Indexation et recherche d'images par combinaison d'informations textuelles et visuelles

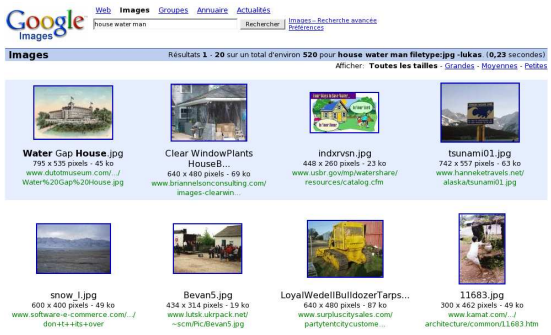
Sabrina TOLLARI

Paris, le 4 octobre 2007

Plan

- Motivation et problématique :
  - Trouver des méthodes de combinaison d'informations textuelles et visuelles efficaces et efficientes pour améliorer la recherche d'images
- Méthodologie et corpus
- Contributions
  - DIMATEX : un système rapide d'auto-annotation
  - Recherche des caractéristiques visuelles d'un mot
  - Sélection des dimensions sur une base d'images mal annotées
- Perspectives
- Résumé des résultats de la tâche2 de ImagEVAL

Systèmes de recherche d'images sur le Web



Indexation d'images

- Indexation textuelle
  - Manuelle : coûteuse, subjective
  - Automatique à partir du nom, de la légende ou du texte entourant l'image
  - Ne décrit pas le contenu de l'image, beaucoup d'erreurs d'indexation, mais apporte des informations sémantiques
- Indexation visuelle
  - Couleurs, formes, textures
  - Localisation, régions d'intérêt, segmentation
  - Décrit le contenu visuel de l'image, mais extraction de la sémantique difficile !

Indexation visuelle et fossé sémantique



(a) (b) (c)

Les images (a) et (b) ont des descripteurs de couleurs similaires, mais un sens différent.

Les images (b) et (c) ont des descripteurs de couleurs différents, mais un sens similaire.

« The semantic gap is the lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data have for a user in a given situation » (Smeulders et al., 2000).

Problématique générale

- Trouver des méthodes efficaces de fusion des informations textuelles et visuelles
  - pour améliorer les systèmes de recherche d'images
  - à partir d'une base d'images généralistes annotées pour lesquelles les descripteurs visuels sont connus d'avance
- Difficulté à prendre en compte :
  - le passage à l'échelle
    - Les techniques classiques de recherche et d'apprentissage ne sont pas forcément efficaces et efficientes sur de grandes bases d'images

Méthodologie et corpus

Méthodologie générale

- Pour pouvoir mesurer la capacité de nos systèmes à fusionner informations textuelles et visuelles, nous proposons de les évaluer pour différentes tâches :
  - Pour la tâche d'auto-annotation d'images à partir du contenu visuel
  - Ou pour la tâche de classification d'images
- Nous utiliserons dans les deux cas le même ensemble d'images généralistes annotées :
  - le corpus COREL
  - et le score normalisé

  Tous deux utilisés par de nombreuses équipes de recherche

Le corpus COREL (1/3)

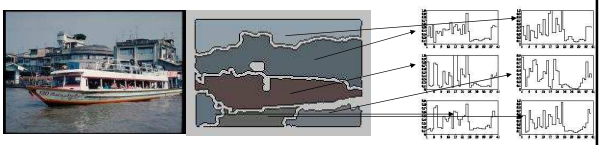
- 10 000 images généralistes fournies par James Z. Wang
  http://wang.ist.psu.edu
- Chaque image est :
  - Annotée par de 1 à 5 mots-clés choisis manuellement dans un lexique de 250 mots-clés environ
  - Segmentée en 10 régions maximum
  - Chaque région (appelée « blob ») est décrite par un vecteur de 40 composantes visuelles
- Les annotations, les segmentations et les vecteurs visuels proviennent des données utilisées dans l'article :

  Kobus Barnard, P. Duygulu, N. de Freitas, D. Forsyth, D. Blei, and M. I. Jordan, *« Matching Words and Pictures »* , Journal of Machine Learning Research, Vol 3, pp 1107-1135, 2003.

