

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

Sabrina TOLLARI

Paris, le 4 octobre 2007

1

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

2

Motivation et problématique

Systèmes de recherche d'images sur le Web

The screenshot shows a Google Images search interface. The search query is "house water man filetype:jpg -lukes". The results page displays a grid of image thumbnails with their respective titles and source URLs. The first row includes "Water Gap House.jpg", "Clear Window Plants HouseB...", "indrxvsh.jpg", and "tsunami01.jpg". The second row includes "snow_1.jpg", "Bevan5.jpg", "LoyalWedellBulldozerTeps...", and "11683.jpg".

Motivation et problématique

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 !

4

Motivation et problématique

Indexation visuelle et fossé sémantique



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).

5

Motivation et problématique

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

6

Méthodologie et corpus

7

- 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
- 8

- Méthodologie et corpus
- ## 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.
- 9

Méthodologie et corpus

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

10

- Méthodologie et corpus
- ## 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
- 11

Méthodologie et corpus

Le score normalisé (NS)

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

wrong : nombre d'éléments non-pertinents retrouvés
 right : nombre d'éléments pertinents retrouvés
 N éléments
 n éléments pertinents

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

12

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
- Perspectives
- Résumé des résultats de la tâche2 de ImagEVAL

13

DIMATEX

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 2ⁿ clusters
 - Chaque vecteur visuel de l'ensemble d'apprentissage est codé en une séquence de bits de longueur n

14

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

15

DIMATEX

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

16

DIMATEX

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 :

$$P(w, C_k, A) = \sum_{J \in A} P(J, A) \sum_{b \in J} P(w|C_k, b, J, A) P(C_k|b, J, A) P(b|J, A)$$
 - 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 $P(b|J) = \frac{nb(w|b)}{\sum_{b \in J} nb(w|b)}$
 - P(C_k|b, J, A) = 1 si C_k = C(b), 0 sinon

17

DIMATEX

Associer des mots à une image

- Associer des mots à un blob :

$$P(w|b, A) = P(w|C(b), A) = \frac{P(w, C(b), A)}{P(C(b), A)}$$
- Associer des mots à une image :

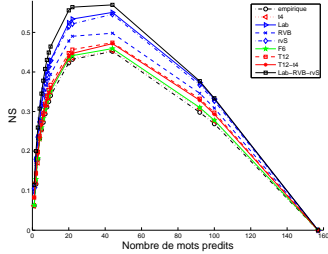
$$P(w|I, A) = \sum_{b \in I} P(w|b, I, A) P(b|I, A)$$

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

18

Expérimentations

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



19



- 1: forest rabbit snow woodland
- 2: water sky tree snow
- 3: sky snow tree water
- 4: plane water jet snow
- 5: birds water fox rock
- 6: birds water fox rock



- 1: cat grass lion mane
- 2: water sky tree people
- 3: cat grass lion tree
- 4: tree rock lizard walls
- 5: closeup insect tree grass
- 6: grass ground shadows closeup

- 1: flower leaf plants
- 2: tree water flower
- 3: buildings people water
- 4: tree water street
- 5: birds rock nest
- 6: face ruins sculpture

- 1: fish reef water
- 2: water tree fish
- 3: people tree water
- 4: tree water buildings
- 5: flowers gardens shop
- 6: birds ground owl

- 1: sky sunset town
- 2: tree sky water
- 3: sky tree water
- 4: pattern people clouds
- 5: building church hills
- 6: birds grass nest rock

- 1: people tables tree
- 2: people flower tree
- 3: grass lion people tree
- 4: tree buildings lizard temple
- 5: birds grass nest rock
- 6: birds grass nest ground

- 1: boat horizon water
- 2: sky water people
- 3: mountains sky water
- 4: mountains water sky
- 5: dock sailboat building
- 6: dock sailboat closeup

- 1: dune grass sand valley
- 2: water sky people snow
- 3: people sky tree
- 4: people snow water
- 5: animal rocks people
- 6: building fence snow

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

20

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%

21

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.

22

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.

