Bases de données multimédia
VII – Bag of words

ENSIMAG

2014-2015

Matthijs Douze & Karteek Alahari

Video-Google

- **LA référence :**
  Josef Sivic and Andrew Zisserman
  « Video Google: A Text Retrieval Approach to Object Matching in Videos »
  International Conference on Computer Vision, 2003 (ICCV 2003)
  - peu vidéo, peu Google...
  - recherche d'une zone sur une image

Video-Google

- **LA référence :**
  Josef Sivic and Andrew Zisserman
  « Video Google: A Text Retrieval Approach to Object Matching in Videos »
  International Conference on Computer Vision, 2003 (ICCV 2003)
  - peu vidéo, peu Google...
  - recherche d'une zone sur une image

- Idée : calquer le fonctionnement d'un moteur de recherche d'images sur celui d'un moteur de recherche de texte/page web

Video-Google

- **LA référence :**
  Josef Sivic and Andrew Zisserman
  « Video Google: A Text Retrieval Approach to Object Matching in Videos »
  International Conference on Computer Vision, 2003 (ICCV 2003)
  - peu vidéo, peu Google...
  - recherche d'une zone sur une image

- Idée : calquer le fonctionnement d'un moteur de recherche d'images sur celui d'un moteur de recherche de texte/page web

- Voir le demo: **http://www.robots.ox.ac.uk/~vgg/research/vgoogle/**

Video-Google

- **LA référence :**
  Josef Sivic and Andrew Zisserman
  « Video Google: A Text Retrieval Approach to Object Matching in Videos »
  International Conference on Computer Vision, 2003 (ICCV 2003)
  - peu vidéo, peu Google...
  - recherche d'une zone sur une image

- Idée : calquer le fonctionnement d'un moteur de recherche d'images sur celui d'un moteur de recherche de texte/page web

- La plupart des systèmes de recherche d'images/de vidéos de l'état de l'art capable d'indexer de très gros volumes de données sont dérivés de cette approche
  - D. Nister and H. Stewenius : CVPR 2006
  - H. Jegou, M. Douze and C. Schmid : ECCV 2008

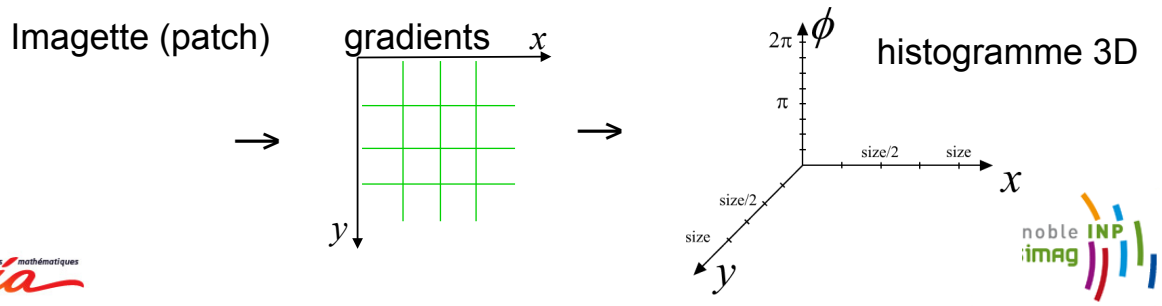
Étapes

- Description locale : détecteurs + descripteurs
- Filtrage spatial et temporel
- Bag-of-words
- Fichier inversé pour la requête
- Re-ranking après vérification de la cohérence spatiale

Video-Google : description locale

- Extracteurs
  - MSER
  - Shape adapted (point d'intérêt + sélection d'échelle + forme elliptique)

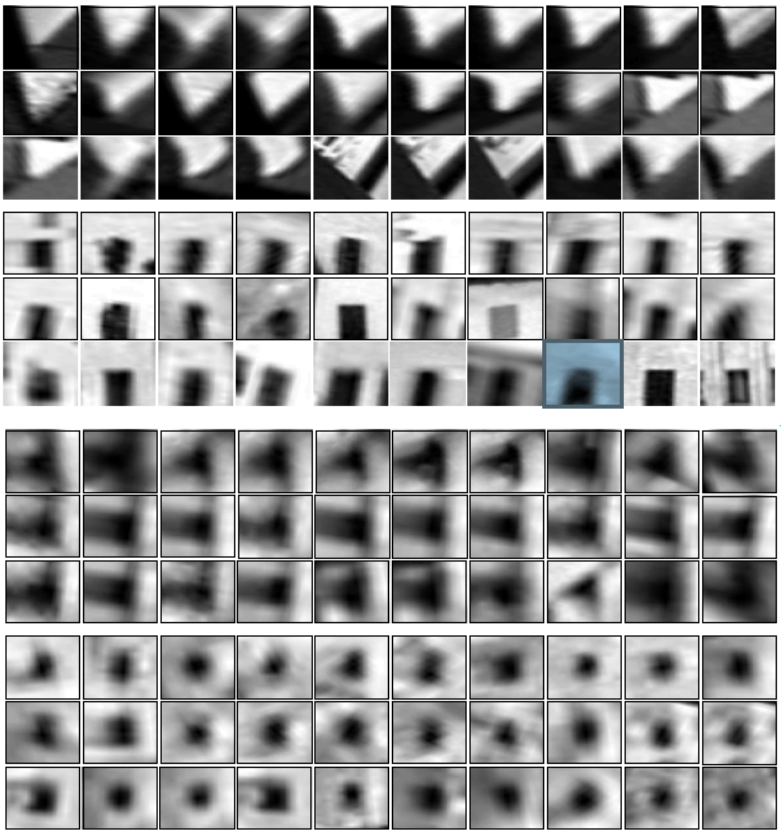
- Descripteurs SIFT (scale-invariant feature transform) [Lowe 04]



Video-Google : bag-of-words (début)

- L'objectif est d'obtenir l'équivalent des "mots" en requête textuelle
- Pour cela, quantification vectorielle des descripteurs via un *k-means* (rappel)
  - index de quantification = "mot visuel"
  - représentation grossière du vecteur de description local
  - pas d'information géométrique

Video-Google : bag-of-words (suite)



Exercice : où sont
- MSER
- Harris

Video-Google : bag-of-words (fin)