Le corpus COREL (2/3)

- Algorithme de segmentation utilisée par K. Barnard et al. :

  J. Shi, J. Malik, « Normalized Cuts and Image Segmentation », IEEE Patterns Analysis and Machine Intelligence, vol.22, n°8, 2000



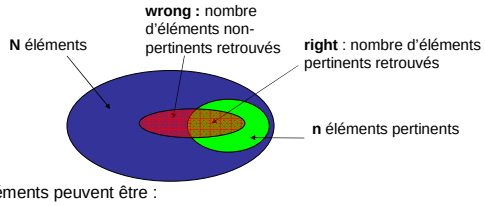
water boat harbor building

Le corpus COREL (3/3)

- Chaque blob de l'image est décrit par un vecteur visuel de 40 composantes extraites par K. Barnard et al. :
  - 6 dimensions de formes (aire du blob…)
  - 18 dimensions de couleurs (RVB, rvS, Lab),
  - 16 dimensions de textures (filtres gaussiens…).
- Nous avons normalisé le corpus :
  - par estimation MLE de distributions Gamma des vecteurs visuels pour la génération de distributions de probabilités et supprimer les artefacts. Les valeurs sont comprises entre 0 et 1.
- **Nous appelons par la suite cet espace à 40 dimensions l'espace U**

Le score normalisé (NS)

$$-1 \leq NS = \underbrace{right/n}_{\text{sensibilité}} - \underbrace{wrong/(N-n)}_{\text{1-spécificité}} \leq 1$$



Les éléments peuvent être :

- Les mots prédits pour chaque image dans le cas de l'auto-annotation
- Les images dans le cas de la classification

Plan

- Motivation et problématique :
  - Trouver des méthodes de fusion d'informations textuelles et visuelles efficaces et efficientes pour améliorer la recherche d'images
- Méthodologie et corpus
- **Contributions**
  - **DIMATEX : un système rapide d'auto-annotation**
  - **Recherche des caractéristiques visuelles d'un mot**
  - **Sélection des dimensions sur une base d'images mal annotées**
- Perpectives
- Résumé des résultats de la tâche2 de ImagEVAL

Construction de la table de distributions jointes

- Principe des VA-Files (Weber et al., 1998):
  - Chaque dimension de l'espace visuel est séparée en deux segments
  - L'espace est partitionné en 2^n clusters
  - Chaque vecteur visuel de l'ensemble d'apprentissage est codé en une séquence de bits de longueur n

Cluster 0000110110001 Cluster 0111010110000 Cluster 1111111111110

DIMATEX : un système rapide d'auto-annotation d'images

DIMATEX : un système rapide d'auto-annotation d'images

- Définition : l'annotation automatique (ou auto-annotation) consiste à associer un groupe de mots à une image uniquement à partir du contenu visuel de cette image
- Principe de notre modèle:
  - Construction d'une table de distributions jointes entre informations textuelles et visuelles à partir des données d'apprentissage à l'aide d'une technique issue des bases de données (VA-Files)
  - Ajout d'un modèle probabiliste simple afin de prédire une distribution de mots pour une nouvelle image

Construction de la table de distributions jointes

- La table de distributions jointes est estimée ainsi :
  - pour tout mot w et pour tout cluster C_k :

$$\mathbf{P}(w, C_k|A) = \sum_{J \in A} \mathbf{P}(J|A) \sum_{b \in J} \mathbf{P}(w|C_k, b, J, A) \mathbf{P}(C_k|b, J, A) \mathbf{P}(b|J, A)$$