23

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

24

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

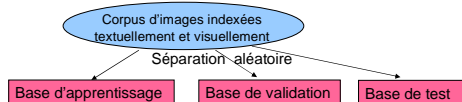
27

Construction de clusters visuels par CAH

Corpus d'images indexées
textuellement et visuellement

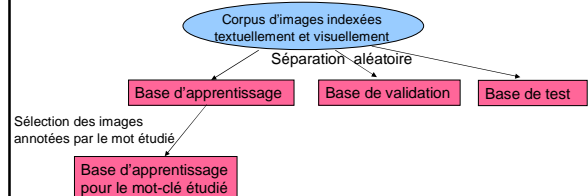
28

Construction de clusters visuels par CAH

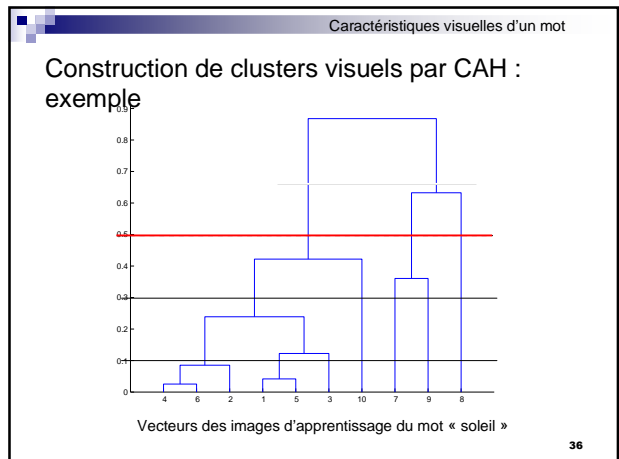
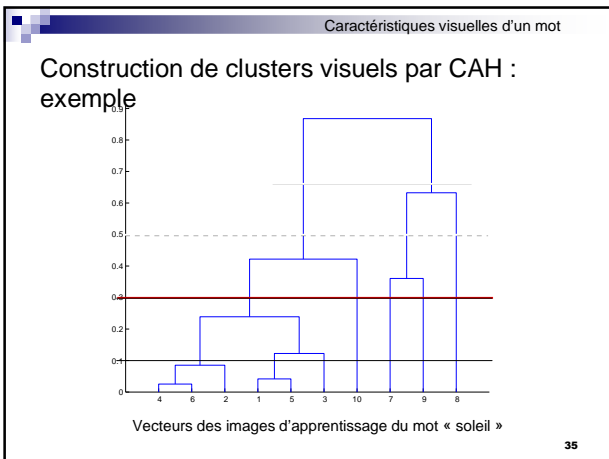
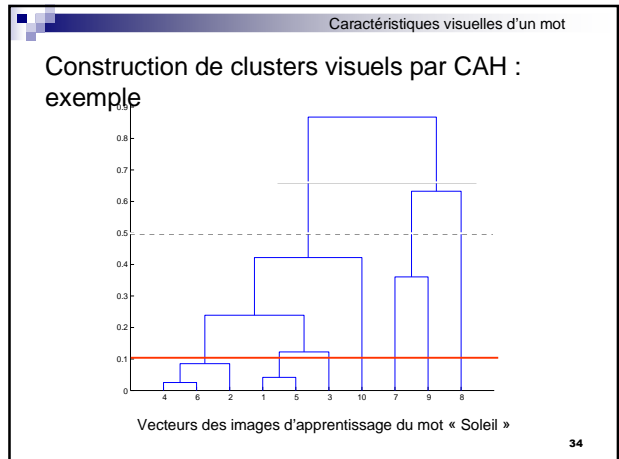
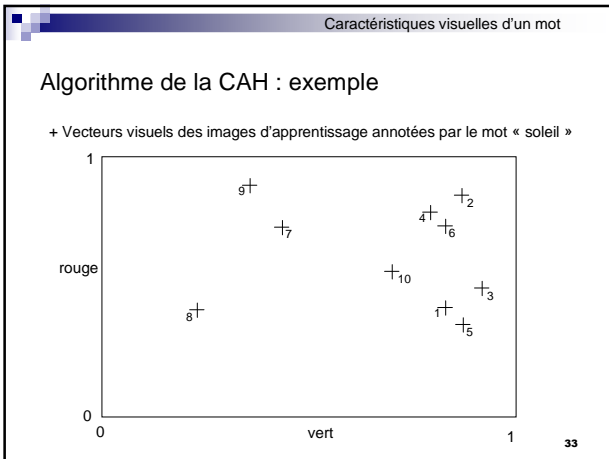
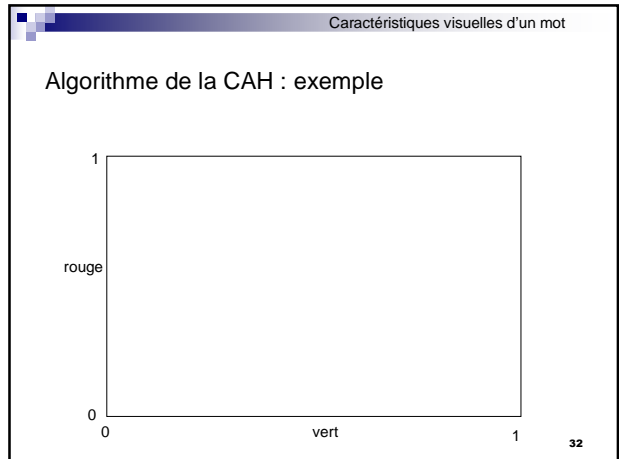
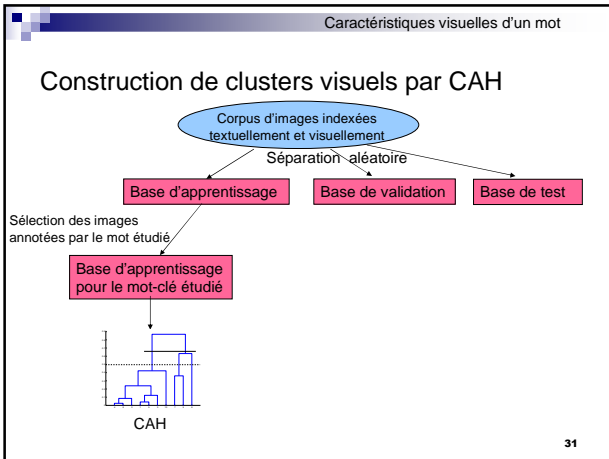


