

Annotation, indexation et recherche d'images par le texte et le contenu visuel

Sabrina Tollari,
 Université Pierre et Marie CURIE – Paris 6
 Laboratoire LIP6
sabrina.tollari@lip6.fr
 Nancy, le 26 juin 2009

Partie 1 / 3

1

ANR-06-MDCA-012

Plan

- **Problématique :**
 - améliorer la recherche d'images en utilisant le texte associé à l'image en combinaison avec le visuel
- **Problèmes à prendre en compte :**
 - fossé sémantique, passage à l'échelle, aspect « en ligne »
- **Méthodes proposées :**
 - Un modèle rapide d'annotation automatique d'images
 - Un système de recherche d'images combinant textes et images amélioré par sélection de la dimension
 - L'utilisation de concepts visuels pour améliorer la recherche d'images

2

Motivation : exemple de recherche d'images par le texte



3




Motivation

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
 - Segmentation, localisation, points d'intérêt
 - Décrit le contenu visuel de l'image, mais extraction de la sémantique difficile !
- **Les deux informations sont complémentaires**

4

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)

5

Autres difficultés

- **Le passage à l'échelle**
 - Malédiction de la dimension :
 - Les espaces de grande dimension possèdent des propriétés particulières qui font que les intuitions géométriques peuvent se révéler fausses.
 - La recherche par similarité visuelle et l'apprentissage sont donc plus difficiles et moins efficaces sur des espaces de grande dimension
 - Grand nombre de données
 - Problème de stockage des matrices de distances entre images
 - Recherche des k images les plus proches difficiles
 - Inversion de matrices très longues...
 - ...
- **L'aspect « en ligne » de la recherche d'images**
 - L'utilisateur ne veut pas attendre pour obtenir le résultats de sa requête, l'extraction des descripteurs visuels et les calculs nécessaires doivent être réalisés en un temps raisonnable
- => Les méthodes proposées doivent être efficaces, mais aussi efficientes

6

Plan

- **Problématique :**
 - améliorer la recherche d'images en utilisant le texte associé à l'image en combinaison avec le visuel
- **Problèmes à prendre en compte :**
 - fossé sémantique, passage à l'échelle, aspect « en ligne »
- **Méthodes proposées :**
 - Un modèle rapide d'annotation automatique d'images
 - Un système de recherche d'images combinant textes et images amélioré par sélection de la dimension
 - L'utilisation de concepts visuels pour améliorer la recherche d'images

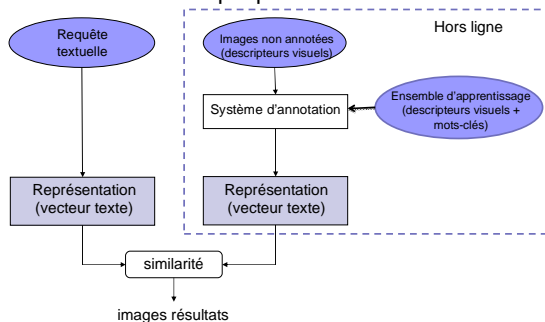
7

Annotation automatique d'images

- **Définition :** l'annotation automatique consiste à associer un groupe de mots à une image au moyen d'un système informatique
- **On distingue :**
 - L'annotation à partir du texte associé à l'image (mêmes méthodes que pour les documents textuels)
 - L'annotation à partir du contenu visuel de cette image
 - Utile quand il n'y a pas de texte associé à l'image
 - Utile pour vérifier la pertinence des mots par rapport au contenu visuel de l'image
- Les systèmes d'annotation automatique d'images par le contenu visuel peuvent être vus comme des sous-modules d'un système de RI dont le but est d'annoter les images avec du texte cohérent par rapport au contenu visuel

8

Moteur de recherche d'images utilisant un système d'annotation automatique par le contenu visuel



9

Annotation automatique d'images à partir du contenu visuel

- **Principe :**
 - D'abord, le système « apprend » à annoter des images à partir d'exemples déjà annotés
 - Puis, il est capable d'annoter une nouvelle image dont on ne connaît que les descripteurs visuels
- Pour les modèles probabilistes, l'annotation automatique consiste à estimer la probabilité *a posteriori* :
 - $P(w|I)$ où I représente l'information connue sur l'image (par exemple, l'ensemble des vecteurs visuels de l'image)
 - Si l'image est segmentée, une première étape peut être d'estimer la probabilité *a posteriori* :
 - $P(w|b)$ où b représente l'information connue sur une région d'image (par exemple, le vecteur décrivant le contenu visuel de la région d'image)

