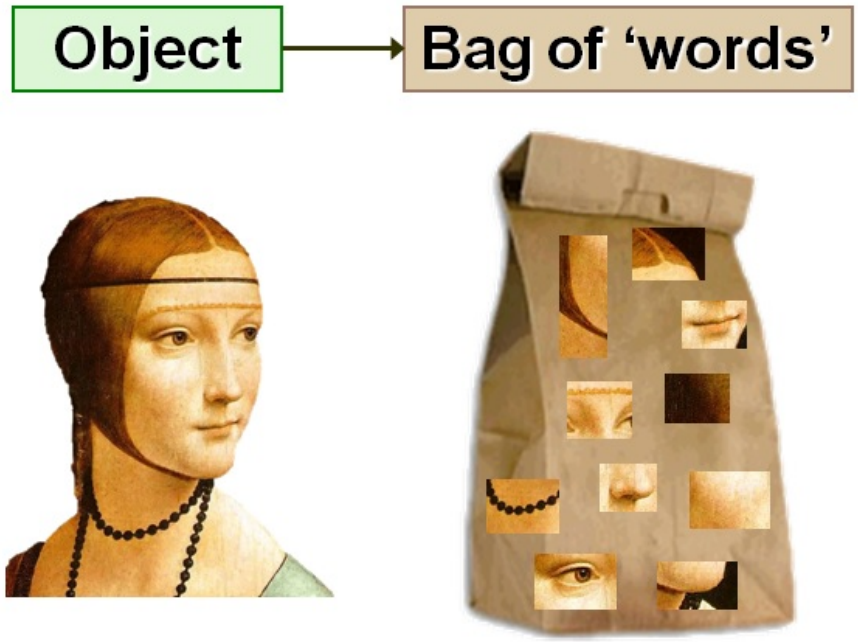


Voir la page terrassier. Un problème de transport optimal. Permet de comparer des distribs qui n'ont pas le même support

voir [CS02] pour une revue des métriques utilisables pour les histogrammes

Représentation par sac de mots (Bag of words)

- Utiliser ou définir un vocabulaire. Définir un vocabulaire visuel
- Représenter une image par un histogramme des mots visuels



- voir Videogoogle [SZ03], le tutoriel de Fei Fei Li.

Quatrième partie IV

Apprendre des représentations et des métriques de similarité

De nouveaux descripteurs appris grâce aux réseaux convolutionnels

l'approche par réseaux convolutionnels (CNN) : extraction des caractéristiques et entraînement du classifieur **ne sont pas dissociés** :