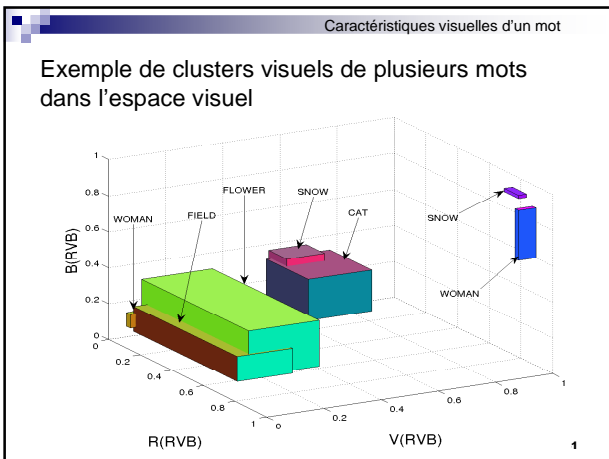
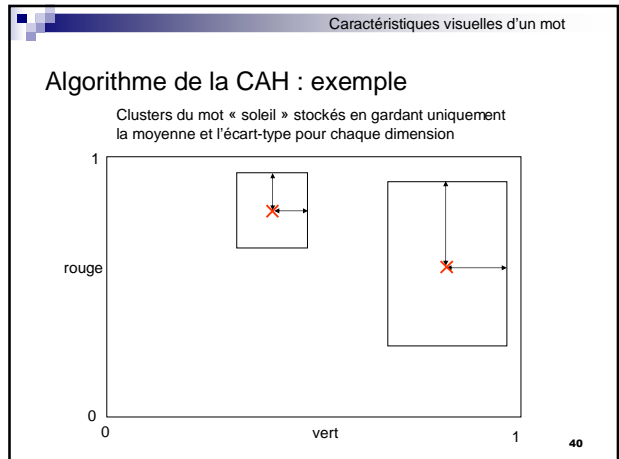
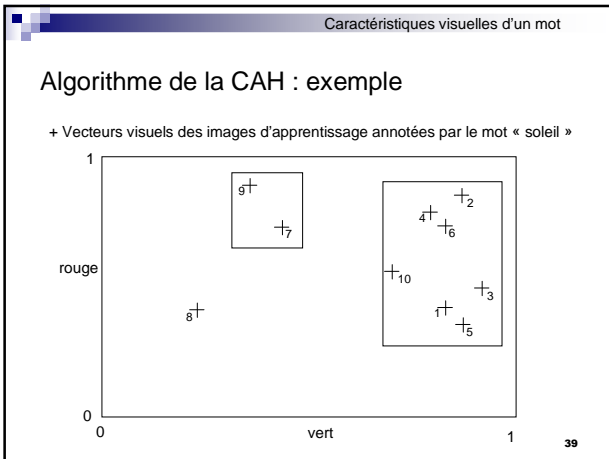
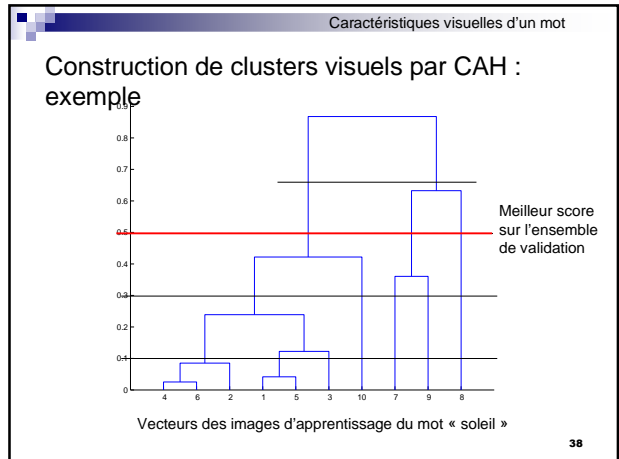
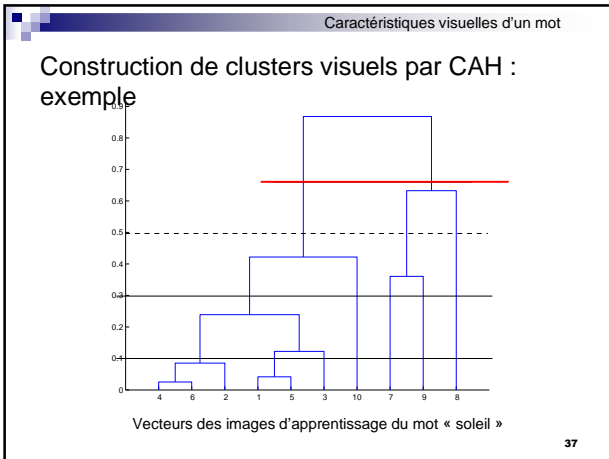
29