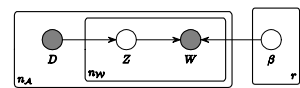
10

Annotation automatique d'images à partir du contenu visuel

- **Grand nombre de modèles :**
 - Modèles basés sur l'analyse de la sémantique latente (LSA, PLSA...)
 - Modèles basés sur la distribution de Dirichlet (MoM-LDA, Corr-LDA...)
 - Modèles de traduction de langues...
- **Différentes façons de combiner informations textuelles et visuelles pour l'annotation**
 - Fusion précoce des espaces textuel et visuel :
 - LSA, PLSA...
 - Combinaisons indépendantes des espaces :
 - MoM-LDA, GM-LDA, MoM-HAM I-2...
 - Combinaisons dépendantes des espaces :
 - Corr-LDA, MoM-HAM D-2...
 - Combinaisons à différents niveaux d'une hiérarchie :
 - MoM-HAM, MoM-LDA, Mix-Hier...

11

Modèle LSA et PLSA



- **Rappel sur LSA :**
 - Matrice termes-documents, décomposition en valeur singulière
 - Le sens d'un mot est défini par rapport à son contexte
 - Deux mots sont similaires s'ils apparaissent dans le même contexte
- **Dans (Monay et al., 2003) :**
 - Une image est représentée par un vecteur concaténant 149 dimensions pour le texte et 648 dimensions (espace RVB) pour le visuel
 - LSA donne de meilleurs résultats que PLSA ! Peut-être à cause du trop grand nombre de dimensions visuelles
- **Dans (Monay et al., 2004) :**
 - Un espace latent est construit pour chaque modalité
 - 1. Les probabilités $p(w|z)$ et $p(z|d)$ sont apprises sur les mots-clés
 - 2. Un autre modèle PLSA est appris sur le visuel $p(v|z)$, mais en gardant la probabilité $p(z|d)$ apprises sur les mots-clés
 - Ce modèle PLSA donne de meilleurs résultats que LSA ou que PLSA avec $p(z|d)$ appris indépendamment

12

Associer des mots à une image

- Associer des mots à une région d'images :

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

- Associer des mots à une image :

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

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

- Ce modèle :
 - ne nécessite aucun apprentissage
 - ne possède aucun paramètre à optimiser
- Permet une annotation très rapide grâce à la binarisation

Comparer les modèles d'annotation automatique

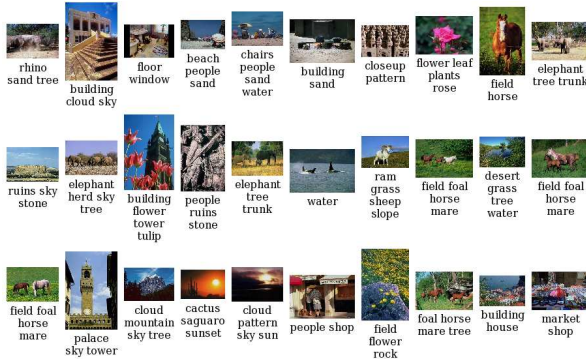


- Corpus COREL :

- 10 000 images <http://wang.ist.psu.edu>
- Chaque image annotée par de 1 à 5 mots parmi 250
- Segmentée en 10 régions maximum
- Chaque région (appelée « blob ») est décrite par un vecteur de 40 composantes visuelles

- 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.

COREL



1. Annotation manuelle 2. DIMATEX 3. PLSA-WORDS

4. PLSA-WORDSFEATURES 5. DIRECT 6. LSA (Monay & Gatica-Perez, 2004)

Base d'images COREL : <http://wang.ist.psu.edu>

Comparaison des systèmes 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%

Evolution des scores de rappels et de précision des modèles de l'état de l'art

Référence	Principes	mP	mR	n _{wf0}
Duygulu et al., 2002	Modèle de traduction	0.04	0.06	49
Jeon et al., 2003	Cross-Media Relevance Models (CMRM)	0.10	0.09	66
Lavrenko et al., 2003	Continuous Relevance Models (CRM) (distribution multinomiale)	0.19	0.16	107
Feng et al., 2004	Distribution multiple de Bernoulli	0.25	0.24	122
Cameiro et Vasconcelos, 2005	Apprentissage supervisée (estimation de densité)	0.29	0.23	137
Gao et al., 2006	Apprentissage multi-classes (maximal figure-of-merit)	0.27	0.25	133
Liu et al., 2007	Dual Cross-Media Relevance Model (DCMRM)	0.28	0.23	135

