# LEC01 : Extraction

lundi 29 juin 2020 16:34

---

Introduction
- Contexte
- Bases de Données Multimédia
- Domaines d'application
- Indexation
- Outil OpenCV

---

**Contexte :**

Indexer : extraire une information synthétique des images afin de faciliter l'accès à leur contenu.
Le sujet de ce cours est la recherche automatique de documents visuels dans des bases de données de grande taille, à partir de requêtes relatives au contenu de ces documents.

**Types d'indexation :** Le type de requête donne le type d'indexation, indexation manuelle ou automatique?

> Recherche par contenu :

Qu'est-ce que ces images ont exactement en commun?

> Recherche par texte :

Récupérer toutes les images montrant : "tournesol"

# Indexation manuelle :

- Annotation
  * Approche classique consiste à indiquer des mots-clés attachés au document décrivant, dans un vocabulaire restreint, les caractéristiques principales et bien identifiables des documents stockés.
  * Par quoi indexer : mots-clés, métadonnées
- Inconvénients
  * Même image peut avoir plusieurs annotations différentes
  * Ambiguïté de l'annotation
  * Dépendance du contexte
  * Le coût d'annotation manuel est très important (10 fois la durée de document)
  * Approche la plus ancienne et la plus répandue

# Indexation automatique :

- Indexation par contenu
  * L'algorithme d'indexation attache des données de bas niveau sémantique, relatifs aux contenus géométrique, spectral, de l'image, à un niveau local ou global.
- Recherche par contenu (CBIR : Content-Based information Retrieval)
  * Les requête se font en général par l'exemple, ou par modèle.
  * Extraire automatiquement d'une image des descripteurs significatifs et compacts, qui seront utilisés pour la recherche ou la structuration.
- BDMM (Base de Données Multimédia)
  * Offline : production d'indexes issus de l'analyse du contenu des images (extraction de caractéristiques pertinentes, organisation…)
    BDMM -> Extraction Signatures -> Base de signatures <-> Index
  * Online : gestion des requêtes des utilisateurs
    Document multimédia requête -> Extraction de signatures -> Recherche par similarité <-> Index

## Définitions : Indice, Descripteur, Signature

* Indices visuels : caractéristiques de l'image, au sens de perception humaine, que l'on cherche à utiliser pour la tâche considérée
  . Principaux indices visuels : couleur, forme, texture, régions, mouvement
* Descripteur d'image : méthode d'extraction du contenu visuel de l'image
  . Exemple : histogramme couleur
* Signature d'image (caractéristiques) : vecteur numérique représentant le contenu visuel de l'image
  . Exemple : 1 vecteur de dimension 216 pour l'histogramme couleur
* Espace de description (de représentation) des images
  . 1 image = 1 ou plusieurs points dans un espace multidimensionnel
* Espace de recherche dans la base d'images
  . Structuration de l'espace de description pour une recherche efficace (index)

## Mesure de similarité :

* Distance

. En mathématiques, on appelle distance sur un ensemble E une application d définie sur le produit $E^2 = E \times E$ et à valeurs dans l'ensemble $R^+$ des réels positifs :

$$d : E \times E \to R^+$$

. Vérifiant les propriétés suivantes :

Symétrie : $\forall (a,b) \in E^2,\ d(a,b) = d(b,a)$

Séparation : $\forall (a,b) \in E^2,\ d(a,b) = 0 \Leftrightarrow a = b$

Inégalité triangulaire : $\forall (a,b,c) \in E^3,\ d(a,c) \leq d(a,b) + d(b,c)$

. Un ensemble muni d'une distance et un espace métrique

. Dans R^n, on peut définir de plusieurs manières la distance entre deux points

. Soient deux points de $E$, $(x_1, x_2, \dots, x_n)$ et $(y_1, y_2, \dots, y_n)$, on exprime les différentes distances ainsi :

distance de Manhattan (taxicab distance) : 1-distance

$$\sum_{i=1}^{n} |x_i - y_i|$$

distance euclidienne (distance vol oiseau) : 2-distance

$$\sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

distance de Minkowski (distance général) : p-distance

$$\sqrt[p]{\sum_{i=1}^{n} |x_i - y_i|^p}$$

distance de Tchebychev : inf-distance

$$\lim_{p \to \infty} \sqrt[p]{\sum_{i=1}^{n} |x_i - y_i|^p} = \max_{1 \leq i \leq n} |x_i - y_i|$$

## OpenCV :

### Introduction :

* Bibliothèque de traitement d'images et vision par ordinateur.
* OpenCV regroupe plusieurs modules :

# Core :

contient les fonctionnalités de base, accès aux pixels, changement de luminosité, de contraste, le changement d'espace couleur, la possiblité de dessiner sur les images, etc.

# Imgproc :

Permet d'appliquer différents filtres (moyen, gaussien, médian)

Permet d'appliquer différentes opérations morphologiques (dilatation,

érosion, ouverture, fermeture, gradient morphologique etc)
Contient différents algorithmes de seuillage et de détection de contours (Sobel, opérateur de Laplace, Canny Edge Detector), mais aussi des détecteur de droites ou d'ellipse (Hough Line/Cercle Transform)

# Highgui : permet l'ajout de composants graphiques de base, mais aussi avancés

# Calib3D : permet la calibration des caméras et la reconstruction 3D

# Feature2D : contient des descripteurs 2D souvent utilisés basés sur les couleurs, la forme, la texture, les points d'intérêt ... et des algorithmes de mises en correspondance.

# Video : contient les fonctionnalités de base de traitement de vidéos (algorithmes de détection de mouvement, de suivi, d'extraction de plan principal etc...)

# Objdetect : contient des algorithmes de détection d'objet, notamment la détection de visages

# ML : contient les algorithmes d'apprentissage et de classification

# GPU : contient les algorithmes permettant l'utilisation de la carte graphique afin d'accélérer le temps d'exécution grâce au parallélisme des GPU

Classes de base :

# Point_

Structure de données générique pour représenter des points dans l'espace de dimension 2

# Size_

Représente la taille d'un objet rectangulaire à 2 dimensions

# Vec

Représente un vecteur générique de faible dimension ($\leq 10$)

# MAT

La classe Mat permet de stocker l'image sous forme matricielle
Chaque objet Mat possède deux parties : L'entête, les données

# LEC02 : CBIR1

lundi 29 juin 2020 16:34

---



---

## Indexation d'image par contenu (CBIR)

### Principe :

**Hors-ligne :** Indexation

Calcul des signatures (indices) de description pour toutes les images de la base

**En-ligne :** Recherche

Calcul de signature pour l'image inconnue (image requête)
Mesure de similarité
Résultat : adresse des meilleures images au sens de la mesure de similarité

**À définir?**

- Soient $B = \{X_i = f(I_i), 1 \leq i \leq n\}$ et $X' = f(I')$ les signatures des images $I_i$ et de l'image requête $I'$ respectivement
- Mesure de similarité (exp : distance euclidienne) entre descripteurs $d(X_i, X')$
- Résultat : $R = \{I_i / d(X_i, X') < e\}$

Ce qui nécessite :

- Choix d'un espace de représentation et d'un **descripteur** (fonction)
- Calcul de **signature** de **l'image** dans cet espace
- Définition d'une **mesure de similarité (distance)** dans cet espace

### Image :

**Image numérique :** C'est une matrice de N*M pixels (picture element) correspondant à l'échantillonnage et la quantification d'un signal acquis avec un capteur

**Format d'image :**

Image à niveaux de gris (intensité ou luminance)

- Chaque pixel est codé sur N bits, ce qui lui confère des valeurs entières comprises entre 0 (noir) et $2^N - 1$ (blanc)

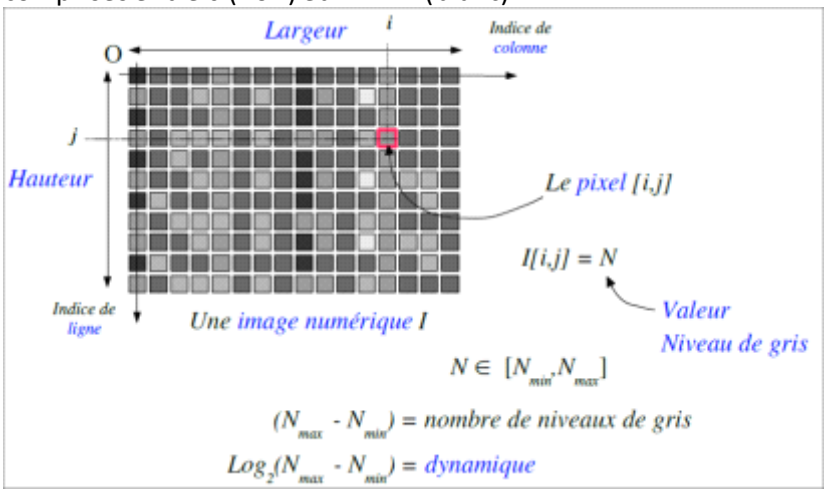



Image couleur

- Une image couleur correspond à la synthèse additive de 3 images, rouge, vert et bleu. Chaque pixel est donc codé sur 3xN bits
- 3 grilles de valeurs, 1 grille par composante de couleur
- RGB : 8 bits de quantification pour chaque couleur

- 24 bits par pixels (pixel = élément du support du signal)

## Descripteurs d'image

### Types :

#### Description globale de l'image : Description approximative de toute l'image

- Considère l'image dans son ensemble
- Caractérise l'image en utilisant des statistiques calculées sur l'image entière
- Une description moins fine de l'image notamment de recherche des objets

#### Description locale de l'image

Considère l'image comme composée d'un ensemble d'objets

Détection de points d'intérêt et calculs éventuels d'invariants autour de ces points d'intérêt