Construction de clusters visuels par CAH



30



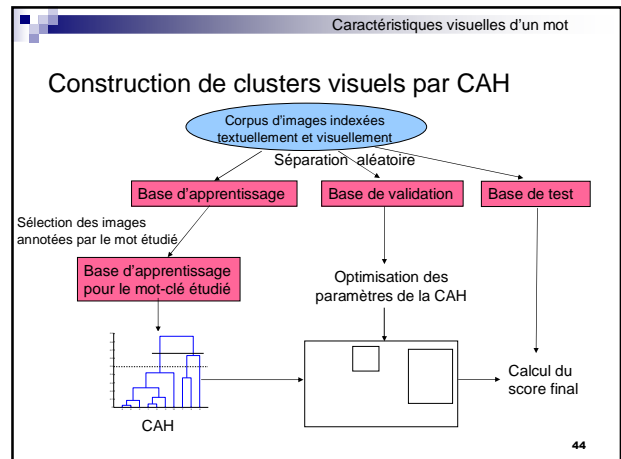
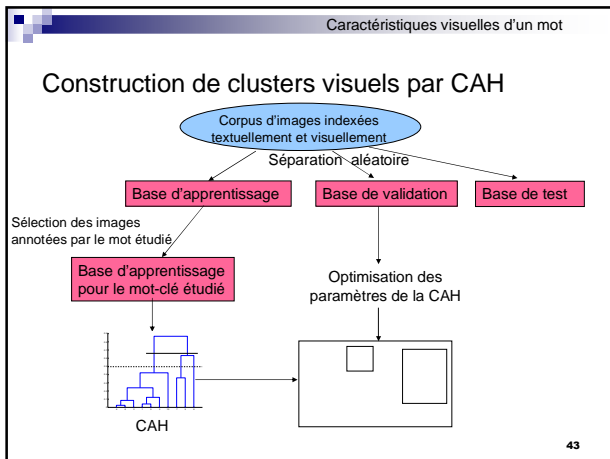


Caractéristiques visuelles d'un mot

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

42



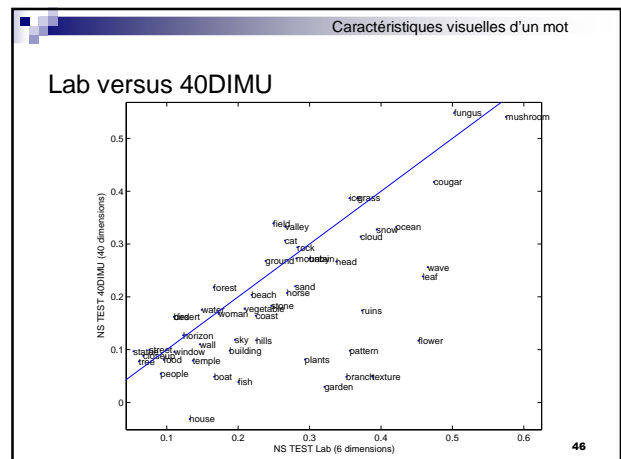
Caractéristiques visuelles d'un mot

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

45



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