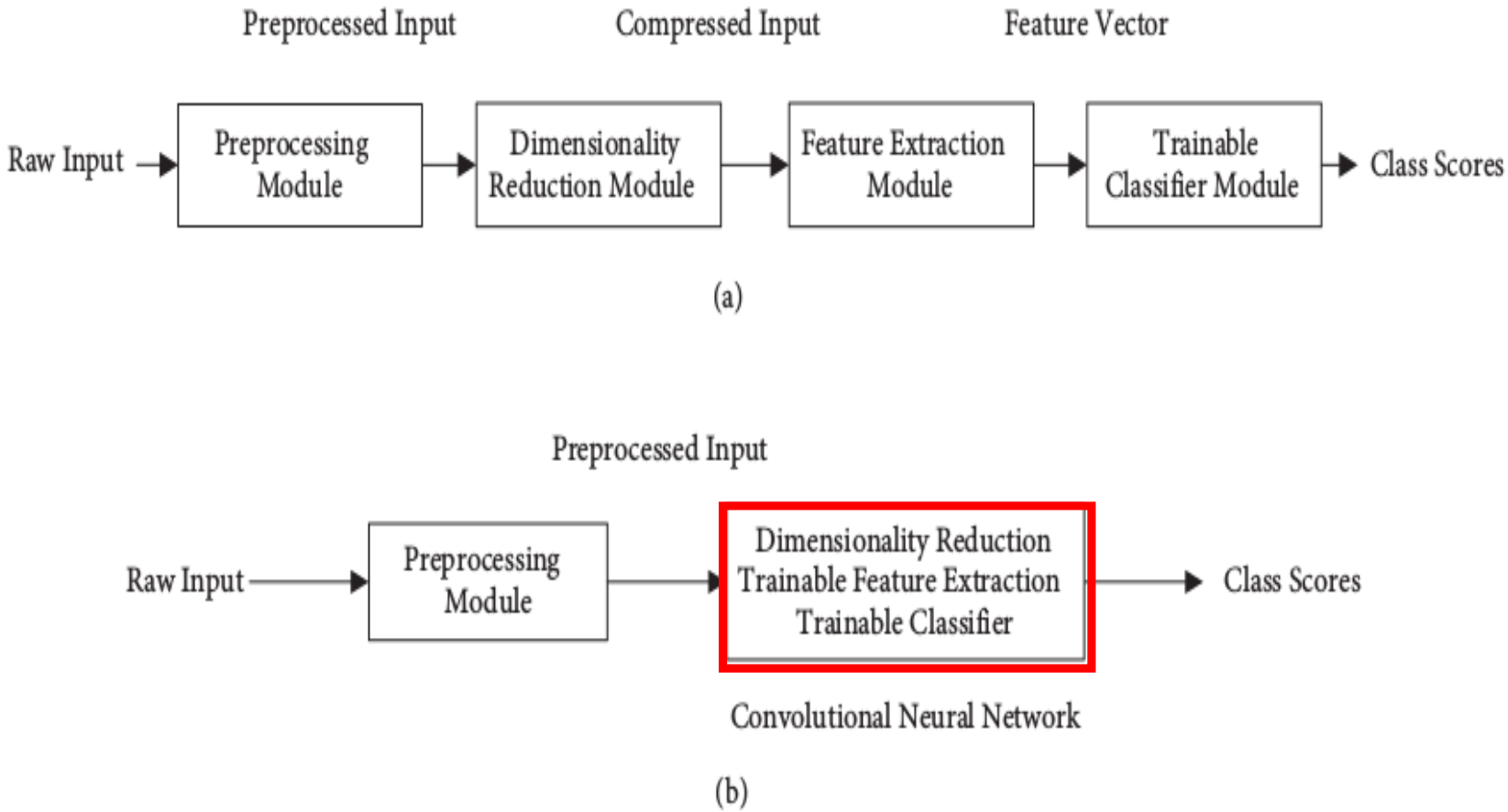


Figure 1. Pattern recognition approaches: (a) conventional, (b) CNN-based.

Un réseau convolutionnel

Exemple d'un des premiers réseaux convolutionnels pour la classification des chiffres : [LBBH98]

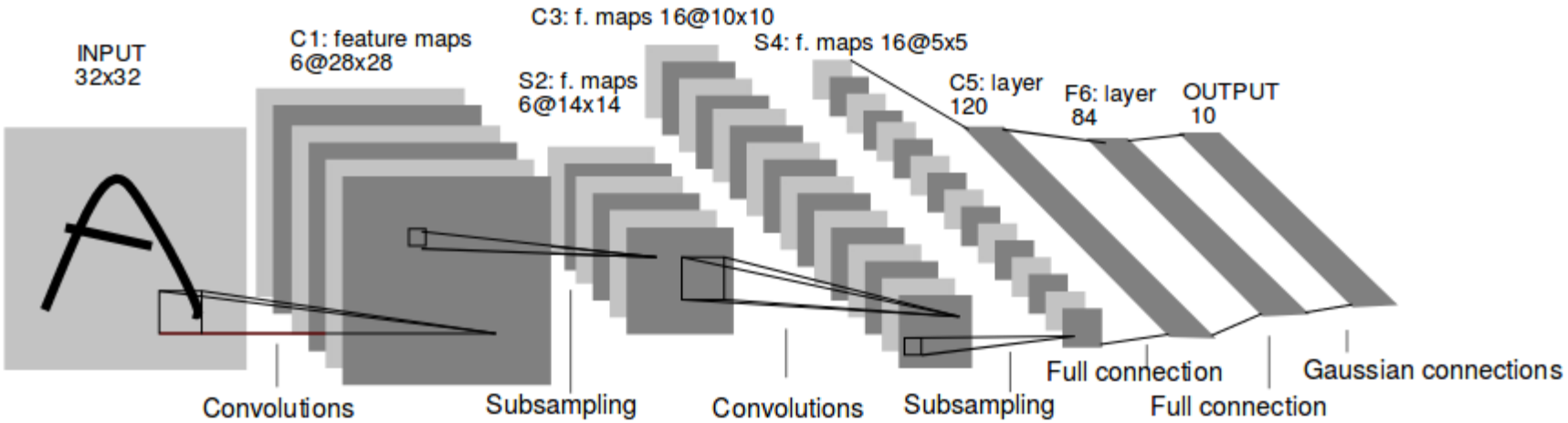


Fig. 2. Architecture of LeNet-5, a Convolutional Neural Network, here for digits recognition. Each plane is a feature map, i.e. a set of units whose weights are constrained to be identical.