#### Description spécifique (essentiellement biométrie)

Empreintes digitales : Minuties

Visages : Eigenfaces

### Type de caractéristiques

#### Caractéristiques globaux

Couleur

Forme

Texture

#### Caractéristiques locales

Points d'intérêts

Régions d'intérêts

#### Caractéristiques spécifiques

Eigenfaces

Minuties

## Descripteurs de couleur

### Moments de couleur

- la moyenne, moment d'ordre un : $\mu \triangleq m_1 = \mathbb{E}(I)$
- la variance, moment centré d'ordre deux : $\mathrm{V}(I) \triangleq \mu_2 = \mathbb{E}[(I-\mu)^2]$
- $\cdots$, ainsi que sa racine carrée l'écart type : $\sigma \triangleq \sqrt{\mathrm{V}(I)} = \sqrt{\mu_2}$
- le coefficient d'asymétrie, moment centré réduit d'ordre trois :

$$\gamma_1 \triangleq \beta_1 = \mathbb{E}\left[\left(\frac{I-\mu}{\sigma}\right)^3\right]$$

- le kurtosis non normalisé, moment centré réduit d'ordre quatre :

$$\beta_2 = \mathbb{E}\left[\left(\frac{I-\mu}{\sigma}\right)^4\right]$$

#### Couleur moyenne

- La valeur moyenne (dans l'espace RGB): somme des valeurs RGB de tous les pixels, normaliser par le nombre de pixels

$$R_{avg} = \frac{1}{N \times M}\sum_{i=0}^{N}\sum_{j=0}^{M} R(i,j),\ G_{avg} = \frac{1}{N \times M}\sum_{i=0}^{N}\sum_{j=0}^{M} G(i,j)$$

$$B_{avg} = \frac{1}{N \times M}\sum_{i=0}^{N}\sum_{j=0}^{M} B(i,j)$$

- Comparaison de deux images $x$ et $y$ par la couleur moyenne en utilisant la distance euclidienne

$$d^2_{avg}(x,y) = (R_{avg}x - R_{avg}y)^2 + (G_{avg}x - G_{avg}y)^2 + (B_{avg}x - B_{avg}y)^2$$

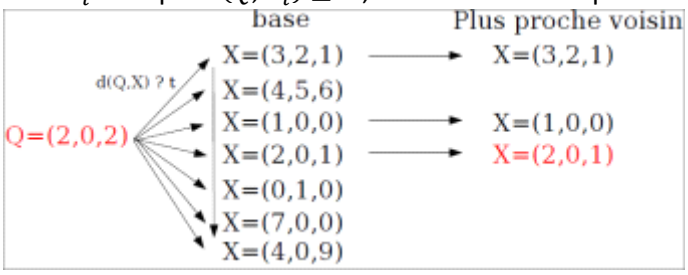
#### Mesure de similarité

Q signature (moment de couleur) d'une image requête $I_Q$

$B = \{X_i (1 < i < N)\}$ Base de signatures des images de la base

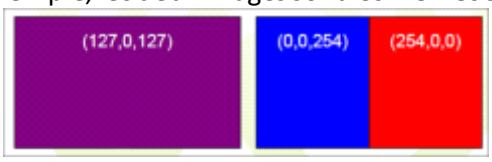
Quelle est l'image de la base la plus similaire à l'image requête?

Sont $X_i$ tels que $d(Q, X_i) \leq \varepsilon$ , avec ε déterminé par l'utilisateur




Limites :

- Mesure de similarité non précise
- Par exemple, les deux images sont les mêmes selon la moyenne de couleur :




Mais :

Rapide et facile à calculer et comparer

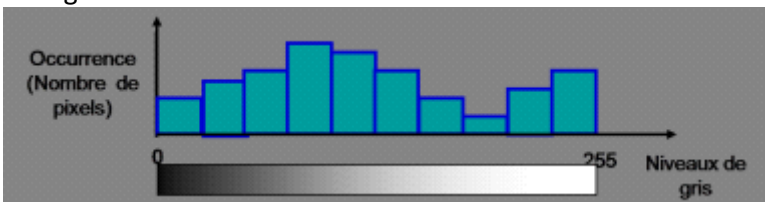
Meilleur pour l'utiliser comme un filtre : exclure images

Couleur dominante influence la moyenne de couleur

## Histogramme

L'histogramme est une fonction Hist(i) permettant de donner la fréquence d'apparition des différents niveaux de gris (couleur) qui composent l'image

- En abscisse on représente les niveaux de gris (couleur) et en ordonnée leurs fréquences d'apparition
- L'histogramme des niveaux de gris (couleur) nous informe sur la concentration de l'image




Pour une image couleur, il y a un histogramme par composante

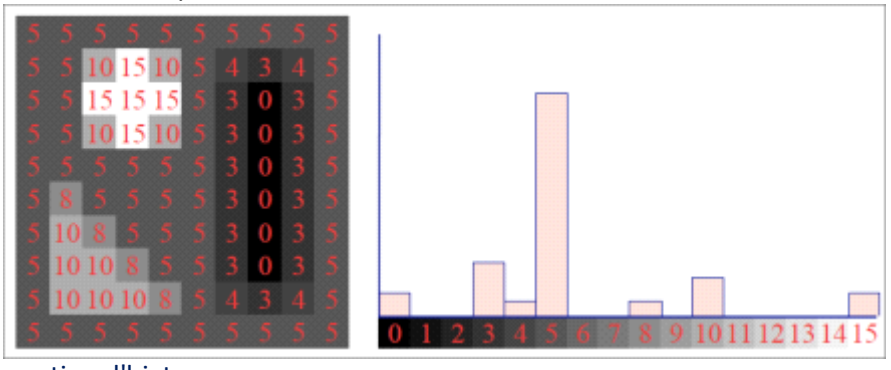
### Formule :

Fonction discrète de 0 à L-1

Hist(i) = n

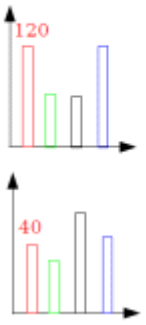
i : ton de gris

n : nombre de pixels de ton i




### L'intersection d'histogrammes

$$d(X,X') = 1 - \frac{\min(x_1, x'_1) + \min(x_2, x'_2) + \ldots + \min(x_D, x'_D)}{\min(x_1 + x_2 + \ldots + x_D, x'_1 + x'_2 + \ldots + x'_D)}$$




Min(120,40)=40 → Au moins 40 pixels rouges dans chacune des deux images

\+ Au moins 35 pixels verts dans chacune des deux images

\+ Au moins 34 pixels noirs dans chacune des deux images

\+ Au moins 50 pixels noirs dans chacune des deux images

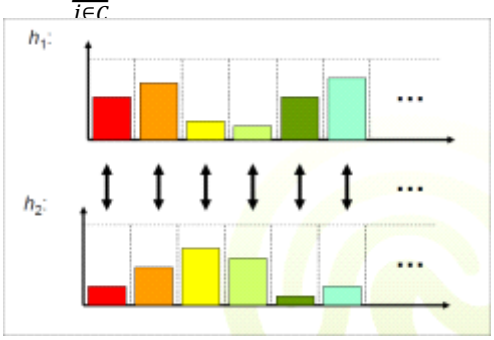
159 pixels ayant une couleur commune dans les deux images

### Distance de Minkowski

Soient h1 et h2 deux histogrammes

La distance de Minkowski avec le paramètre r :

$$d_r = \sum_{i \in C} |h_1(i) - h_2(i)|^r$$




La distance entre une image rouge et une image rouge vif est la même que entre une image rouge et image bleu

Limité dans le cas de changements de couleur parce que toutes les colonnes sont comparées individuellement

### Distance quadratique :

Évaluer la relation entre les différentes couleurs

Soit A une matrice qui exprime la similarité des paires entre la couleur i et la couleur j ($a_{ii} = 1$ et $a_{ij} = a_{ji}$)

$$d_A(h_1 - h_2) = (h_1 - h_2)^T * A * (h_1 - h_2)$$

$$= \sum_j \sum_i (h_1(i) - h_2(i))(h_1(j) - h_2(j))$$

### Autres distances : la distance de Mahalanobis, Chi2

### Problèmes des histogrammes

- Choix de la représentation de la couleur
- Distance entre couleurs

### Limites

Indice visuel insuffisant si utilisé seul, car insuffisamment discriminant : Pomme rouge vs Ferrari

### Avantages des histogrammes

Robuste à certaines transformations géométriques de l'image

**Remarque :** une image en noir et blanc suffit à un humain pour effectuer la tâche demandée

## Descripteurs de texture

### Méthodes de description de la texture

#### Méthodes statistiques

- **Matrice de cooccurrence (Mesures de Haralick)**
- **Histogramme**

#### Méthodes à base de modèle

- Décomposition de Wold
- Modèles Fractals
- Modèles AR (Autoregressive Models)

### Méthodes fréquentielles (traitement du signal)

**Fourier**

Gabor

Ondelettes

Il existe plusieurs méthodes pour analyser la texture, mais le plus difficile est de trouver une bonne représentation (paramètres) pour chaque texture

## Matrice de cooccurrence (Mesures de Haralick)

L'idée de cette méthode est d'identifier les répétitions de niveaux de gris selon une distance (pas) et une direction

Matrice de cooccurrence : matrice de taille $Ng \times Ng$

$Ng$ étant le nombre de niveaux de gris de l'image (256 * 256)

On réduit souvent à des tailles 8 * 8, 16 * 16 ou 32 * 32

Pour un voisinage $(dx, dy)$, la matrice de cooccurrence $M(dx, dy)$ est donnée par :

$$M[dx,dy](u,v) = \frac{1}{(N_x dx)(N_y dy)} \sum_{i,j} 1[I(i,j) = u \;\&\; I(i+dx, j+dy) = v]$$

$(Nx, Ny)$ : Taille de l'image

(u,v) : Niveaux de gris de l'image (valeur quantifiée)

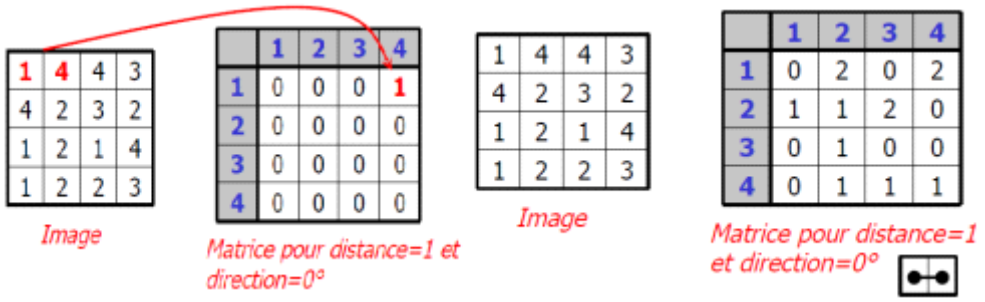
Plusieurs matrices, pour chaque distance (pas) et direction

Distance : 1,2,3 ...