- Une fois le clustering défini (une fois pour toutes), un document est représenté par un histogramme de ses mots visuels
- Application d'une pondération "TFIDF" aux composantes de cet histogramme
  - objectif : donner plus d'importance aux mots visuels rares qu'à ceux qui sont fréquents
  - même pondération que celle utilisée en indexation de documents textuels

Nb d'occurrences de i dans doc d

Nb de documents dans la base

$$t_i = \frac{n_{id}}{n_d} \log\left(\frac{N}{n_i}\right)$$

Nb de mots dans doc d

Nb d'occurrences de i dans toute la base

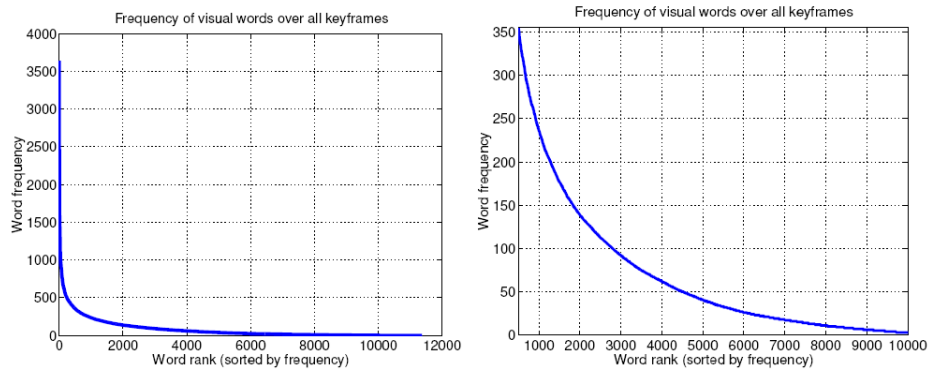
Video-Google : fichier inversé

- Utilisé pour comparer les vecteurs de fréquence épars
- Mesure de similarité usuellement utilisée : produit scalaire

- Analogie avec le texte
  - description d'un document
  - pondération des mots
  - utilisation de *stoplists*

Stoplists

- On supprime 5% en haut, 10% en bas

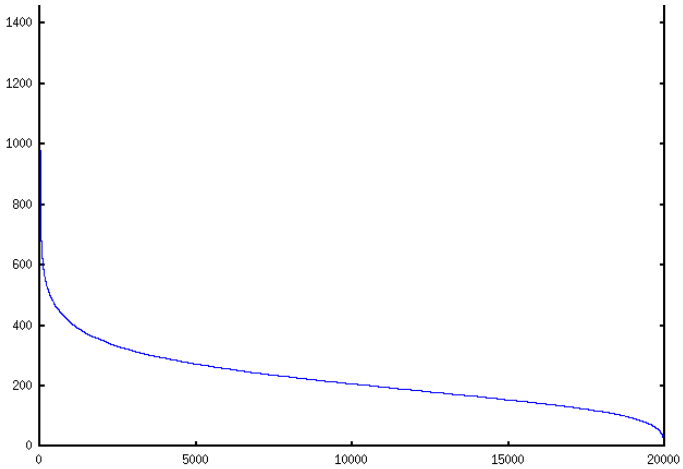


(a) (b)

Figure 5: Frequency of MS visual words among all 3768 keyframes of Run Lola Run (a) before, and (b) after, application of a stoplist.

Avec vocabulaire mieux équilibré

- Obtenu avec plus de données d'apprentissage pour k-means
- Stoplist moins utile



- Cas idéal : répartition uniforme.
  - complexité de la recherche en fichier inversé en $O(d\ s^2)$

Video-Google : filtrage spatial et temporel

- Idée : obtenue par analogie avec Google
  - Google améliore le score des pages pour lesquelles les mots-clés se retrouvent à des positions proches dans le texte
- Filtrage spatial
  - on ne garde un match que si ses voisins géométriques matchent aussi
  - variante : vérification géométrique plus forte (voir chapitre IV)

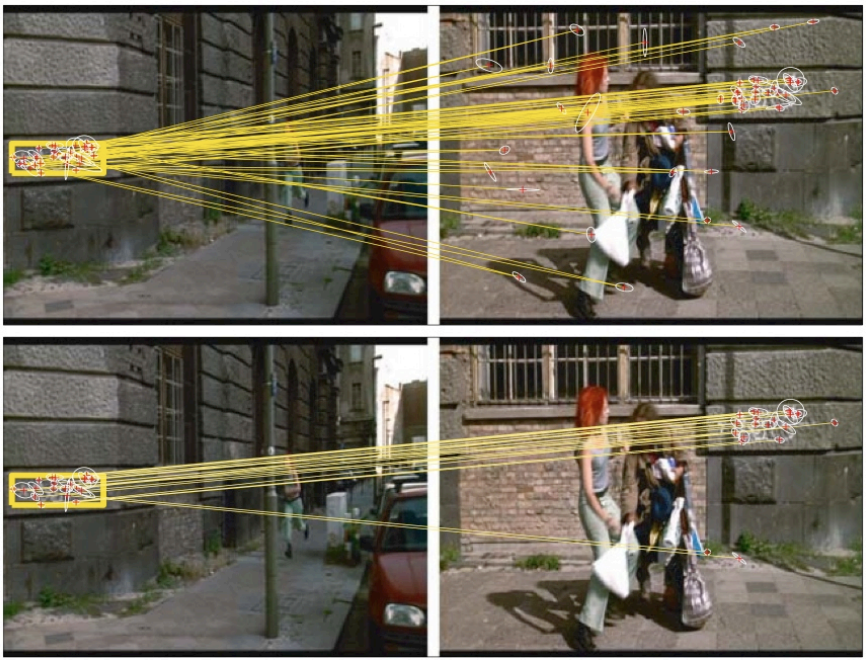
    → calcul de la transformée affine
- Filtrage temporel (pour les vidéos uniquement)
  - utilisation d'un modèle dynamique simple d'une frame à l'autre
  - on ne garde une région de description que si elle est suffisamment stable d'une frame sur l'autre