  - où l'on peut supposer que :
    - P(J|A) suit une distribution uniforme
    - P(w|C_k,b,J,A)=1 si w appartient aux mots annotant J, 0 sinon
    - P(b|J,A) ≅P(b|J) et $\mathbf{P}(b|J) = \frac{\mathbf{aire}(b)}{\sum_{b_p \in J} \mathbf{aire}(b_p)}$
    - P(C_k|b,J,A)=1 si C_k=C(b), 0 sinon

Associer des mots à une image

- Associer des mots à un blob :

$$\mathbf{P}(w|b, A) = \mathbf{P}(w|C(b), A) = \frac{\mathbf{P}(w, C(b)|A)}{\mathbf{P}(C(b)|A)}$$

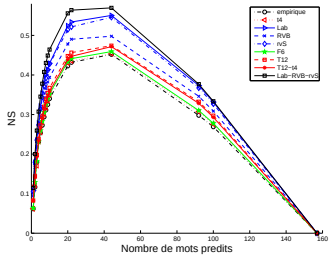
- Associer des mots à une image :

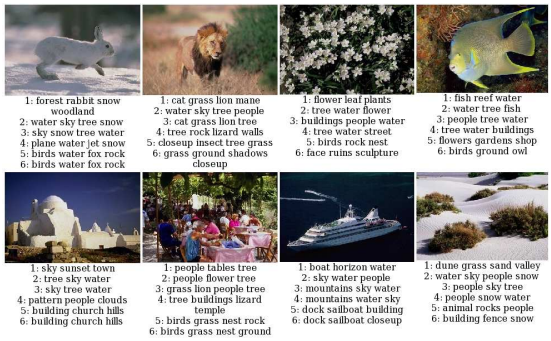
$$\mathbf{P}(w|I, A) = \sum_{b \in I} \mathbf{P}(w|b, I, A) \mathbf{P}(b|I, A)$$

où $\mathbf{P}(b|I, A) \cong \mathbf{P}(b|I) = \frac{\mathbf{aire}(b)}{\sum_{b_p \in I} \mathbf{aire}(b_p)}$

Expérimentations

- Corpus COREL : 7000 images d'apprentissage, 3000 images de test, lexique composé des 157 mots annotant au moins 20 images d'apprentissage





1. Annotation manuelle 2. DIMATEX 3. PLSA-WORDS
4. PLSA-WORDSFEATURES 5. DIRECT 6. LSA (Monay & Gatica-Perez, 2004)

Comparaison avec les modèles de l'état de l'art

Références	Modèles	NS	ΔNS	Gain NS
(Barnard et al., 2003)	empirique	0.425	-	-
	binary-D-2-region-cluster	0.604	0.179	+42%
	MoM-LDA	0.536	0.107	+25%
(Monay & Gatica-Perez, 2004)	empirique	0,427	-	-
	LSA	0.540	0.113	+26%
	PLSA-WORDS	0.571	0.144	+34%
DIMATEX (2005)	empirique	0.453	-	-
	Lab-RVB-rvS	0.583	0.132	+29%

Complexité

- Le modèle hiérarchique *binary-D-2-region-cluster* (Barnard et al., 2003) nécessite pour être optimal en moyenne 10 itérations de l'algorithme EM pour un total de 511 nœuds. Chaque nœud nécessite l'apprentissage de plusieurs paramètres.
- Le modèle *PLSA-WORDS* (Monay & Gatica-Perez, 2004) nécessite l'apprentissage de plusieurs distributions de probabilités pour chacune des modalités, chacune nécessite plusieurs itérations de l'algorithme EM.

Complexité de DIMATEX

- Le modèle DIMATEX
  - ne nécessite aucun apprentissage
  - ne possède aucun paramètre à optimiser
- Sa complexité moyenne est celle des VA-Files. C'est-à-dire :
  - O(1) pour insérer un vecteur visuel dans la table
  - O(1) pour annoter une image
- Une seule difficulté : la taille de la table de distributions jointes croît de manière exponentielle avec le nombre de dimensions visuelles.
  - De manière expérimentale, nous montrons que pour un nombre de dimensions supérieur à 15, les performances du système diminuent.