Un réseau convolutionnel

Une succession de couches : on pourra consulter cette page de Stanford qui est un pense bête sur la structure des réseaux de neurones convolutionnels

- couches de convolution pour extraire des caractéristiques locales des images, à différentes échelles
- couche de pooling : réduire la taille tout en préservant les informations les plus importantes (average ou max pooling : garder la valeur maximale d'une fenêtre 4x4)
- couches Relu : fonction d'activation visant à introduire des non-linéarités
- couche entièrement connectée (FC) : elle s'aplique sur une entrée "aplatie", donc à la fin des architectures de CNN. Chaque entrée est connectées à tous les neurones. Typiquement utilisée pour calculer les scores de classe.

Comment choisir les paramètres de toutes ces couches ?

Entraînement d'un réseau

Exemple de l'apprentissage supervisé :

- On dispose d'un ensemble de données $\{Z^p\}$ dont la classe D^p (vérité terrain, étiquette) est connue
- Soit W les paramètres ajustables du système (convolution,biais...)
- Pour une entrée Z, le réseau calcule une valeur $F(Z, W)$ qui fournit le mieux possible l'étiquette correcte de Z en sortie
- Les paramètres W du système sont "appris" en minimisant $E_{train}(W) = \sum_p dist(D^p, F(Z^p, W))$, c'est-à-dire l'écart entre la donnée prédite et la donnée attendue [1] sur la base de donnée des exemples disponible pour l'apprentissage.

1. On précisera la forme que prend *dist* un peu plus tard

Les descripteurs "convNet"

- La dernière couche des réseaux avant FC a permis de bien classifier les données !
- C'est donc a priori un bon candidat pour décrire une forme : c'est le descripteur CNN/convnet landmarks
- des descripteurs *génériques*[2] proviennent de réseaux appris sur des grandes base de données Imagenet Large Scale Visual recognition challenge. Alexnet [KSH12], googleLeNet [SLJ+14]... sont des réseaux couramment utilisés pour produire ces descripteurs
- On peut rendre ces descripteurs *invariants* en fournissant des données d'apprentissage d'une forme dans des conditions variées (rotation de l'image, changement des conditions d'illumination (c'est une invariance expérimentale et non formelle).

2. C'est la taille et la diversité des bases de données d'apprentissage qui rend d'une certaine façon le descripteur générique

Utilisation en reconnaissance de lieux

Voir Sunderhof [SSJ+15] : reconnaissance de lieux malgré des points de vue très différents (pour fermeture de boucle)

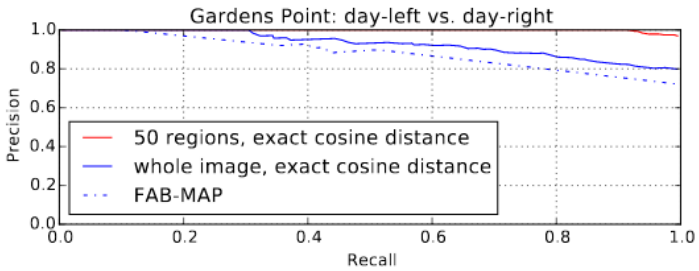


Fig. 3: Results for the Gardens Point Campus dataset. We clearly outperform the whole-image ConvNet-based method proposed by [41] and OpenFABMAP [14].

Fig. 4: Two example scenes from the Gardens Point Campus dataset with extracted and matched ConvNet landmarks. Notice the lateral camera displacement of several meters.