Vérification spatiale

- Avant et après

Vérification spatiale

- Avant et après

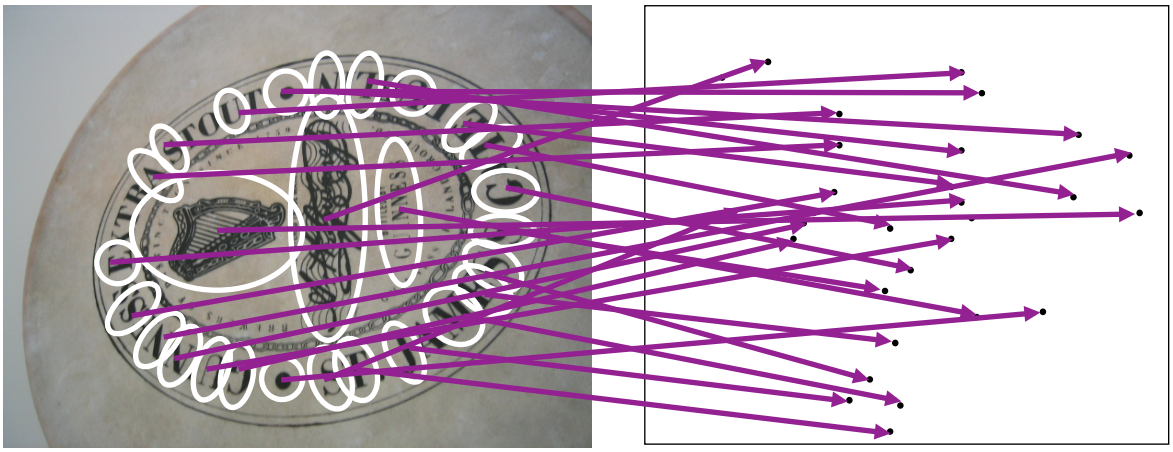


Variante 1 : méthode hiérarchique de Nister et Stewénius

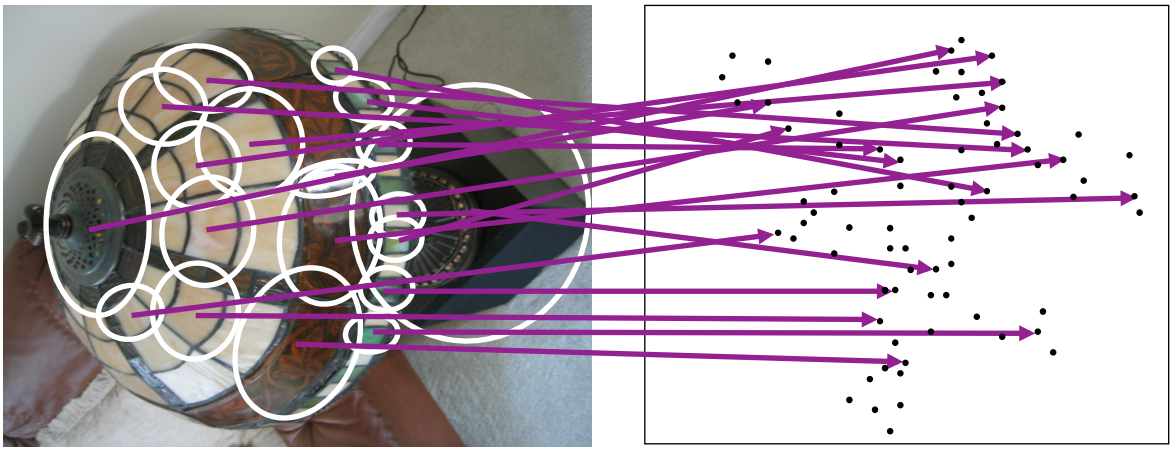
- Motivation : réduire le coût d'assignement des "visual words"
- Comment : utilisation d'un vocabulaire visuel hiérarchique

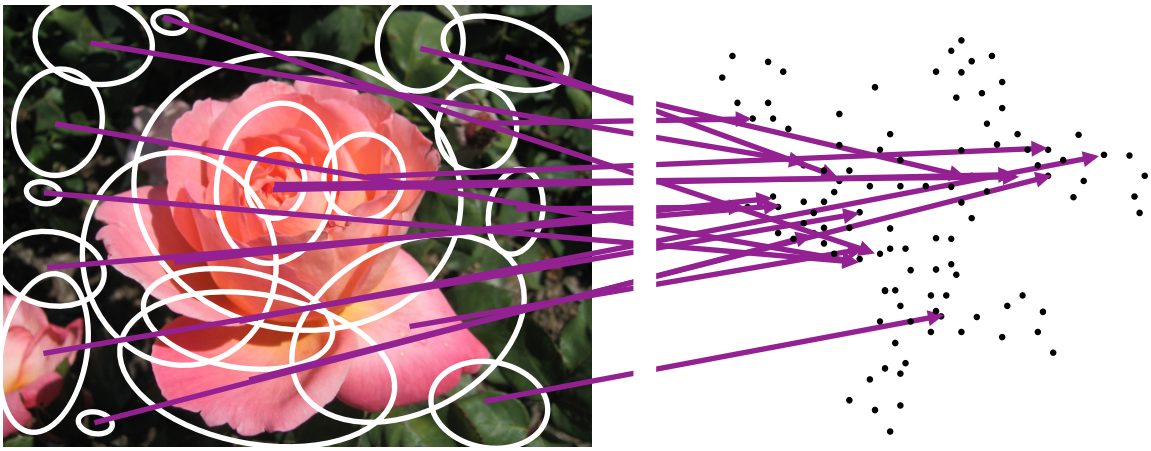
- Autre éléments de la chaîne modifiés :
  - modification de la métrique de comparaison des mots visuels (L1)