Direction : 0,45,90,135 (en degré)

Temps de calcul de ces matrices est assez long

Exemple :




- On parcours l'image et pour chaque couple de pixels formé avec la distance et la direction données, on incrémente la matrice de cooccurrences de 1
- Le 2 de la matrice de cooccurrence (ligne 1 colonne 4) signifie que l'on trouve deux fois un pixel de valeur 1 de distance 1 et de direction 0 d'un pixel de valeur 4

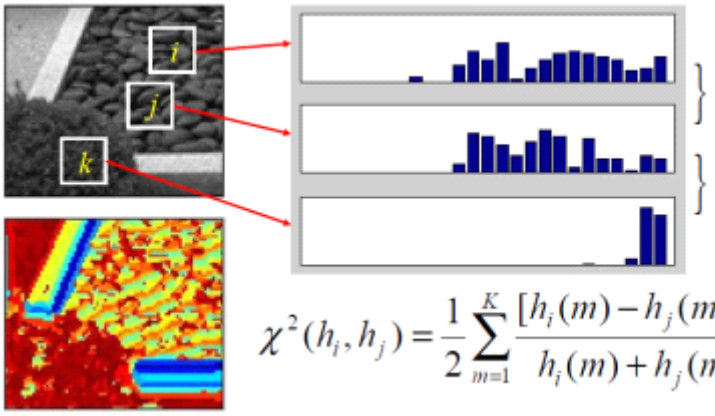
À partire de cette matrice de cooccurrence, il est possible de définir plusieurs descripteurs (Mesures de Haralick), tels que ceux répertoriées dans cette table :

| Opérateur | Formulation |
|---|---|
| Maximum | $\max_{ij}(C_{ij})$ |
| Différence d'ordre $k$ | $\sum_i \sum_j c_{ij} (i-j)^k$ |
| Entropie | $\sum_i \sum_j c_{ij} \, log(c_{ij})$ |
| Uniformité | $\sum_i \sum_j c_{ij}^2$ |

## Histogramme

### Statistiques du premier ordre (histogramme)

Distance du Chi2 entre histogrammes de textures

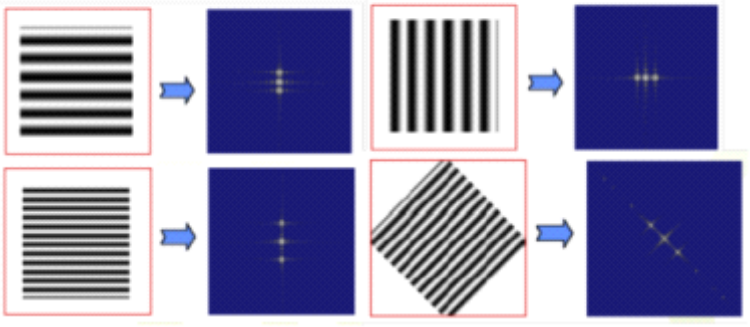



$$\chi^2(h_i,h_j)=\frac{1}{2}\sum_{m=1}^{K}\frac{[h_i(m)-h_j(m)]^2}{h_i(m)+h_j(m)}$$

## Transformée de Fourier discrète

En traitement d'images, on utilise la transformation de Fourier à deux dimensions, sa définition discrète est :

$$\mathrm{F}(u,v)=\sum_{m=0}^{\mathrm{M}-1}\sum_{n=0}^{\mathrm{N}-1} I(m,n)\cdot \mathrm{e}^{-2i\pi\left(\frac{um}{\mathrm{M}}+\frac{vn}{\mathrm{N}}\right)}$$



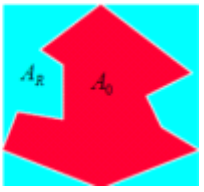
## Descripteurs de forme

Caractérisation des objects contenus dans l'image : représentation de composantes

- Caractéristiques internes : description de la région occupée par l'objet (surface, ...)
- Caractèristiques externes : description du contour de l'objet représenté par le périmètre, circularité, rectangularité, ...

Cette description est invariante en rotation et translation

| | |
|---|---|
| elongation | $\frac{W}{H}$ avec $H$ : hauteur, $W$ : largeur |
| circularité | $\frac{T^2}{4\pi A}$ avec $T$: périmetre, $A$ : surface |
| rectangularité | $\frac{A0}{Ar}$ avec $A0$ : surface de la région, $Ar$ : surface de rectangle |




Décrire les formes nécessite une identification préalable de régions

Segmentation de l'image

Détection de leurs contours

## Slides 36-52 : Code

# LEC03 : CBIR2

lundi 29 juin 2020 16:34

Descripteurs locaux et spécifiques
-> Descripteurs locaux
-> Descripteurs spécifiques

Annexe
-> Filtrage, Gradient

_____

## Descripteurs locaux

### Objectifs

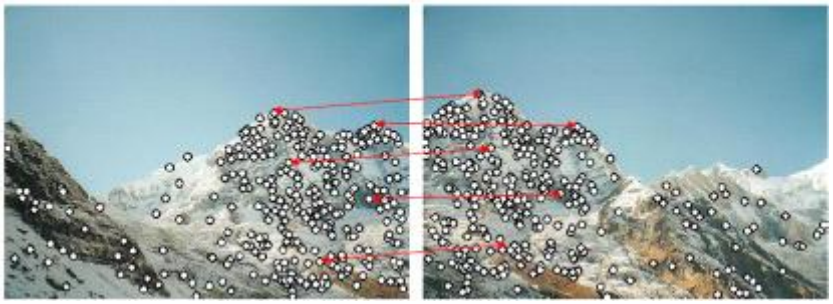
Recherche de zones ou d'objets similaires
Recherche de parties d'image
Requêtes partielles

### Points d'intérêt, Régions d'intérêt

Solutions :

Description à base de régions
Description à base de points

Points d'intérêt



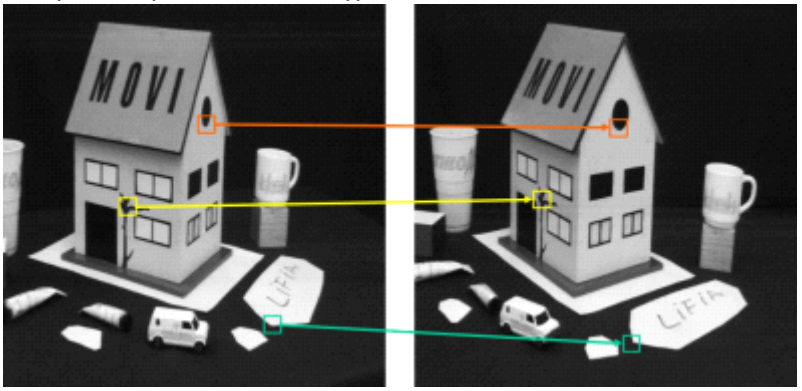
points particuliers des contours, coins (corners)
Coins (corners) : sont les points de l'image où le contour change brutalement de direction, comme par exemple aux quatre sommets d'un rectangle
Un point d'intérêt = site informatif de l'image

Régions d'intérêt

Il s'agit de zones d'intérêt plus générales que les points.

### Points d'intérêt : Difficultés

Dans l'exemple ci-dessous, la scène est soumise `a des déformations géométriques complexes qui invalident l'hypothèse de translation Locale




Il est primordial de disposer de descripteurs qui soit le plus invariants possible aux transformations géométriques : rotation, homothétie, transformation affine

### Points d'intérêt : Méthodes

Détecteur SIFT (Scale-invariant feature transform) : transformation de caractéristiques

visuelles invariante à l'échelle

- Détecteur SURF (Speeded Up Robust Features) : (on peut traduire par) caractéristiques robustes accélérées
- Méthode est basée sur l'analyse des : DoG (Difference of Gaussians), LoG (Laplacian of Gaussian) ou DoH (Difference of Hessians)

Détecteur de Harris

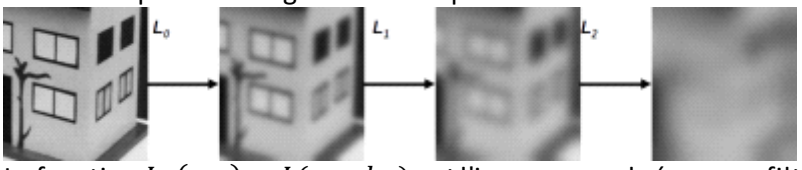
## Points d'intérêt : SIFT, SURF

### Étapes :

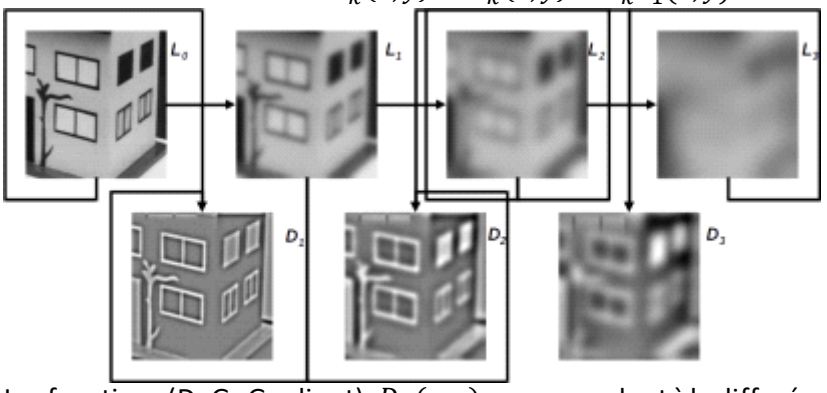
Extraction automatique de points d'intérêt dans chaque image
Description locale de chaque point d'intérêt
Mise en place d'une mesure de similarité
Indexation des descripteurs des images de référence et recherche des correspondances avec les descripteurs de l'image question