Conclusion sur DIMATEX

- Le système DIMATEX obtient des scores similaires aux modèles de l'état de l'art.
- C'est un système d'annotation rapide à condition que le nombre de dimensions de l'espace visuel reste raisonnable
- Il a l'avantage d'être dynamique dans sa phase d'entraînement
  - Perspective : remplir la table de distributions jointes avec des données en provenance du Web pour pouvoir annoter de nouvelles images
- Le système DIMATEX :
  - ne permet pas de savoir quelles sont les caractéristiques visuelles d'un mot
  - n'utilise pas de critères pour sélectionner les dimensions visuelles les plus pertinentes

Une méthode de recherche des caractéristiques visuelles d'un mot

25

Qu'est ce qui caractérise le mot « tiger » ?

tiger stone water

tiger water ground

tiger bengal grass

tiger forest

26

Comment trouver les caractéristiques visuelles d'un mot ?

- Principe :
  - pour trouver les régions d'images similaires qui caractérisent un mot, utiliser une méthode classique d'apprentissage
- Proposition :
  - la classification ascendante hiérarchique (CAH)
    - Avantages :
      - Maîtrise des paramètres
      - Clusters visuels interprétables

27

Construction de clusters visuels par CAH

Corpus d'images indexées textuellement et visuellement

28

Construction de clusters visuels par CAH

Corpus d'images indexées textuellement et visuellement

Séparation aléatoire

Base d'apprentissage | Base de validation | Base de test

29

Construction de clusters visuels par CAH

Corpus d'images indexées textuellement et visuellement

Séparation aléatoire

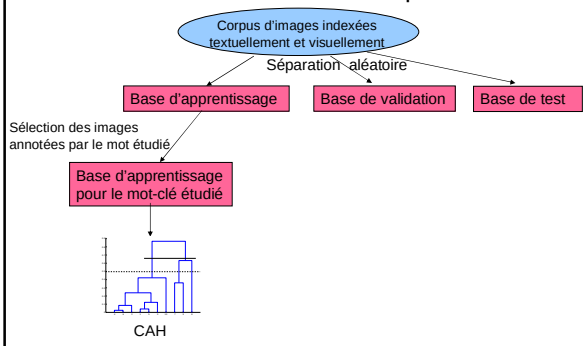
Base d'apprentissage | Base de validation | Base de test

Sélection des images annotées par le mot étudié

Base d'apprentissage pour le mot-clé étudié

30

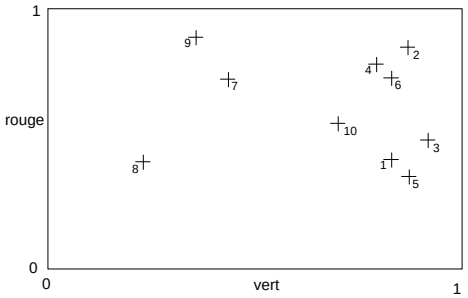
Construction de clusters visuels par CAH



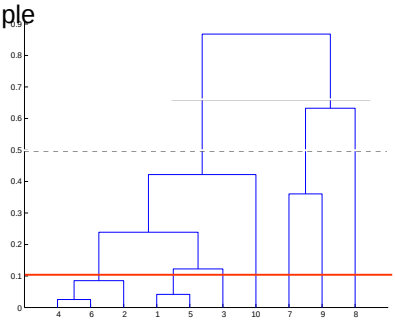
Algorithme de la CAH : exemple

Algorithme de la CAH : exemple

+ Vecteurs visuels des images d'apprentissage annotées par le mot « soleil »