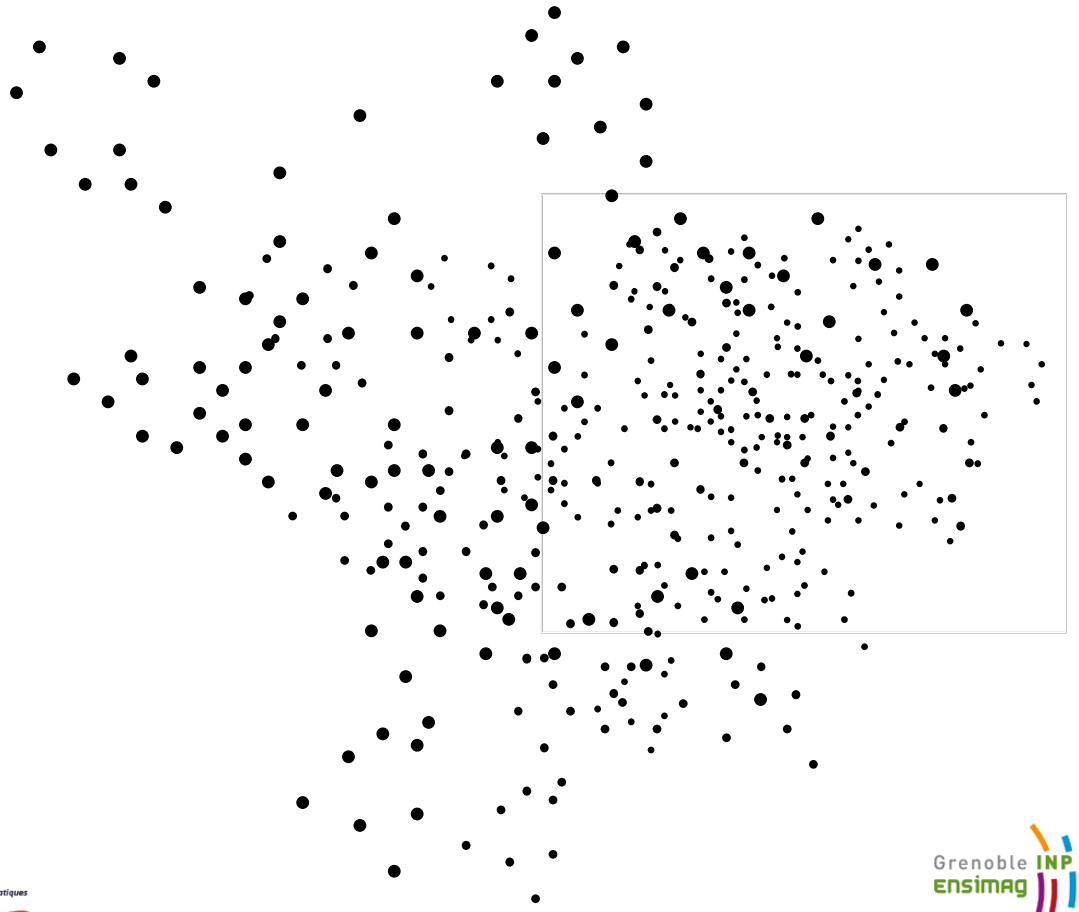
Scalable Recognition with a Vocabulary Tree
David Nistér and Henrik Stewénius, CVPR 06