### Détection d'extrema dans l'espace des échelles

La détection s'effectue dans un espace discret que l'on appelle espace des échelles (scale space) qui comporte trois dimensions (x, y et σ)
L(Gaussienne de facteur d'échelle σ) : $L(x, y, \sigma) = G(\sigma) * I(x, y)$ est l'image convoluée par un filtre gaussien G de paramètre σ



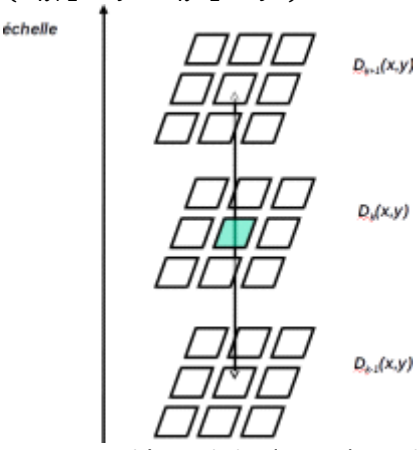

La fonction $L_k(x, y) = L(x, y, k\sigma)$ est l'image convoluée par un filtre gaussien G de paramètre kσ
Par conséquent, la détection des objets de dimension approximativement égale à σ se fait en étudiant l'image appelée différence de gaussiennes (DoG)
DoG définie comme suite : $D_k(x, y) = L_k(x, y) - L_{k-1}(x, y)$




Les fonctions (DoG : Gradient), $D_k(x, y)$ correspondent à la différénce entre 2 gaussiennes adjacentes ($L_k$ $et$ $L_{k+1}$)

### Point-clé candidat (x,y,σ)

Les points sélectionnés par SIFT sont les maxima et les minima locaux de la fonction $D_k(x, y)$, à la fois dans l'échelle courante et dans les échelles adjacentes $\left(D_{k+1}(x, y), D_{k-1}(x, y)\right)$



Un point-clé candidat (x, y, σ) est défini comme un point où un extremum du DoG est atteint par rapport `a ses voisins immédiats, c'est-`a-dire sur l'ensemble contenant 26 autres points défini par :

$\{D(x+\delta_x, y+\delta_y, s\sigma), \delta_x \in \{-1,0,1\}, \delta_y \in \{-1,0,1\}, s \in \{k-1, k, k+1\}\}$

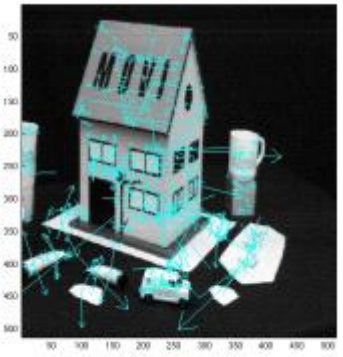
Dans l'image (DoG) ne persistent plus que les objets observables dans des facteurs d'échelle qui varient entre σ et kσ

## Point d'intérêt SIFT (x,y,σ)

### Assignation d'orientation

Pour chaque extrema de l'espace d'échelle des différences de gaussiennes (point d'intérêt SIFT), on calcule la direction associée par :

$$\theta(x,y) = \arctan\left(\frac{D_y^\sigma(x,y)}{D_x^\sigma(x,y)}\right)$$

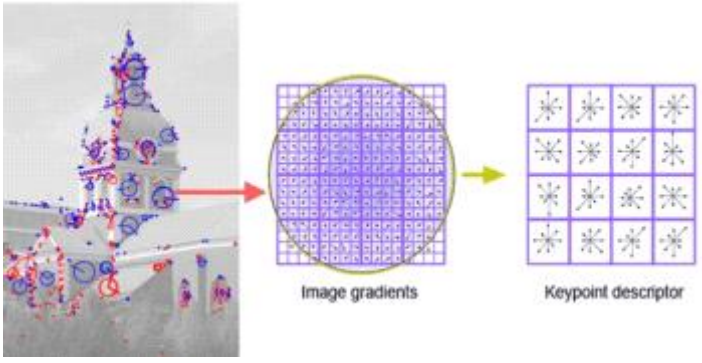


À l'issue de cette étape, un point-clé est donc défini par quatre paramètres (x,y,σ,θ)

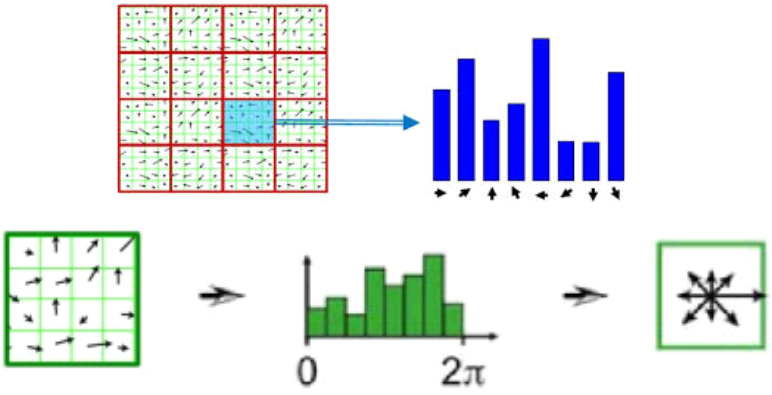
## Descripteur de point d'intérêt SIFT

Les descripteurs associés aux points d'intérêt SIFT sont des histogrammes des orientations locales autour du point d'intérêt

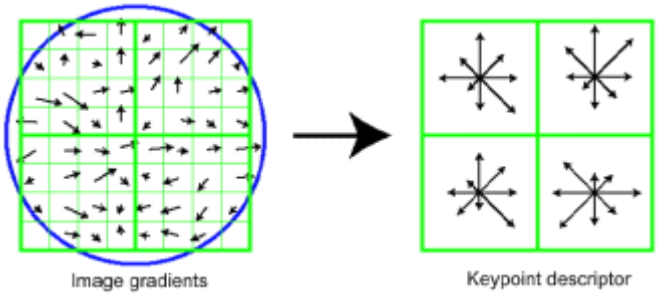
On divise l'espace autour de chaque point d'intérêt (x,y) en $N^2$ carrés 4x4 pixels

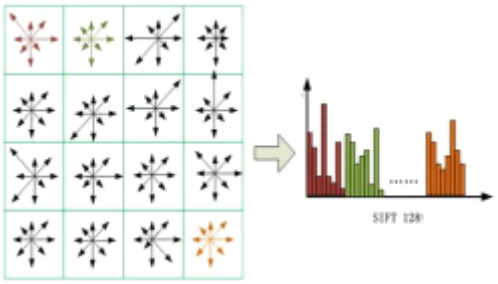



Pour chaque carré 4x4, on calcule un histogramme des orientations quantifiées en 8 directions (8 intervalles), en pondérant par : (1) le module du gradient (2) l'inverse de la distance au point d'intérêt (x,y)

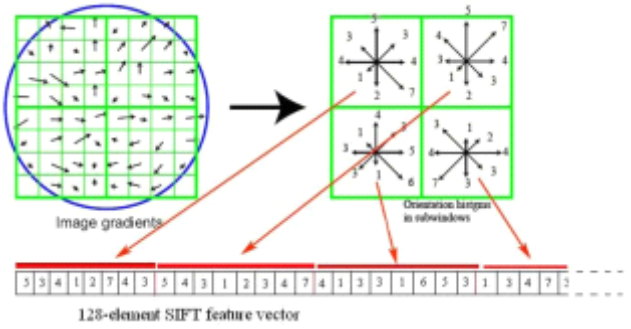



On utilise le gradient calculé $(D_x^\sigma(x,y), D_y^\sigma(x,y))$ pour les 4x4x$N^2$ points (a,b)

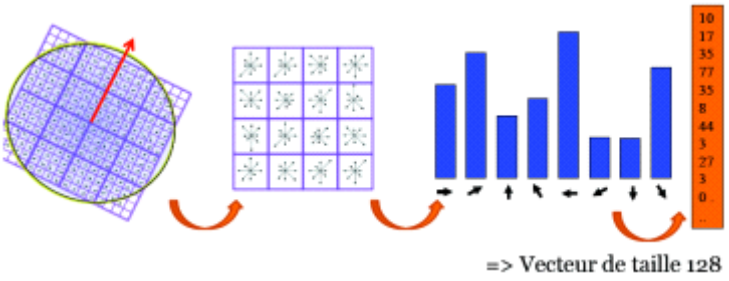


Les descripteurs formés sont donc des vecteurs de taille $8xN^2$

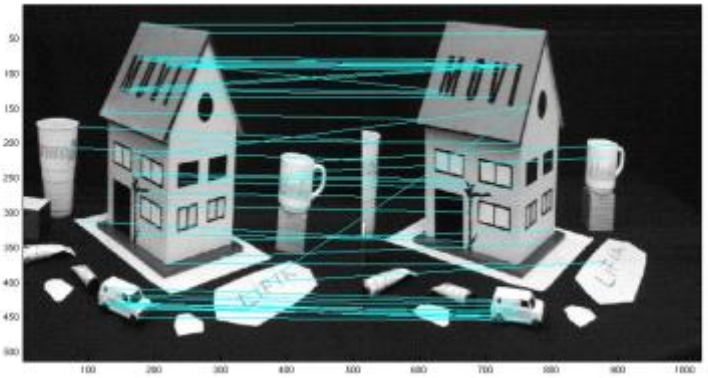



Calcul de plusieurs histogrammes des orientations, relatives à l'orientation dominante : 8bins x 4 x 4 histogrammes = vecteur de taille 128




### Appariement par SIFT

Mise en correspondance deux à deux des points entre chaque paire d'images
Les descripteurs des points formés (vecteurs de taille 128) seront appariés en utilisant une distance (e.g. distance euclidienne)
Mesure de similarité entre 2 images = le nombre de points mis en correspondance



## Points d'intérêts : Harris

### Calcul de coins de Harris

Calcul des images des gradients $I_x$ et $I_y$ de l'image (opérateurs de Sobel, Roberts, Prewitt, ...)
Filtrer ces images résultantes par un filtre Gaussien de taille à définir (3x3, 9x9 ou 12x12 ...)