mP : précision moyenne, mR : rappel moyen, n_{wf0} : nombre de mots prédits

Dual Cross-Media Relevance Model (Liu et al., 2007)

- Modèle traditionnel :

$$w^* = \operatorname{argmax}_w P(w|I) = \operatorname{argmax}_w \sum_j P(j|P)P(w|j)P(I|j)$$

- Modèle DCMRM :

$$w^* = \operatorname{argmax}_w P(w|I) = \operatorname{argmax}_w \sum_w P(w')P(w|w')P(I|w')$$

- $P(w')$ indique l'importance du mot w'
- $P(w|w')$ représente la relation sémantique entre deux mots w et w' (exemple : relation de WordNet)
- $P(I|w')$ modélise comment l'image I est pertinente pour le mot w' (exemple : probabilité de retrouver une image I lors de la requête textuelle w' dans un moteur de recherche d'images textuels)
- => pas d'ensemble d'apprentissage

25

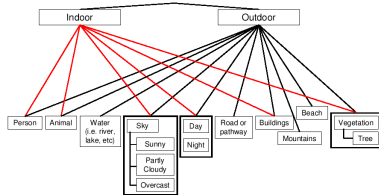
Lacunes des systèmes d'annotation d'images

- En pratique :
 - De bons résultats sont obtenus pour des mots « visuels » (extérieur, intérieur, arbre, mer, portrait, couché de soleil...),
 - Mais pas sur des mots plus généraux (hôtel de ville, table, peuplier, mer méditerranée, Peter Falk...)
- Les systèmes d'annotations sont globalement de plus en plus efficaces
- Mais peu de systèmes sont construits pour prendre en compte :
 - le temps de calcul, le nombre de paramètres, la complexité du modèle
- Nouveaux corpus d'annotation automatique :
 - Visual Concept Detection ImageCLEF 2008 :
 - Image Annotation ImageCLEF 2009

Rappel : COREL 10 000 images, 200 mots

26

ImageCLEF2008 : Visual Concept Detection Task (VCDT)



- 17 classes en partie hiérarchisées
- 2k images d'apprentissage
- 1k images de test
- => Problème multiclassées avec classes imbriquées

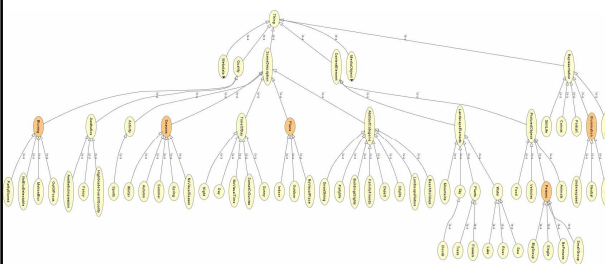
27

ImageCLEF2009 : Large Scale Visual Concept Detection and Annotation Task

- Corpus :
 - MIRFLICKR-25000 Image Collection
 - 5000 images pour l'apprentissage
 - 13000 images pour le test
- Challenges :
 - Est-ce qu'un classifieur d'images peut passer à l'échelle en nombre de concepts et de données ?
 - Est-ce qu'une ontologie (hiérarchie et relations) aide pour l'annotation à l'échelle ?