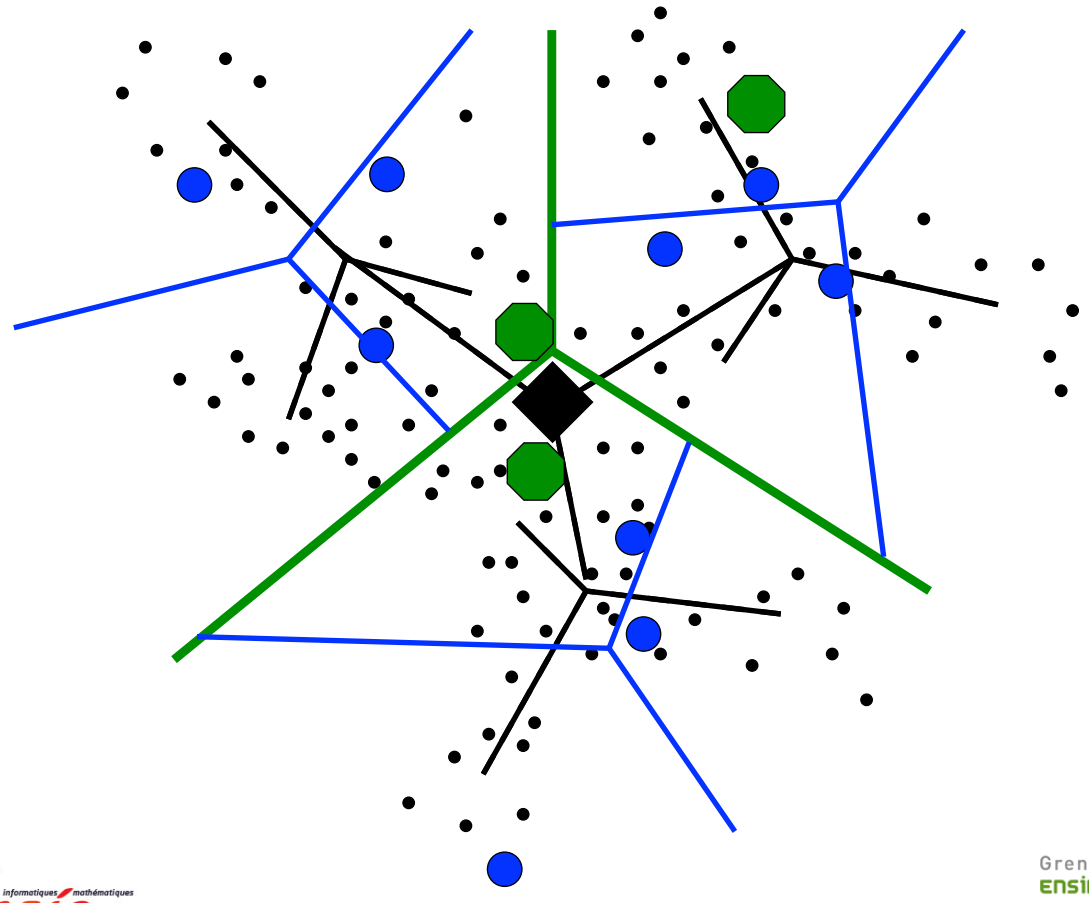


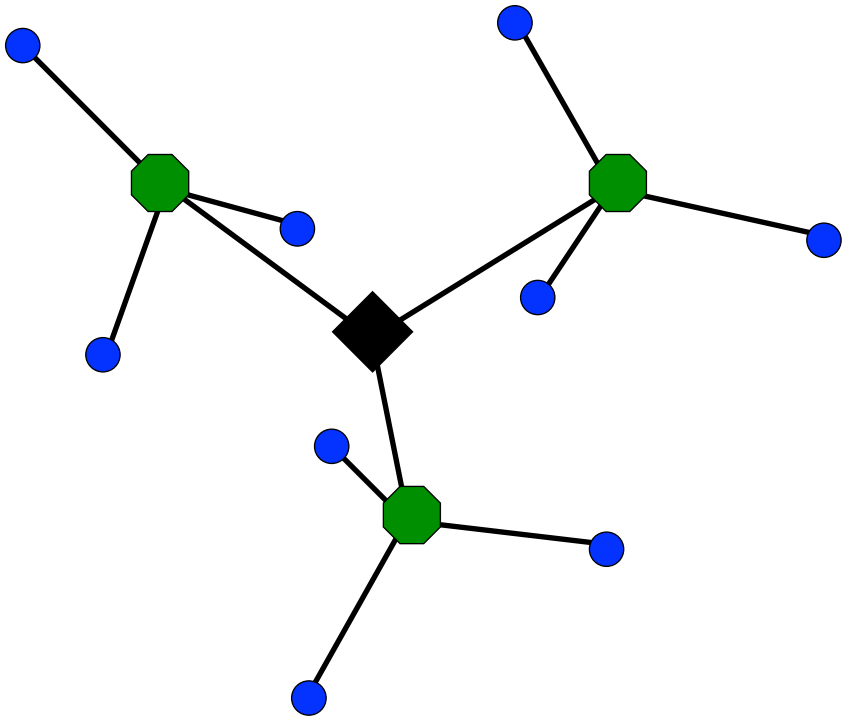
Slides courtesy D. Nister

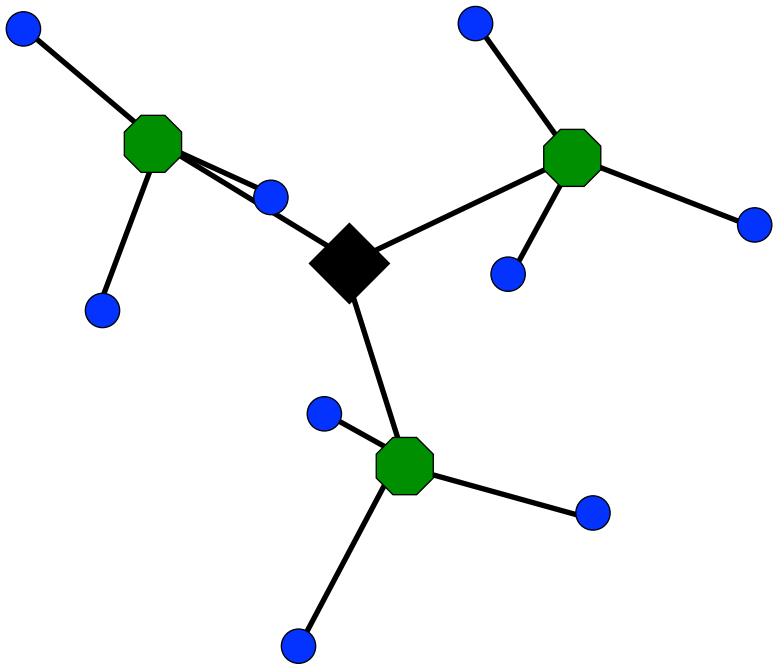


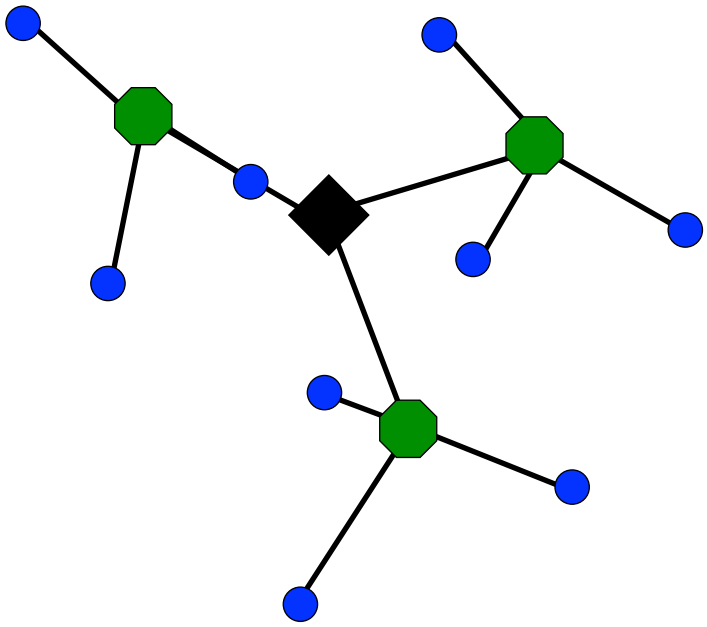


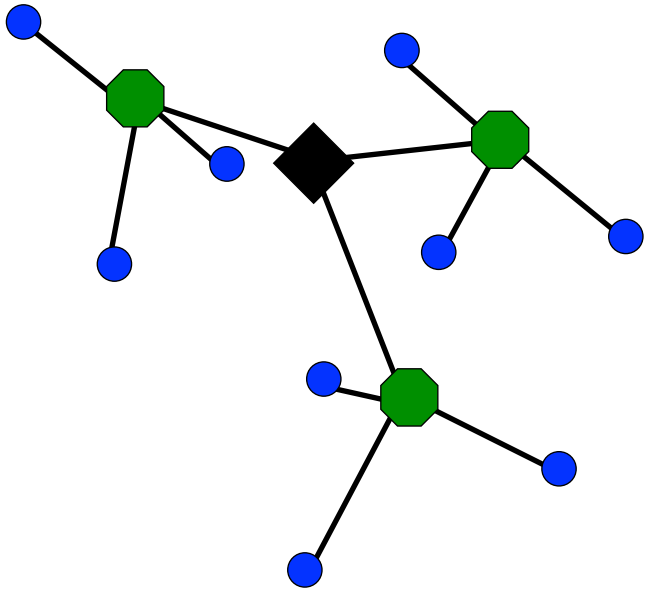


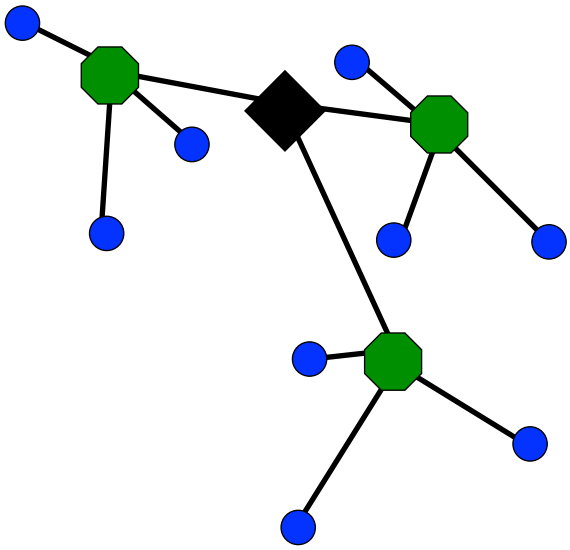


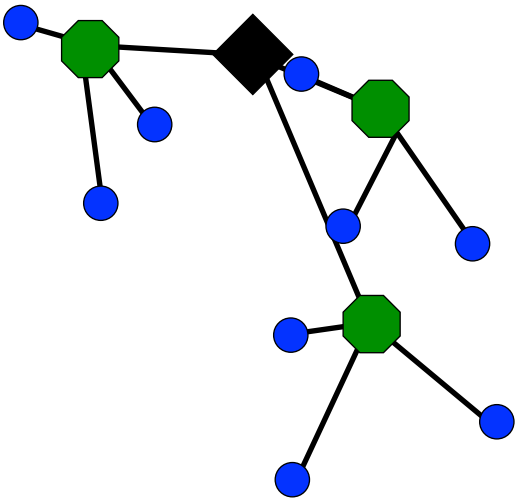


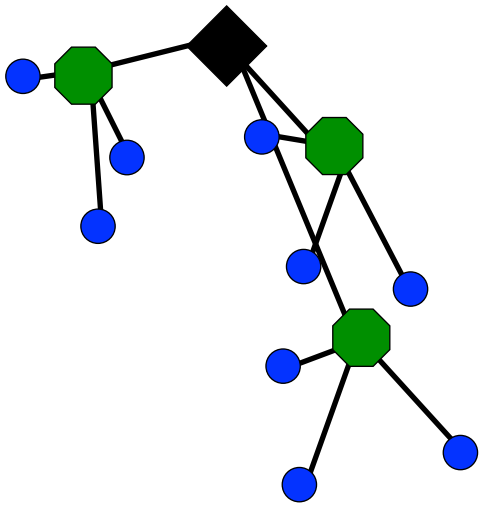


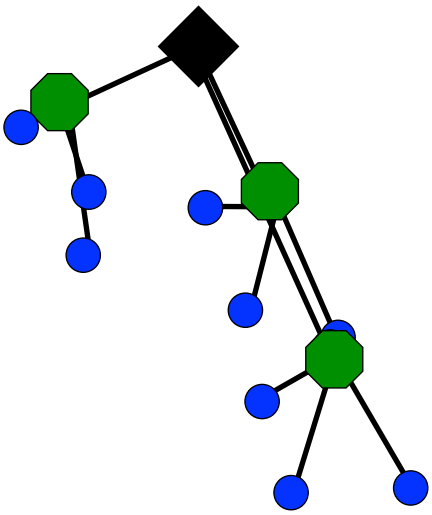


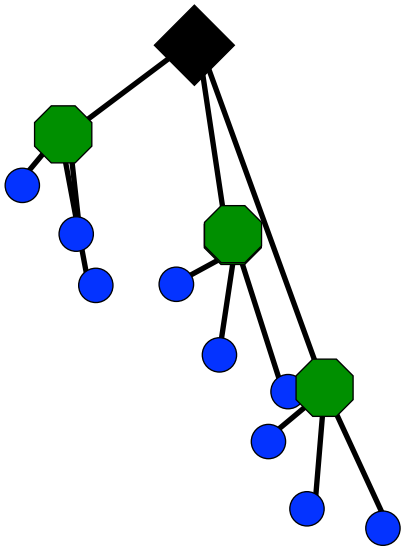


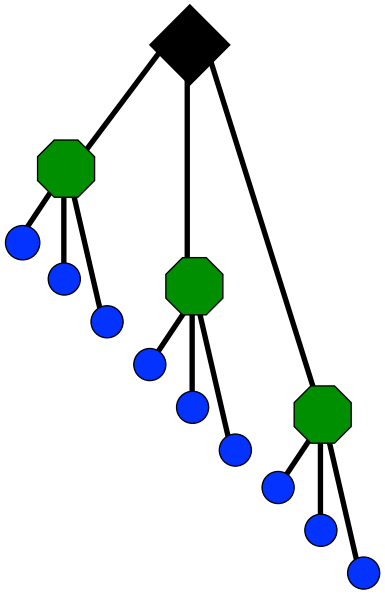