Calcul des gradients $I_x^2$ et $I_y^2$ de l'image (en utilsant pour l'image 2 fois de suite l'opérateur dérivé), puis filtrer par une gaussienne

Calcul des coins de Harris pour chaque pixel de l'image de la façon suivante

Calcul de la matrice de Harris

$$\begin{pmatrix} I_x^2 & I_x I_y \\ I_x I_y & I_y^2 \end{pmatrix}$$

$I_x$ est le gradient suivant l'axe X et $I_y$ est le gradient suivant l'axe Y

$I_x^2$ et $I_y^2$ sont les images convoluées 2 fois avec le gradient X et Y

Calcul de la trace et du déterminant de la matrice (sans calculer les valeurs propres) :

$trace(M) = \lambda_1 + \lambda_2 = M_{1,1} + M_{2,2}$

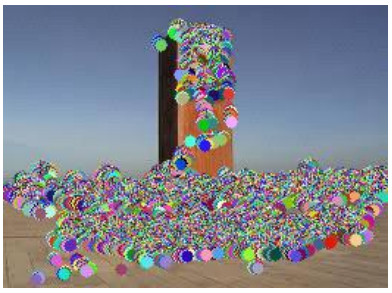
$\det(M) = \lambda_1 \times \lambda_2 = M_{1,1}M_{2,2} - M_{1,2}M_{2,1}$

Calcul de la réponse R du détecteur :

$R = \det(M) - k\big(trace(M)\big)^2$

k est un paramètre à régler, typiquement k=0.04

Extraction des maxima locaux positifs dans un voisinage 3x3 de la réponse R (c'est-à-dire mettre à zéro tous les points négatifs ou dont la valeur n'est pas supérieure à celle des 8 voisins)



Mise en correspondance des points d'intérêts

Extraction de n meilleurs points de Harris (par tri par insertion dans un tableau de taille n), n étant un paramètre à configurer, avec par exemple n = 50

Pour chaque coin trouvé de l'image gauche (n coins au total) :

Calculer la similarité (ZSSD) avec chacun des n coins trouvés dans l'image droite (en utilisant une fenêtre K * K autour du pixel)

Conserver le meilleur point (plus forte similarité) et faire une correspondance

Construire un tableau contenant les correspondances trouvées (coordonnées image gauche, coordonnées image droite, similarités)

## Régions d'intérêt

Descripteur fin de chaque région par un histogramme adaptif

## Détection de visage

Cherche à détecter la présence et la localisation précise d'un ou plusieurs visages dans

une image numérique
C'est un sujet difficile, notamment dû à la grande variabilité d'apparence des visages
La détection de visage est la première étape vers des applications plus évoluées, qui nécessitent la localisation du visage
- La reconnaissance de visage
- La reconnaissance d'expression faciales
- L'évaluation de l'âge ou du sexe d'une personne
- Le suivi de visage
- Ou l'estimation de la direction du regard et de l'attention

Algorithme largement utilisé pour détection de visage : Méthode de Viola et Jones

### OpenCV : SIFT, SURF

#### Étapes :

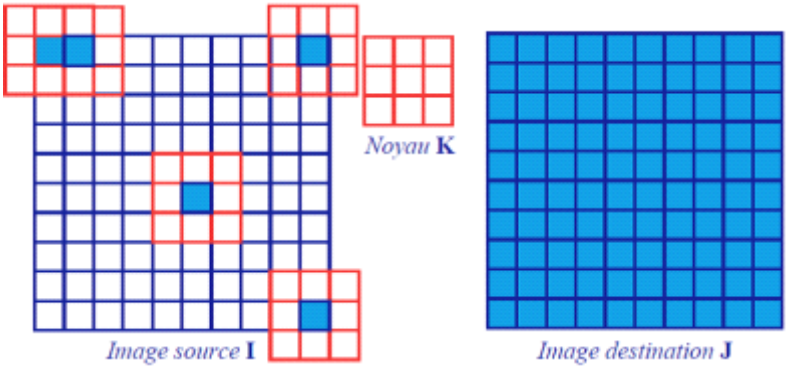
- Extraction automatique de points d'intérêt dans chaque image
  - Détection d'extrema dans l'espace des échelles
  - Extraction des points-clés (des images de référence et de l'image question)
- Description locale de chaque point d'intérêt
  - Exemple : Un point peut être décrit par 8 invariants à la translation et à la rotation
  - Calcul des descripteurs (des images de référence et de l'image question)
- Mise en place d'une mesure de similarité
  - Entre signatures de points d'intérêt
  - Entre ensembles de points
- Indexation des descripteurs des images de référence et recherche des correspondances avec les descripteurs de l'images question

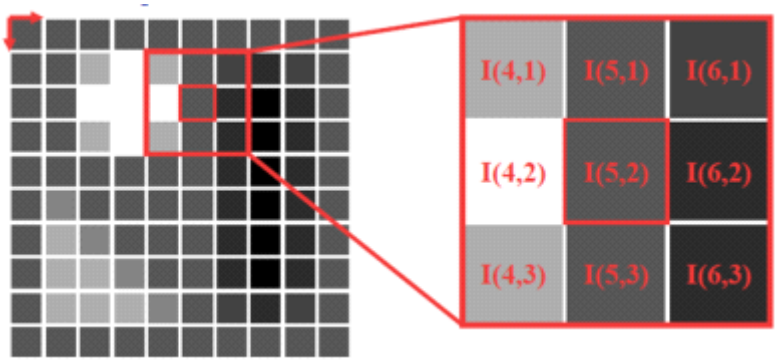
Code : Slides 24-30

## Annexe :

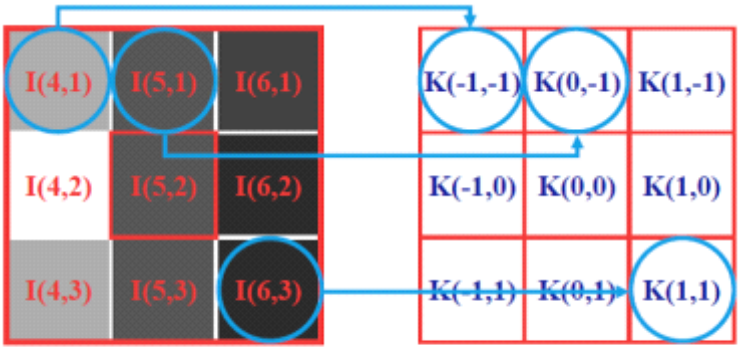
### Filtrage : Éliminer le bruit

Éliminer tous les pixels aberrants sans modifier les autres pixels corrects
Éliminer le bruit numérique est une tâche difficile dans le sens où il faudrait dans l'absolu
En réalité, nous devons chercher à diminuer le bruit numérique
Les techniques classiques consistent à appliquer les filtres ou des transformées permettant de flouter l'image
En pratique, le filtrage d'image (convolution) se fera par application de masque sur l'image



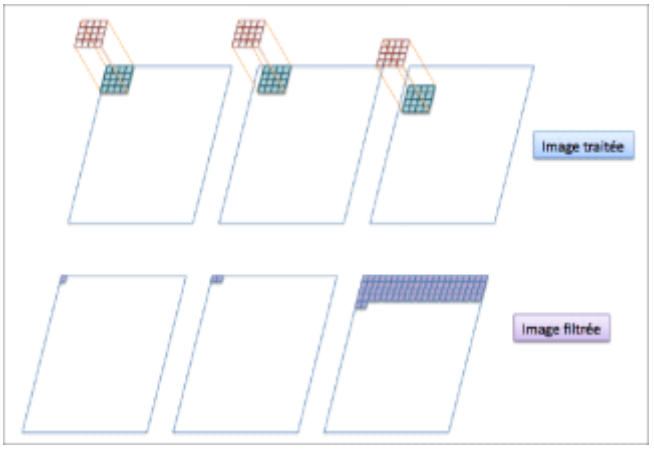



Le masque d'une image se fera par une sommation de multiplications




$$\begin{aligned}\mathbf{J}(5,2) &= \mathbf{I}(4,1)\mathbf{K}(-1,-1) + \mathbf{I}(5,1)\mathbf{K}(0,-1) + \mathbf{I}(6,1)\mathbf{K}(1,-1)\\ &+ \mathbf{I}(4,2)\mathbf{K}(-1,0) + \mathbf{I}(5,2)\mathbf{K}(0,0) + \mathbf{I}(6,2)\mathbf{K}(1,0)\\ &+ \mathbf{I}(4,3)\mathbf{K}(-1,1) + \mathbf{I}(5,3)\mathbf{K}(0,1) + \mathbf{I}(6,3)\mathbf{K}(1,1)\end{aligned}$$

Principe de la fenêtre glissante




Filtre Gaussien : c'est un filtre défini par G :

$$G(x,y,\sigma) = \frac{1}{2\pi\sigma^2} e^{-\left(\frac{x^2+y^2}{2\sigma^2}\right)}$$