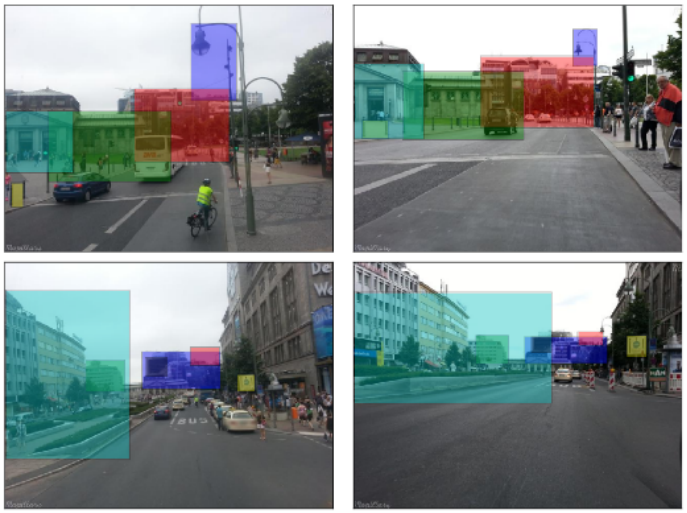


Fig. 5: Examples of successfully matched scenes from the *Berlin Kurfürstendamm* sequence of the Mapillary dataset. Images in a row belong to the same place but have been taken from different viewpoints, i.e. from the bike lane and from the upper deck of a tourist bus. The colored boxes illustrate some of the extracted and correctly matched landmarks.

Reconnaissance de lieux

Couplage

- d'une méthode de *box proposal*(Edge Box [ZD14])
  - afin d'eviter de considérer toutes les boites possibles (sliding window)
  - repérer les fenêtres intéressantes comme étant celles qui contiennent un objet relativement isolé :
    critère :nombre de contours dans la boite - nombre de ces contours qui existent à l'extérieur de la boite
- de mise en correspondance entre ces boites utilisant la ressemblance des descripteurs CNN

Quelques idées sur EdgeBox

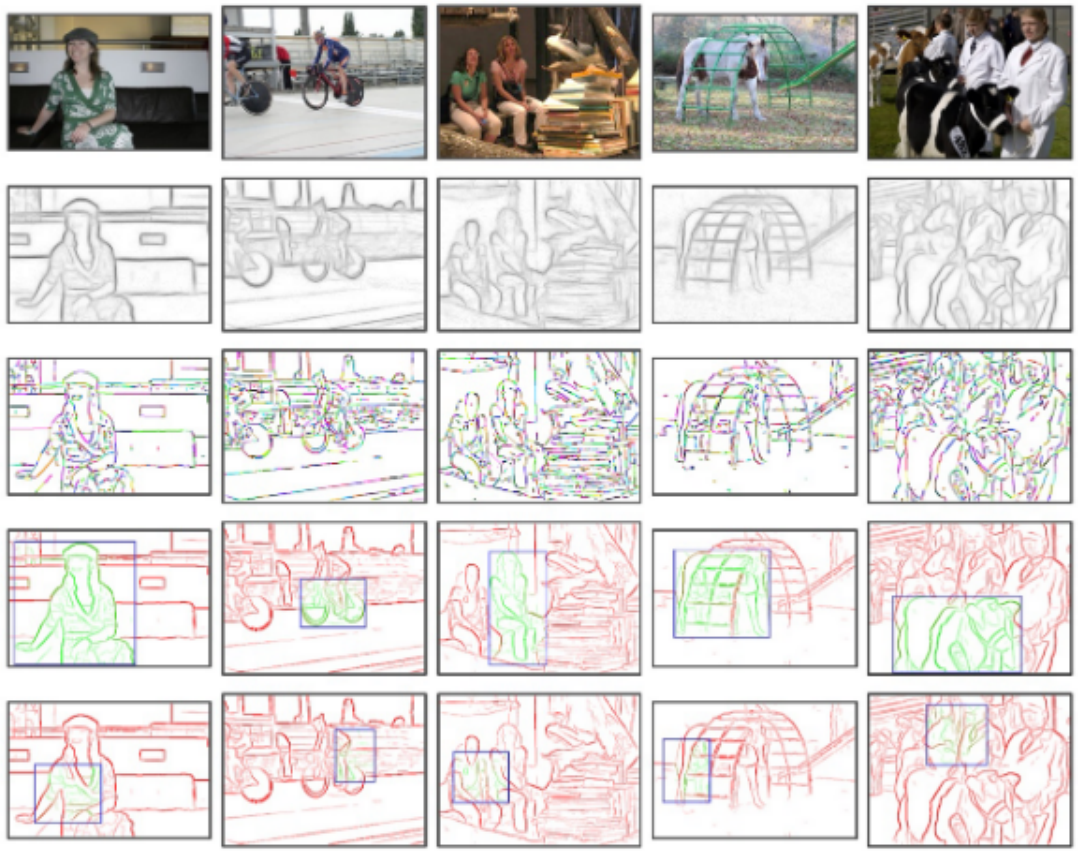


Fig. 1. Illustrative examples showing from top to bottom (first row) original image, (second row) Structured Edges [16], (third row) edge groups, (fourth row) example correct bounding box and edge labeling, and (fifth row) example incorrect boxes and edge labeling. Green edges are predicted to be part of the object in the box ($w_b(s_i) = 1$), while red edges are not ($w_b(s_i) = 0$). Scoring a candidate box based solely on the number of contours it *wholly encloses* creates a surprisingly effective object proposal measure. The edges in rows 3-5 are thresholded and widened to increase visibility.