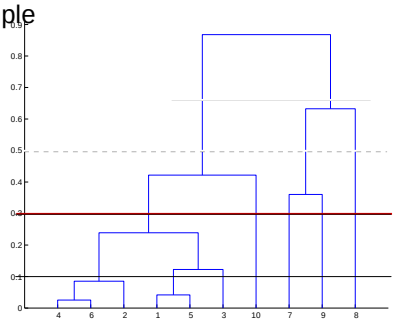


Construction de clusters visuels par CAH : exemple



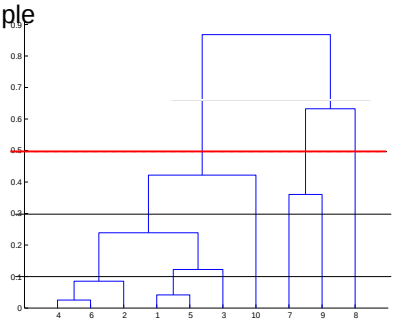
Vecteurs des images d'apprentissage du mot « Soleil »

Construction de clusters visuels par CAH : exemple



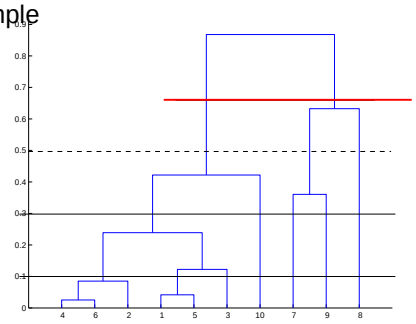
Vecteurs des images d'apprentissage du mot « soleil »

Construction de clusters visuels par CAH : exemple



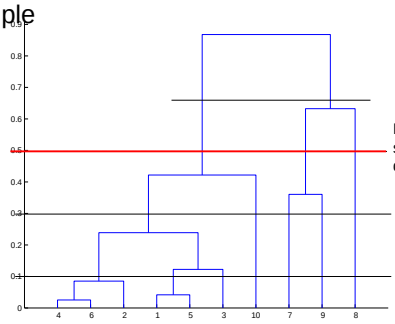
Vecteurs des images d'apprentissage du mot « soleil »

Construction de clusters visuels par CAH : exemple



Vecteurs des images d'apprentissage du mot « soleil »

Construction de clusters visuels par CAH : exemple

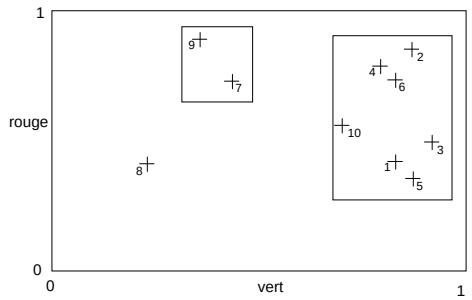


Meilleur score sur l'ensemble de validation

Vecteurs des images d'apprentissage du mot « soleil »

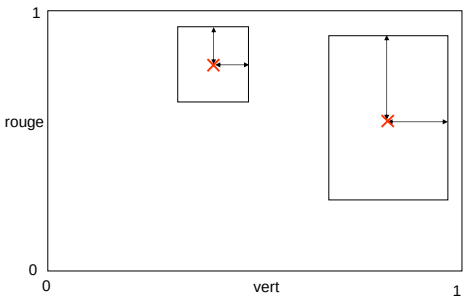
Algorithme de la CAH : exemple

+ Vecteurs visuels des images d'apprentissage annotées par le mot « soleil »

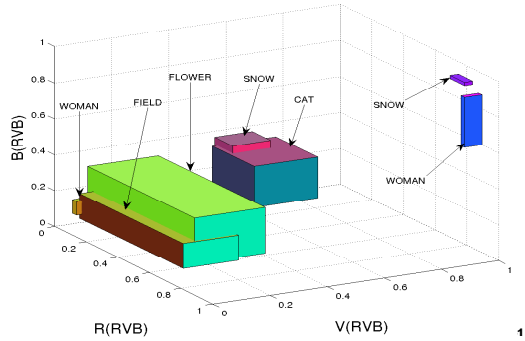


Algorithme de la CAH : exemple

Clusters du mot « soleil » stockés en gardant uniquement la moyenne et l'écart-type pour chaque dimension