Si par exemple σ = 0.8, on a le filtre 3 * 3 suivant :

$$\begin{array}{|c|c|c|}\hline G(-1,-1) & G(0,-1) & G(1,-1)\\ \hline G(-1,0) & G(0,0) & G(1,0)\\ \hline G(-1,1) & G(0,1) & G(1,1)\\ \hline\end{array} \simeq \frac{1}{16} \cdot \begin{array}{|c|c|c|}\hline 1 & 2 & 1\\ \hline 2 & 4 & 2\\ \hline 1 & 2 & 1\\ \hline\end{array}$$

Filtre moyenneur : filtre dont tous les coefficients sont égaux (normalisé par la somme des coefficients)

$$\frac{1}{9} \cdot \begin{array}{|c|c|c|}\hline 1 & 1 & 1\\ \hline 1 & 1 & 1\\ \hline 1 & 1 & 1\\ \hline\end{array}$$

## Gradient

L'image est une fonction 2D

$$I: (x,y) \rightarrow I(x,y)$$

La dérivée première (gradient) de l'image est l'opérateur de base pour détecter les

contours dans l'image

$$\vec{D}(D_x, D_y) = \left(\frac{\partial I(x,y)}{\partial x}, \frac{\partial I(x,y)}{\partial y}\right)$$

Gradient de l'image (dérivée en X + dérivée en Y) : vecteur avec une norme et une direction

Norme : Intensité du gradient en chaque pixel $D = \sqrt{D_x^2 + D_y^2}$

Direction : $\theta = \arctan\left(\frac{D_y}{D_x}\right)$

Pour une image numérique, on cherche donc à approximer les dérivées par différence finies

$$D_x(x,y) = I(x,y) - I(x-n,y)$$

Ou

$$D_x(x,y) = I(x+n,y) - I(x-n,y)$$

Avec en général n=1

Ces dérivées sont calculées par convolution de l'image avec un masque de différences [-1,0,1]

Le calcul des dérivées directionnelles en x et y revient finalement à la convolution avec les noyaux suivants

Sobel :

$$h_1 = \begin{bmatrix} -1 & 0 & 1 \\ -1 & 0 & 1 \\ -1 & 0 & 1 \end{bmatrix}$$

$$h_2 = \begin{bmatrix} -1 & -1 & -1 \\ 0 & 0 & 0 \\ 1 & 1 & 1 \end{bmatrix}$$

Priwitt :

$$h_1 = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}$$

$$h_2 = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix}$$

# LEC04 : CBVR

lundi 29 juin 2020 16:34

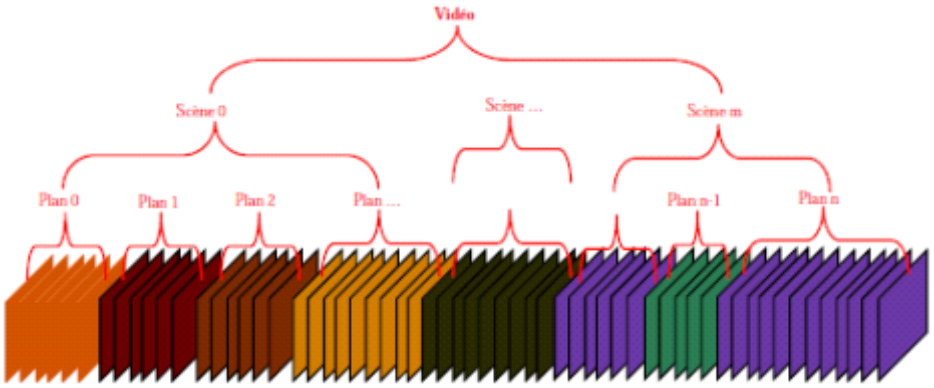
Indexation vidéo par contenu
- CBVR : Introduction
- Analyse de la vidéo
- Segmentation de la vidéo
- Création de résumé de la vidéo

## Qu'est-ce qu'un document vidéo?

### Un document vidéo est constitué

- D'un signal audio : une ou plusieurs bandes son
- D'un signal vidéo qui est constitué d'un nombre important de frames défilées de plus de 20 fps (frames par sec)
- Vidéo = suite d'images + son + texte

### Le modèle cinématographique du signal vidéo




- Les frames (frame) : image de la vidéo représentant le contenu visuel du plan
- Un plan
  - Il est défini par une séquence de frames durant laquelle l'acquisition du signal n'a pas subi d'interruption
- Une scène
  - Elle regroupe par définition l'ensemble des plans consécutifs se situant dans le même espace et au même instant, ou ayant un lien sémantique étroit
  - Elle représente la plus petite unité sémantique d'un film
  - Exemple (un dialogue, une conversation téléphonique, une action observée de plusieurs points de vue, deux évènements se produisant en parallèle, etc)
- Une vidéo peut être décomposée en plusieurs scènes, et chaque scène peut être composée d'un ou de plusieurs plans, et un plan de plusieurs frames

### Représentation d'une vidéo numérique

Le flux de données dans la vidéo numérique est alors défini par :

$$F = Nb * f_y * f_x * f_t \text{ (bits/s)}$$

On caractérise donc la vidéo numérique par les paramètres suivants :

$f_t$ : Fréquence d'image (FPS)
$f_y$ : Lignes par frame
$f_x$ : Pixels par lignes
$Nb$ : Bits à encoder, ce paramètre encode la valeur d'intensité (couleur) d'un pixel

Une vidéo numérique peut être obtenue par la discrétisation et l'encodage de la vidéo analogique ou en utilisant directement une série d'acquisition numérique (capteur numérique)

### Requêtes

#### Requêtes sur :

un ensemble de vidéo
une seule vidéo

#### Requêtes sous quelle forme :

Usage des méta données

Description textuelle d'une scène
Une image d'un film
L'image d'un acteur du film
Requête vidéo

## Indexation de la vidéo par métadonnées

### Métadonnées

Auteurs, réalisateurs, producteurs, acteurs, autres professionnels associés
Date de création
Informations sur les conditions de production