Evaluation de EdgeBox

Fig. 6. Qualitative examples of our object proposals. Blue bounding boxes are the closest produced object proposals to each ground truth bounding box. Ground truth bounding boxes are shown in green and red, with green indicating an object was found and red indicating the object was not found. An IoU threshold of 0.7 was used to determine correctness for all examples. Results are shown for Edge Boxes 70 with 1,000 object proposals. At this setting our approach returns over 75% of object locations.

Définition d'une métrique entre deux images

Soient $\mathcal{S}_1$ et $\mathcal{S}_2$ les ensembles de boites obtenues dans deux images I_1 et I_2

- appariement : chaque boite de $\mathcal{S}_1$, est appariée avec la boite de $\mathcal{S}_2$ qui lui ressemble le plus au sens des descripteurs convNet. Similarité en cosinus entre descripteurs d_i^1 et d_j^2 : $\frac{d_i^1}{||d_i^1||} . \frac{d_j^2}{||d_j^2||}$(produit scalaire)
- Chaque couple de boites appariées est doté d'un score s_{ij} de ressemblance de boites favorisant les couples pour lesquelles les tailles de boites sont similaires
- Calcul de la similarité entre 2 images

$$\mathcal{S}_{1,2} = \frac{1}{n} \sum_{app\ i,j} 1 - d_{ij}.s_{ij}$$

où n : nombre de boites généréés par edgeBox (50 à 100).$d_{ij} = 1 - d_i.d_j$.

Résultats

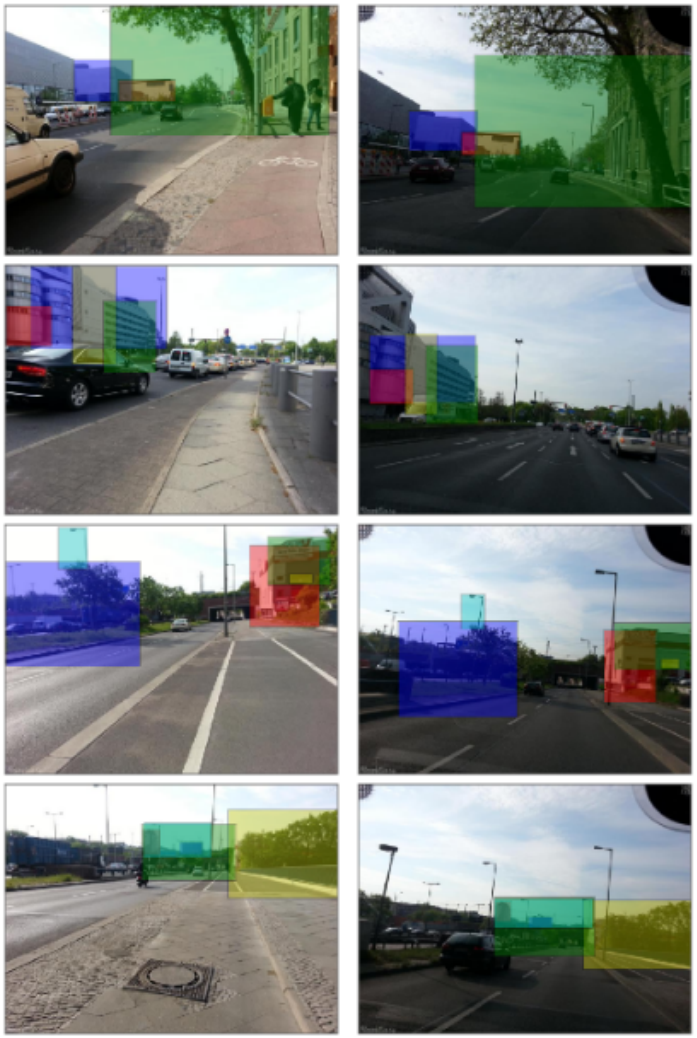


Fig. 6: The images in the *Berlin Halenseestraße* sequence have been recorded by a biker riding on the bike lane (left column) and a dashboard camera in the front of a car (right column). The changes in viewpoint are severe but our proposed method is able to extract landmarks and correctly match them between a large number of scenes