28

Hiérarchie des concepts de la tâche ImageCLEF2009 Annotation



29

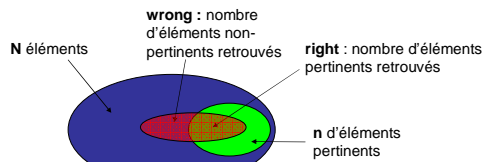
Bibliographie

- Barnard, K., Duygulu, P., de Freitas, N., Forsyth, D., Blei, D., & Jordan, M. I., Matching Words and Pictures, *Journal of Machine Learning Research*, 3, 1107-1135.
- Blei, D. M., & Jordan, M., Modeling Annotated Data, *ACM SIGIR*, 2003
- Monay, F., & Gatica-Perez, D., On image auto-annotation with latent space models, *ACM Multimedia*, 2003
- Monay, F., & Gatica-Perez, D., PLSA-based image auto-annotation: constraining the latent space, *ACM Multimedia*, 2004
- Hervé Glotin, Sabrina Tollari, "Fast Image Auto-annotation with Visual Vector Approximation Clusters", *Workshop on Content-Based Multimedia Indexing (CBMI)*, 2005
- Duygulu, P., Barnard, K., de Freitas, J. F. G., & Forsyth, D. A. 2002. Object Recognition as Machine Translation: Learning a Lexicon for a Fixed Image Vocabulary, Pages 97-112 of *ECCV*
- Jeon, J., Lavrenko, V., & Manmatha, R., 2003. Automatic image annotation and retrieval using cross-media relevance models. Pages 119-126 of: *ACM SIGIR*.
- Lavrenko, V., Manmatha, R., & Jeon, J. 2003. A Model for Learning the Semantics of Pictures. In: *Neural Information Processing Systems (NIPS)*.
- Feng, S. L., Manmatha, R., & Lavrenko, V. 2004. Multiple Bernoulli Relevance Models for Image and Video Annotation, Pages 1002-1009 of: *Proceedings of Computer Vision and Pattern Recognition (CVPR)*.
- Cameiro, G., & Vasconcelos, N. 2005. Formulating Semantic Image Annotation a Supervised Learning Problem, Pages 163-168 of: *IEEE Computer Vision and Pattern Recognition (CVPR)*
- Gao, S., Wang, D.-H., & Lee, C.-H. 2006 Automatic Image Annotation through Multi-Topic Text Categorization. In: *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*
- Liu et al., Dual cross-media relevance model for image annotation, *ACM Multimedia*, 2007

30

Le score normalisé (NS)

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



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

31

Sabrina Tollari,
Université Pierre et Marie CURIE – Paris 6
Laboratoire LIP6
sabrina.tollari@lip6.fr
Nancy, le 26 juin 2009

Partie 2 / 3

32

Plan

- Problématique :
 - améliorer la recherche d'images en utilisant le texte associé à l'image en combinaison avec le visuel
- Problèmes à prendre en compte :
 - fossé sémantique, passage à l'échelle, aspect « en ligne »
- Méthodes proposées :
 - Un modèle rapide d'annotation automatique d'images
 - Un système de recherche d'images combinant textes et images amélioré par sélection de la dimension
 - L'utilisation de concepts visuels pour améliorer la recherche d'images

33

Un système de recherche d'images combinant textes et images amélioré par sélection de la dimension

Sabrina TOLLARI* et Hervé GLOTIN**

* Université Pierre et Marie Curie-Paris6 / UMR CNRS 7606 LIP6

** Université du Sud Toulon-Var / UMR CNRS 6168 LSIS

sabrina.tollari@lip6.fr, glotin@univ-tln.fr

34

Plan

- Description de la tâche 2 de la campagne ImagEVAL
- Description du système de fusion visuo-textuelle
- Amélioration par sélection de la dimension visuelle
 - Utilisation de l'Approximation de l'Analyse Linéaire Discriminante (ALDA)
- Expérimentations sur le corpus d'ImagEVAL
 - Résultats officiels de la tâche 2 d'ImagEVAL
 - Résultats généraux sur le modèle de fusion
 - Amélioration par sélection de la dimension
- Conclusion

35

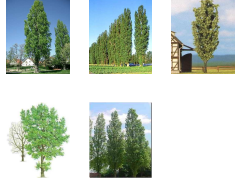
Motivation :
exemple de recherche textuelle

36

Motivation :
utilisation de requêtes visuo-textuelles

- Pour exprimer son besoin d'information, l'utilisateur peut compléter sa requête en utilisant des images qui indiquent visuellement ce qu'il attend
- Exemple :

« peuplier l'arbre » +



requête 24 de la campagne ImagEVAL <http://www.imageval.org/> 37

ImagEVAL

Description de la tâche 2 d'ImagEVAL

- Corpus : 700 urls
 - 700 pages Web
 - 10k images Web
- 25 requêtes : chaque requête est composée de mots-clés et d'images
- But : trouver parmi les 10k images celles qui sont pertinentes pour chaque requête