### Données textuelles

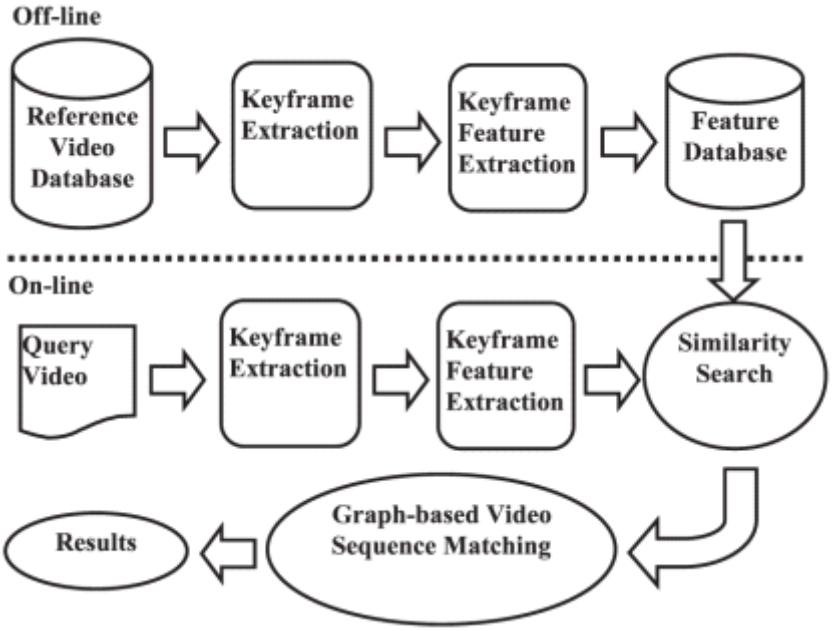
### Flux de données associé = le son

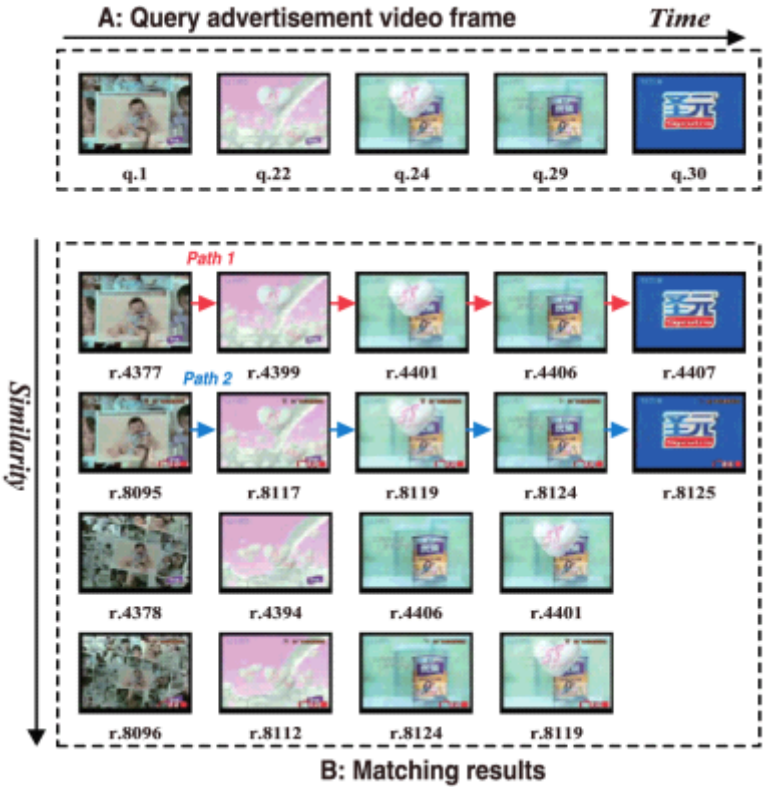
## Indexation de la vidéo par contenu

### Requête image

### Requête vidéo

## Schéma de CBVR




## Analyse de la vidéo numérique

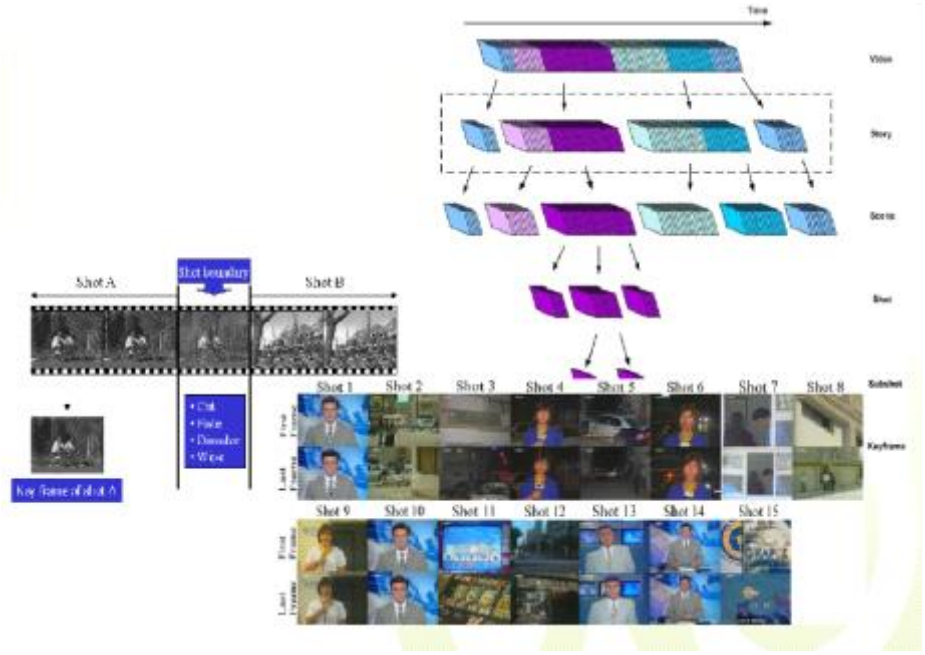
### Segmentation en plan et en scène

La segmentation en plans est l'identification automatique, par des méthodes informatiques, des bornes des plans dans une vidéo

Cela consiste à repérer automatiquement les points de montage définis à l'origine par le réalisateur, en mesurant les discontinuités entre les images successives de la vidéo

Permettra ensuite la sélection d'images-clés

Frame clé (key frame) : Image de la vidéo représentant grossièrement le contenu visuel du plan




## Structure d'un vidéo - transitions entre plans

Comment détecter les transitions entre plans, i.e. discontinuités dans le flux vidéo?

Comment détecter les discontinuités dans le flux vidéo?

- Transition brusque : cut
- Transitions progressives entre plans

- La détection de rupture de plan est basée sur la segmentation (partitionnement) de la séquence vidéo en une succession de frames représentant une action spatiale et temporelle continue
- Les images au voisinage d'une transitions sont très différentes
  - On cherche à repérer les discontinuité dans le flux vidéo
- De façon générale, on effectue les étapes suivantes :
  - Extraire une signature (primitive globale) sur chaque image
  - On définit ensuite une distance (ou mesure de similarité)^entre les signatures
  - Entre frame successifs, l'application de cette distance, sur l'ensemble de flux vidéo, produit un signal unidimensionnel
  - On cherche alors les variations extrêmes (pics du signal) qui correspondent aux instants de faible similarité

## Intensité

Imaginons la technique la plus simple :

Signature : intensités des images

Mesure de similarité : distance moindres carrés

$$D(t) = \sum_x \sum_y \big(I(x,y,t) - I(x,y,t-1)\big)^2$$

La technique est très sensible aux :

- Mouvements d'objets ou de caméra, bruit, ombrage, changement d'illumination

Il faut donc trouver des nouvelles primitives ou mesures plus robustes

## Histogramme

- Un plan contient grossièrement les mêmes objets en mouvement sur un fond statique
- La distribution de l'intensité à l'intérieur des frames d'un même plan est donc semblable
- On représente cette distribution par l'histogramme d'intensité des frames

Mesure de similarité :

Comment calculer une mesure de similarité avec des histogrammes?
On calcul une distance entre les histogrammes normalisés de chaque deux frames successifs de la vidéo et on pose un seuil pour l'une des distance suivantes :

- Distance euclidienne

$$D_e(H_t, H_{t-1}) = \sum_{u=0}^{m} (H_t(u) - H_{t-1}(u))^2$$

- Intersection d'histogramme

$$D_i(H_t, H_{t-1}) = 1 - \frac{\sum_{u=0}^{m} \min\left(H_t(u) - H_{t-1}(u)\right)}{Min(\sum H_t(u), \sum H_{t-1}(u))}$$

Avantage de l'histogramme d'intensité :

- Insensible au mouvement des objets : représentation qui ignore les arrangements spatiaux du contenu des frames
- Insensible aux petits mouvements de caméra : la représentation est insensible aux secousses de caméra

## Création de résumé

### Pourquoi

On veut fournir des informations pertinentes et concises afin d'aider l'utilisateur à indexer, à naviguer ou à organiser des fichiers vidéos plus efficacement.

- Résumé statique (video summary) : sélectionner les images les plus représentatives de la vidéo
  - Ces images appelée images clés se présentent en général sous la forme d'un scénarimage (storyboard)
- Résumé dynamique (video skimming) : version courte de la vidéo originale
  - Ex : construction automatique d'une bande-annonce du film

### Méthodes

On peut regrouper les familles de résumés statiques en quatre catégories distinctes :

- Méthodes reposant sur l'échantillonnage
- Méthodes reposant sur les plans
- Méthodes reposant sur les scènes
- …

### Résumé statique par échantillonnage

Dans le résumé statique par échantillonnage, il faut choisir les images clés en sous-échantillonnant uniformément ou aléatoirement la séquence originale

Inconvénient :

- Certaines parties de la vidéo ne seront pas représentées
- Redondance de certaines images clés avec un contenu similaire

Cette approche n'a pas d'attache au contenu de la vidéo

### Résumé statique par plan

Dans la résumé statique par plan, la détection des plans est réalisée pour mieux ajuster la sélection des images clés au contenu de la vidéo.
Pour ce faire, on va extraire des images-clés des différents plans extraits de la vidéo

- Cas 1 : Première image

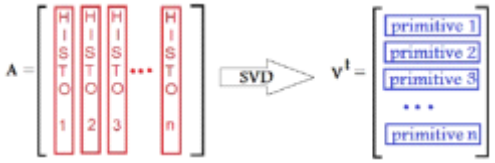
  On extrait la première image du plan comme image clé ou les première et dernière images du plan
- Cas 2 : utilisation de l'histogramme

  La première image du plan est sélectionnée comme image clé. Puis, si la

distance entre l'histogramme d'intensité de la dernière image clé sélectionnée et l'image courante est supérieure à un seuil, alors l'image courante est la nouvelle image clé.

### Résumé statique par scène

- L'idée est d'assigner chaque image à un groupe, puis à réunir les groupes les plus similaires de manière itérative jusqu'à un critère d'arrêt
- Une matrice est créée ou chaque colonne contient un vecteur caractéristique (l'histogramme) associé à un groupe d'images de la vidéo




Une SVD est réalisée pour réduire l'espace des caractéristiques.

#### Étapes :

- On crée la matrice d'histogrammes A avec tous les frames de la scène
- On réduit A avec la SVD (singular value decomposition) pour avoir l'espace de primitive réduit $V^t$
- Partitionnement : On choisit une image-clé de départ, puis
  - On classe en ordre les vecteurs-primitives ($V_p$) selon la distance avec le vecteur-primitive de l'image clé $S_1$
  - Pour chaque $V_p$ non classé, on trouve l'image ayant la distance minimum $d(S_{min})$
  - Si $d(S_{min}) < T$, on associe le $V_p$ à l'image $S_{min}$. Sinon, on crée une nouvelle image clée $S_c$
  - On arrête lorsque tous les $V_p$ sont associées à une image clé $S_c$
- On retourne les images clés $S_c$

#### Remarques :

- On travaille sur la totalité des frames de la scène
- Le temps de calcul peut être long

#### Solution :

Approche hiérarchique (séquentielle)

#### Définition du résumé hiérarchique à deux niveaux

- Premier niveau : on applique l'approche précédente avec une séquence fixe d'images (20 et plus)
- Deuxième niveau : on l'approche précédente, mais en utilisant les images clés sorties du niveau précédent

Cette approche permet de paralléliser la première étape et de réduire le coût de la décomposition SVD

# Lasse9

jeudi 2 juillet 2020 12:02

## Introduction

Indexer = attacher à une image/vidéo un ensemble de descripteurs de leur contenu
= extraire une information (ou clé d'accès à l'information = index) synthétique d'une image

But :
Mesurer la ressemblance avec les descripteurs de notre requête pour la recherche automatique des documents visuels

Bases de données multimédia
BD + Multimédia = BDMM
-> Stockage, organisation et interrogation des données multimédia (texte, son, image, vidéo)

Comment indexer et interroger des documents multimédia?
Indexation manuelle : L'approche la plus ancienne et répandue
- Par texte : annotation
- Par mots-clés, métadonnées

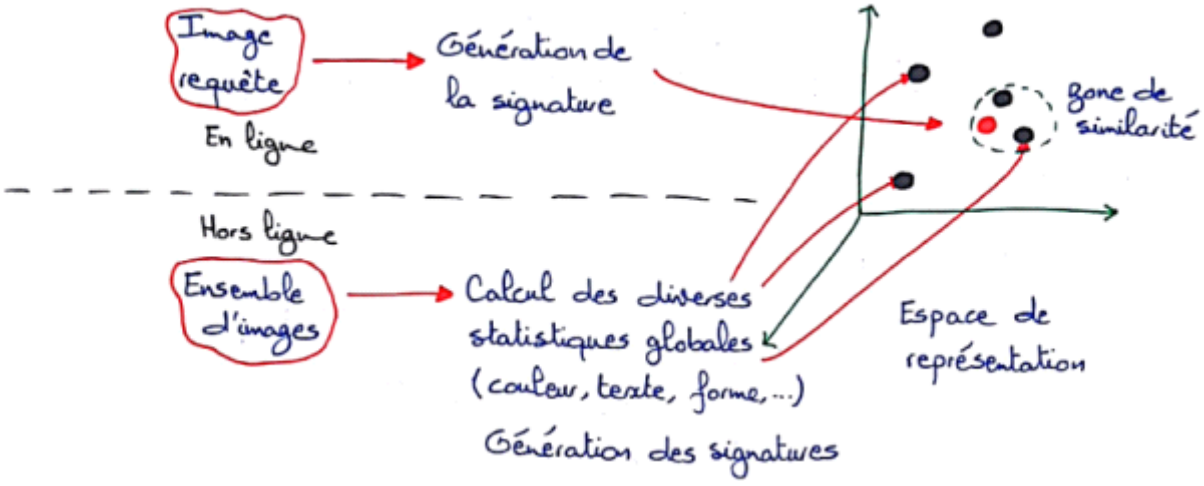
Indexation automatique : CBIR
- Offline : Production d'indexes issus de l'analyse du contenu des images
- Online : Gestion de requêtes des utilisateurs