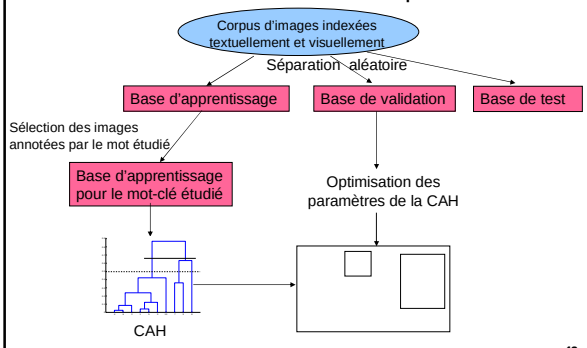
Exemple de clusters visuels de plusieurs mots dans l'espace visuel



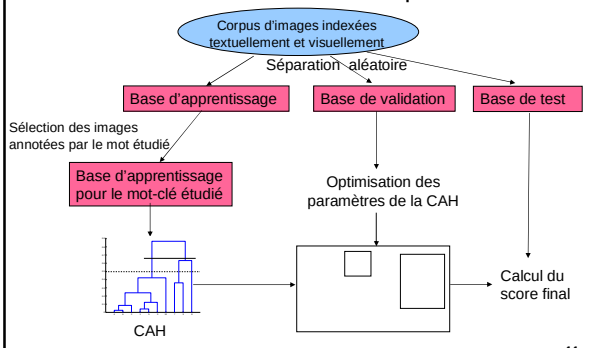
Évaluation de la qualité des clusters obtenus

- Un blob est annoté par un mot s'il appartient à l'un des clusters de ce mot
- Une image est annotée par un mot si au moins B blobs de cette image sont annotés par ce mot

Construction de clusters visuels par CAH



Construction de clusters visuels par CAH

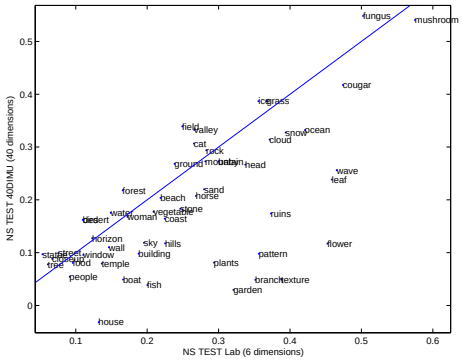


Expérimentations

- Corpus COREL
  - 5000 images d'apprentissage
  - 2500 images de validation
  - 2500 images de test
- Espaces visuels :
  - Lab : 6 dimensions visuelles de couleurs
  - 40DIMU : 40 dimensions visuelles de l'espace U

	40DIMU	Lab
Nombre de dimensions	40	**6**
NS moyen sur 52 mots (validation)	0.236	**0.290**
NS moyen sur 52 mots (test)	0.192	**0.248**

Lab versus 40DIMU



Application : Filtrage de l'indexation textuelle d'images par le contenu visuel

- On suppose que tous les mots du lexique représentent le texte associé à l'image (ici 52 mots supposés extraits d'une page web)
- On filtre les mots avec les clusters visuels des mots obtenus par CAH
- On calcule le score NS à partir des mots associés initialement à l'image

Image 172052 (10 blobs)
Légende (3 sur 3)
water(OK) mountain(OK) coast(OK)
sensi=1.00 specif=0.65
preci=0.15 NS=0.65

20 mots associés par le système
desert(7) water(6) sky(6) wave(6) hills(6) closeup(6) mountain(6) coast(6) tree(6) beach(6) boat(5) branch(5) temple(5) fish(4) sand(4) forest(4) cloud(4) people(4) horizon(3) valley(3)

32 mots non associés
snow(2) statue(2) vegetable(1) rock(1) bird(1) wall(1) flower(1) head(1) building(1) window(1) woman(1) street(1) plants(1) field(1) cat cougar food fungus garden grass ground horse house ice leaf mushroom ocean pattern ruins stone texture

Total: 52 mots