- Pour le test officiel :
 - 300 images doivent être rendues
 - Les MAP (Mean Average Precision) sont calculés par le logiciel standard treceval
 - Les images pertinentes sont inconnues (entre 10 et 100 par requête)
 - Nous n'avons pas utilisé d'ensemble d'apprentissage/validation

38


ImagEVAL

Exemples de requêtes visuo-textuelles de la campagne ImagEVAL

« poisson clown » +



« tournevis » +



39

ImagEVAL

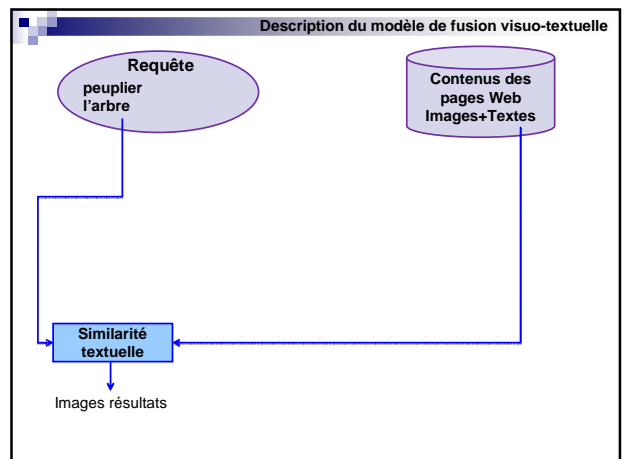
Description du modèle de fusion visuo-textuelle

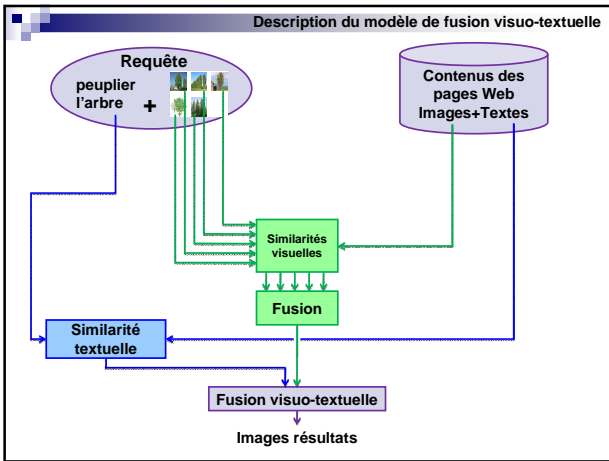
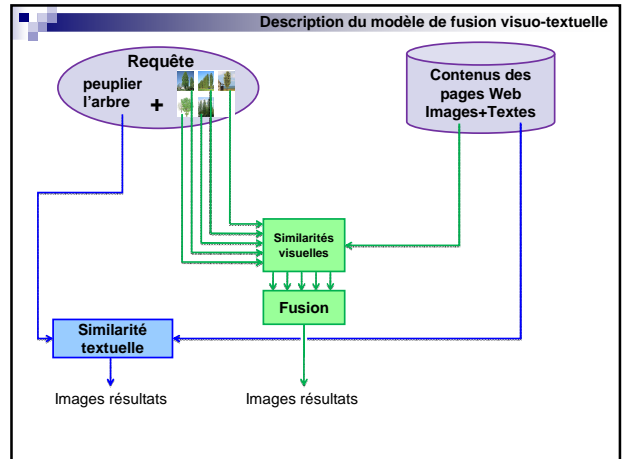
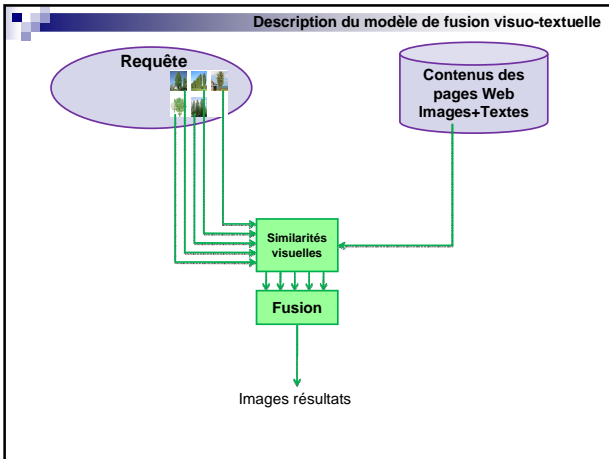
40