Quelques définitions :
- Indices visuels : Caractéristiques de l'images (Couleur, forme, texture)
- Descripteurs : méthodes d'extraction du contenu visuel (Histogramme couleur ...)
- Signature : Vecteur numérique représentant le contenu visuel d'une image
- Espace de description des images
- Espace de recherche dans la base d'images

Distances :
Distance de Manhattan :
$$\sum_{i=1}^{n} |x_i - y_i|$$
Distance Euclidienne :
$$\sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$
Distance de Minkowski :
$$\sqrt[p]{\sum_{i=1}^{n} |x_i - y_i|^p}$$
Distance De Tchebychev :
$$\lim_{p \to \infty} \sqrt[p]{\sum_{i=1}^{n} |x_i - y_i|^p} = \max_{1 \le i \le n} |x_i - y_i|$$

## CBIR : Descripteurs globaux

CBIR = Content Based Image Retrieval (Indexation d'images par contenu)




Principe :

a. Choix d'un espace de représentation et d'un descripteur
b. Calcul de signature de l'image dans cet espace
c. Définition d'une mesure de similarité (distance) dans cet espace

Image numérique

3 grilles de valeurs, 1 grille par composante de couleur
RGB : 8 bits de quantification pour chaque couleur
=> 24 bits par pixel

Descripteurs d'images :

Description globale : considérer l'image dans son ensemble
Description locale :

Considérer l'image comme composée d'un ensemble d'objets
Détecter ls points d'intérêt et calculer les invariants autour de ces points d'intérêt

Description spécifique : Empreintes digitales - Visages - ...

Caractéristiques globales :

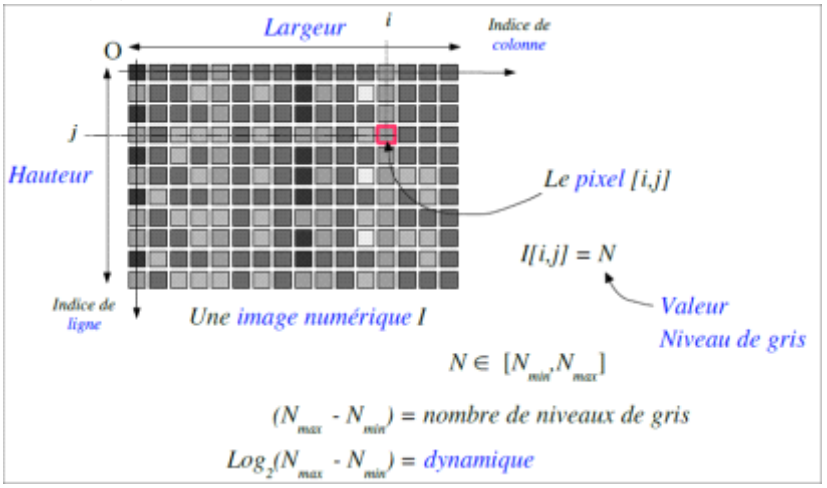
Couleur :

La valeur moyenne dans l'espace RGB :

Pixel [i,j] : Colonne i, Ligne j

$$X_{avg} = \frac{1}{N.M}\sum_{i=0}^{N}\sum_{j=0}^{M} X(i,j)$$

$X \in \{R, G, B\}$




Comparaison en utilisant la distance euclidienne :

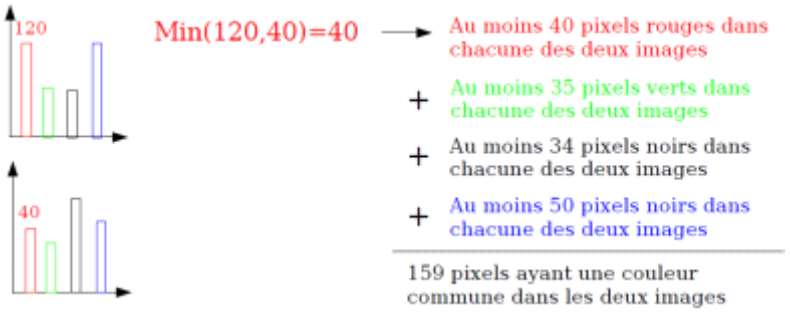
$$d^2_{avg}(x,y) = \left(R_{avg}x - R_{avg}y\right)^2 + \left(G_{avg}x - G_{avg}y\right)^2 + \left(B_{avg}x - B_{avg}y\right)^2$$

Histogramme :

L'intersection d'histogrammes :

$x_i$ : nombre de pixels de valeur $i$ dans l'histogramme 1
$y_i$ : nombre de pixels de valeur $i$ dans l'histogramme 2

$$d(X,Y) = 1 - \frac{\min(x_1,y_1) + \cdots + \min(x_d,y_d)}{\min(x_1 + \cdots + x_d, y_1 + \cdots + y_d)}$$




Distance de Minkowski :

$$d_r(h_1,h_1) = \sum_{i \in C} |h_1(i) - h_2(i)|^r$$

Distance quadratique :

$$d_q(h_1,h_2) = \sum_{j \in C} \sum_{i \in C} (h_1(i) - h_2(i))(h_1(j) - h_2(j))$$

Texture : "Zone homogène en un certain sens"

Matrice de cooccurrence (Mesures de Haralick) :

C'est une matrice $N_g \times N_g$

Où $N_g$ est le nombre de niveaux de gris d'une image

Idée : identifier les répétitions de niveaux de gris selon une distance (pas) et une direction

Distance : 1,2,3 ... pixels

Direction : 0°,45°,90°,135° (en degré, 0° signifie à droite du pixel, 90° signifie en bas du pixel)

| | | | |
|---|---|---|---|
| 1 | 4 | 4 | 3 |
| 4 | 2 | 3 | 2 |
| 1 | 2 | 1 | 4 |
| 1 | 2 | 2 | 3 |

*Image*

| | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1 | ? | ? | ? | ? |
| 2 | ? | ? | ? | ? |
| 3 | ? | ? | ? | ? |
| 4 | ? | ? | ? | ? |

*Matrice pour distance=1 et direction=0°*

Calcul :

| | | | |
|---|---|---|---|
| **1** | **4** | 4 | 3 |
| 4 | 2 | 3 | 2 |
| 1 | 2 | 1 | 4 |
| 1 | 2 | 2 | 3 |

*Image*

| | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1 | 0 | 0 | 0 | **1** |
| 2 | 0 | 0 | 0 | 0 |
| 3 | 0 | 0 | 0 | 0 |
| 4 | 0 | 0 | 0 | 0 |

*Matrice pour distance=1 et direction=0°*

| | | | |
|---|---|---|---|
| 1 | 4 | 4 | 3 |
| 4 | 2 | 3 | 2 |
| 1 | 2 | 1 | 4 |
| 1 | 2 | 2 | 3 |

*Image*

| | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1 | 0 | 2 | 0 | 2 |
| 2 | 1 | 1 | 2 | 0 |
| 3 | 0 | 1 | 0 | 0 |
| 4 | 0 | 1 | 1 | 1 |

*Matrice pour distance=1 et direction=0°*

Descripteurs :

Maximum :

Différence d'ordre k

Entropie

Uniformité

Histogramme de textures :

Distance $X^2$ :

$$X^2(h_1,h_2) = \frac{1}{2} \times \sum_{m=1}^{k} \frac{[h_1(m) - h_2(m)]^2}{h_1(m) + h_2(m)}$$

Forme

- Caractéristiques internes : surface, …
- Caractéristiques externes : périmètre, circularité, rectangularité, …

## CBIR : Descripteurs locaux

### Type de descriptions :

a. Description à base de régions
b. Description à base de points

### Définitions :

#### Points d'intérêts : coins, points particuliers des contours, sites informatifs de l'image

##### Méthodes :

1) SIFT, SURF
2) Détecteur de Harris

#### Régions d'intérêts : zones d'intérêts plus générales que les points

##### Méthodes :

### SIFT, SURF :

#### Étapes :

Détection des points d'intérêts dans chaque image
Description locale de chaque point d'intérêt
Mise en place d'une mesure de similarité
Mise en correspondance deux à deux les points entre chaque paire d'images

#### p.s :

Application en TD

### Détecteur de Harris

## CBVR :

### Définition :

- CBVR = Content Based Video Retrieval (Indexation de vidéos par contenu)
- Vidéo = Suite d'images + Son + texte
- Vidéo = Ensemble de scènes
- Scène = Ensemble de Plans
- Plan = Ensemble de Frames

### Le flux de données dans la vidéo numérique est (en bits/s) :

$F = Nb * f_y * f_x * f_t$

$Nb$ : Bits à encoder
$f_y$ : Lignes par frame
$f_x$ : Pixels par lignes
$f_t$ : Fréquence d'image (frames per second FPS)

### Segmentation en plan et en scène

- Intensité
- Histogramme

### Création de résumé :

- Résumé statique (video summary) -> Storyboard
  - Par échantillonnage
  - Par plan
  - Par scène
- Résumé dynamique (video skimming) -> Trailer