Résultats : courbes precision-recall

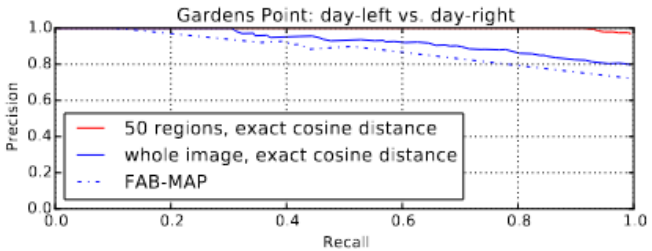


Fig. 3: Results for the Gardens Point Campus dataset. We clearly outperform the whole-image ConvNet-based method proposed by [41] and OpenFABMAP [14].

A match is considered a positive if it passes a ratio test (ratio of the distances of the best over the second best match found in the nearest neighbor search) , and a negative otherwise. Every scene in the datasets has a ground truth match. A match is a true positive when it is within 1..5 frames of the ground truth (depending on the frame rate of the recorded dataset) and a false positive otherwise. The running parameter for creating the PR curves is the threshold on the ratio test.

Apprendre des métriques

- Avec les decripteurs convNet, on apprend un descripteur adapté au problème, mais la métrique reste la distance Euclidienne
- On peut aller plus loin et apprendre une métrique .
  - idée de base : donner des paires d'objets semblables (i.e. de la même classe) ou dissemblables (i.e. de classes différentes) pendant l'apprentissage
  - outil : les réseaux siamois
  - on verra cela un peu plus tard dans le cours

Conclusion

descripteur ou métrique fait main (handcrafted) ou appris ?

- les CNNs ont apporté des progrès indéniables en terme d'invariance aux conditions environnementales (lumière, point de vue,...)
- l'utilisation des CNNs dépend de la disponibilité de données d'apprentissage. Pour pallier au manque de données :
  - re-training d'un réseau existant avec de nouvelles données
  - utilisation de données synthétiques réalistes comme complément (augmentation de données)
- ces descripteurs ne supplantent pas forcément les descripteurs faits main. Voir par exemple : *Comparative Evaluation of Hand-Crafted and Learned Local Features* [SHSP17]
- il existe de nombreuses approaches hybrides mixant CNN et conventionnel

Bibliographie I

Sung-Hyuk Cha and Sargur N. Srihari.

On measuring the distance between histograms.

Pattern Recognition, 35(6) :1355 – 1370, 2002.

Navneet Dalal and Bill Triggs.

Histograms of oriented gradients for human detection.

In *In CVPR*, pages 886–893, 2005.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton.

Imagenet classification with deep convolutional neural networks.

In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, pages 1097–1105, USA, 2012. Curran Associates Inc.

Yann Lecun, Léon Bottou, Yoshua Bengio, and Patrick Haffner.

Gradient-based learning applied to document recognition.

In *Proceedings of the IEEE*, pages 2278–2324, 1998.

Bibliographie II

G. Mori, S. Belongie, and J. Malik.

Shape contexts enable efficient retrieval of similar shapes.

In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, 2001.

Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas.

The earth mover's distance as a metric for image retrieval.

Int. J. Comput. Vision, 40(2) :99–121, November 2000.

Johannes L. Schönberger, Hans Hardmeier, Torsten Sattler, and Marc Pollefeys.

Comparative evaluation of hand-crafted and learned local features.

In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6959–6968, 2017.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich.

Going deeper with convolutions.

CoRR, abs/1409.4842, 2014.

Bibliographie III

Niko Sünderhauf, Sareh Shirazi, Adam Jacobson, Feras Dayoub, Edward Pepperell, Ben Upcroft, and Michael Milford.

Place recognition with convnet landmarks : Viewpoint-robust, condition-robust, training-free.

In *Robotics : Science and Systems*, 2015.

Sivic and Zisserman.

Video google : a text retrieval approach to object matching in videos.

In *Proceedings Ninth IEEE International Conference on Computer Vision*, pages 1470–1477 vol.2, Oct 2003.

C. Lawrence Zitnick and Piotr Dollár.

Edge boxes : Locating object proposals from edges.

In *ECCV*, 2014.