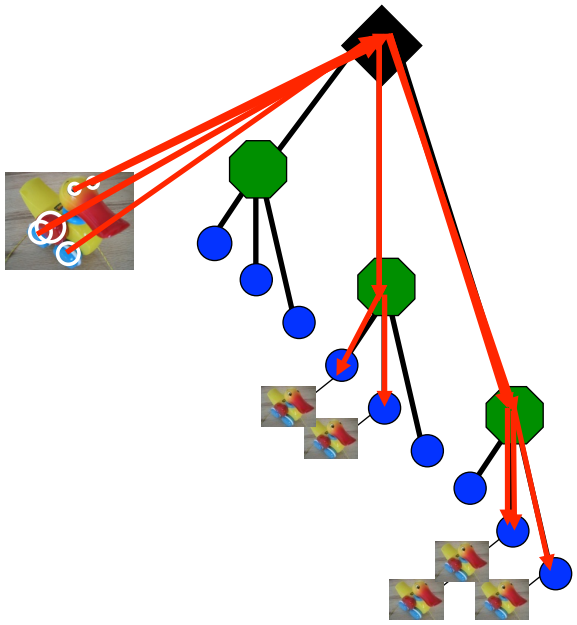


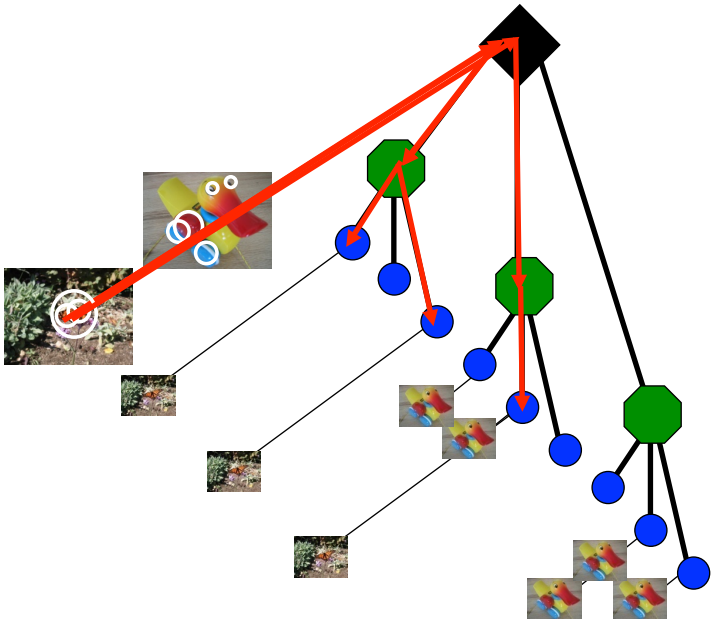


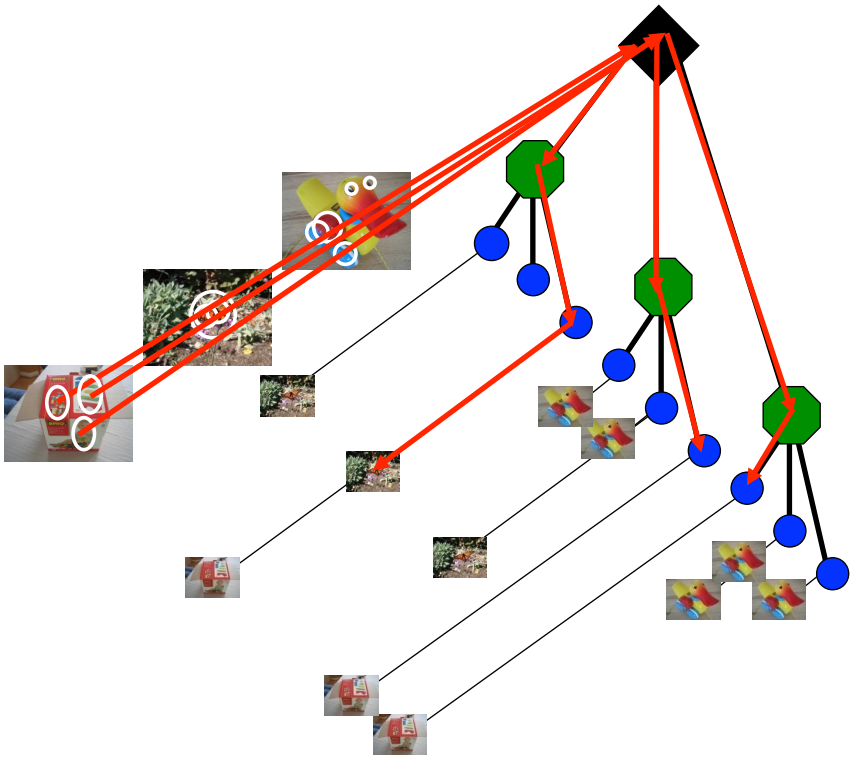


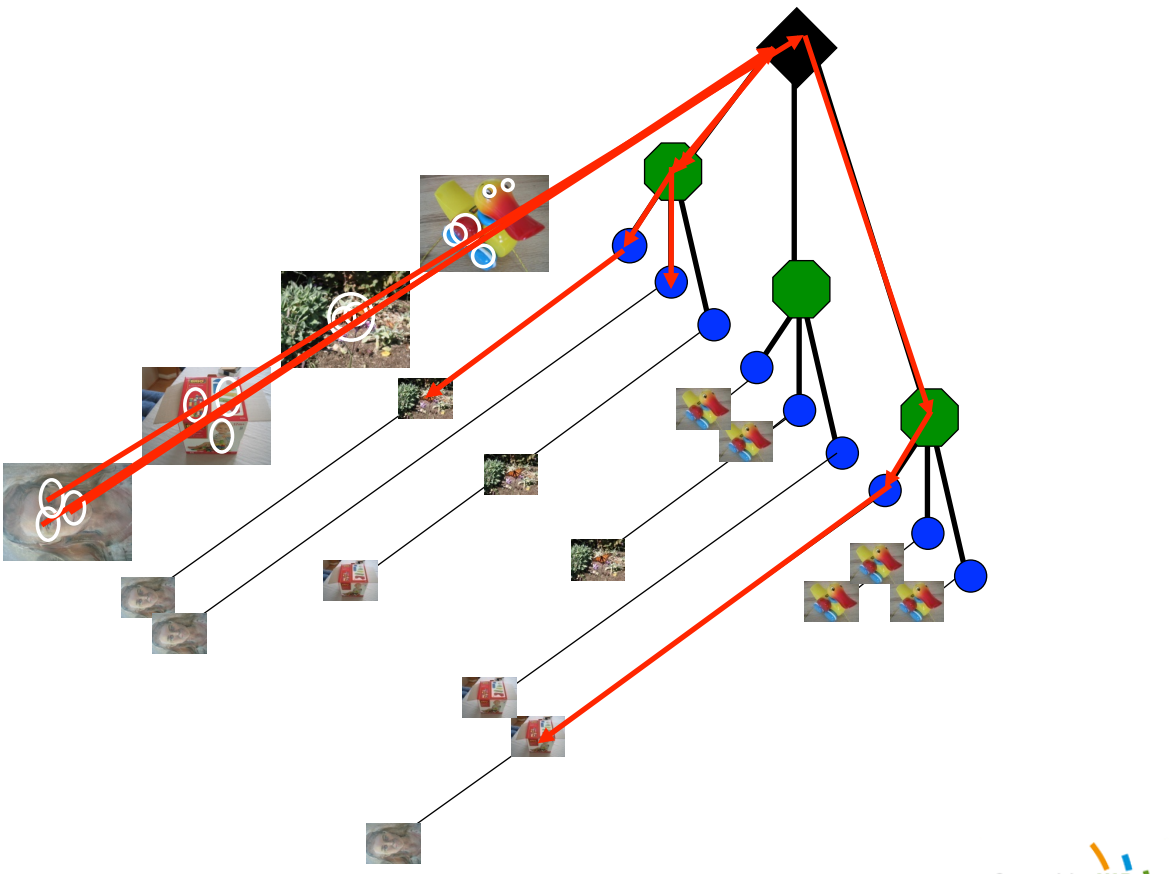


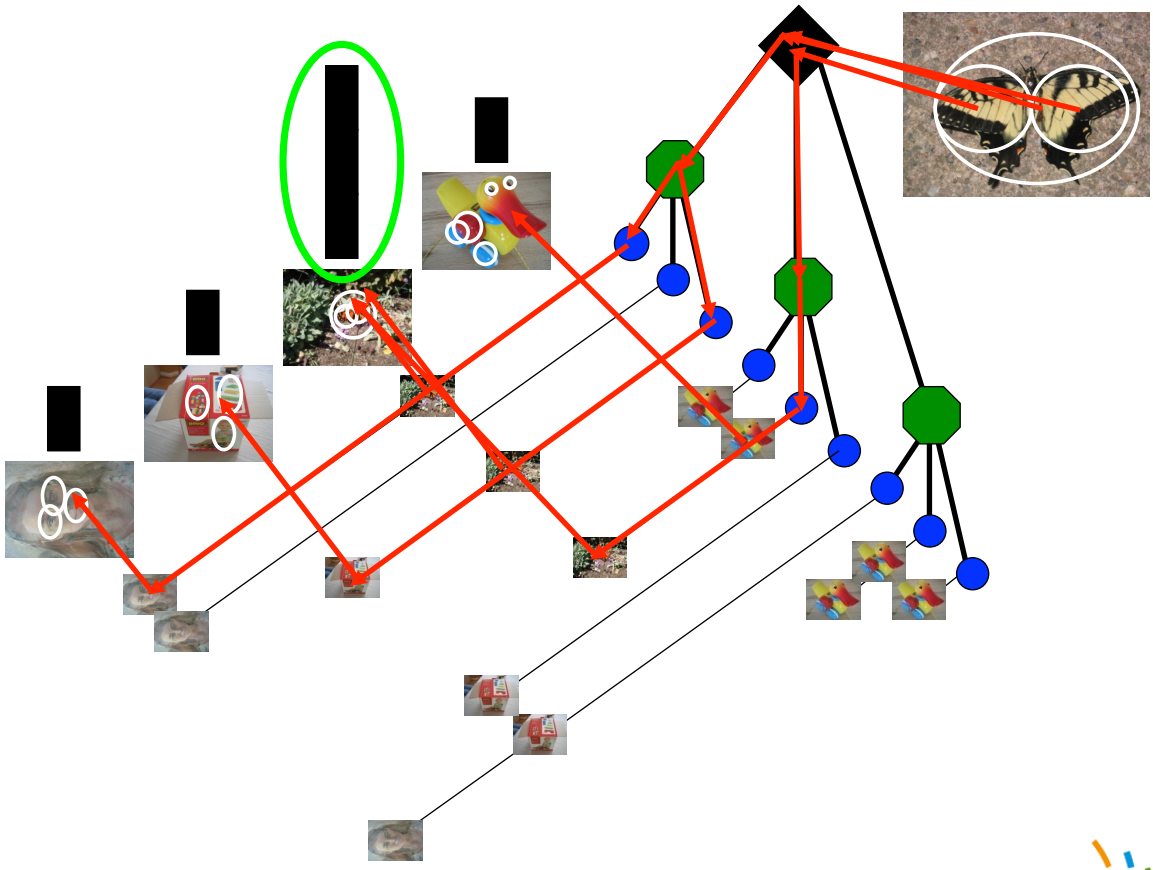












Variante 2 : Jégou, Douze et Schmid

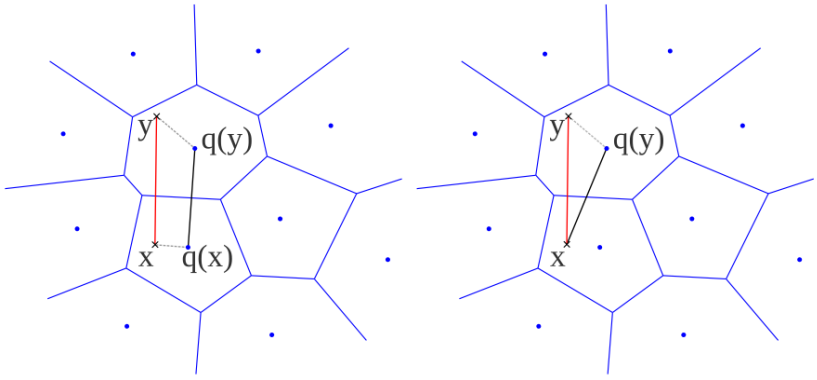
- Principales idées :
  - interpréter le système de sac-de-mots comme un système de recherche approximative de plus proche voisins
  - Augmenter la sélectivité de celui-ci en raffinant la description du descripteur
  - Utilisation de la “Weak geometry consistency” pour intégrer de l’information spatiale

- Méthode état de l’art en indexation image et vidéo
  - utilisée dans la compétition de copyright vidéo Trecvid 08
  - voir exposé séparé dans le chapitre suivant sur l’indexation video

Hamming embedding and weak geometric consistency for large scale image search
H. Jegou, M. Douze and C. Schmid, ECCV 08

Indexation de descripteurs globaux

- Exemple: utiliser LSH
  - trop grande dimension
- Indexation par quantification
  - quantification en “mots visuels” (cette fois-ci pour toute l'image)
  - trouver les mots visuels communs
    - peu discriminant
- Recherche assymétrique
  - sans quantification du descripteur de l'image requête



Conclusion

- Bag of words
  - principe de base pour l'indexation dans des bases de taille moyenne (< 10 M d'images)
  - descripteurs locaux → global
- Nombreuses améliorations
  - finesse de la représentation
  - géométrie