Choix du type de fusion

- Fusion précoce : On construit un espace contenant à la fois des descripteurs textuels et des descripteurs visuels
 - Avantage : Plus facile à intégrer dans un système de recherche d'information
 - Inconvénients :
 - Difficile de déterminer combien de dimensions textuelles et visuelles doivent être utilisées pour obtenir de bons résultats
 - Utilise des espaces à grande dimension
- Fusion tardive :
 - Une recherche est effectuée en utilisant uniquement le texte de la requête et les descriptions textuelles
 - Une autre recherche est effectuée en utilisant uniquement les images
 - Une fusion tardive des valeurs ou des rangs permet d'obtenir un nouvel ordre de pertinence des résultats

41





Description du modèle de fusion visuo-textuelle

Fusion visuo-textuelle

Nous combinons le visuel et le texte en utilisant une moyenne pondérée des distances visuelles et textuelles :

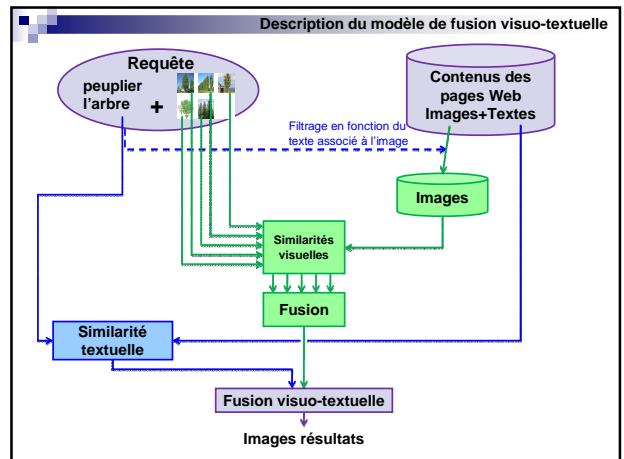
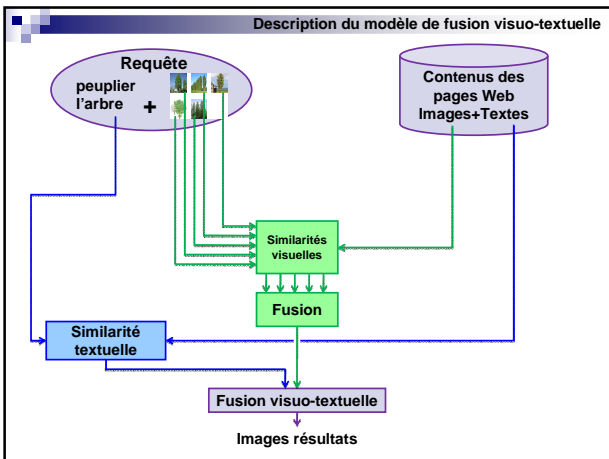
$$D(Q,I,t) = t \times D_T(Q,I) + (1 - t) \times D_V(Q,I)$$

où

- D_T est la distance normalisée basée sur tfidf,
- D_V est la distance visuelle entre les images de la requête Q et une image I et
- t représente le poids de texte dans la fusion.

Hypothèse : s'il n'y a aucun mot-clé de la requête dans la page Web de l'image alors les images de cette page Web ne sont pas prises en considération (filtrage textuelle)

46



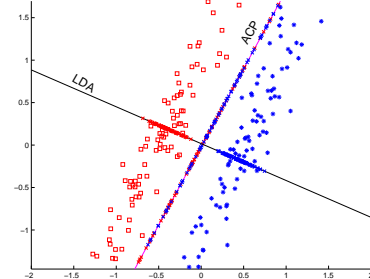
Amélioration par sélection de la dimension

- Problèmes :
 - Réaliser des distances visuelles est très coûteux en temps de calcul
 - Le problème de la malédiction de la dimension réduit la qualité des résultats
- Solution proposée :
 - Sélectionner les dimensions les plus discriminantes par ALDA
 - Problème : comment obtenir des données pour appliquer cette méthode supervisée ?

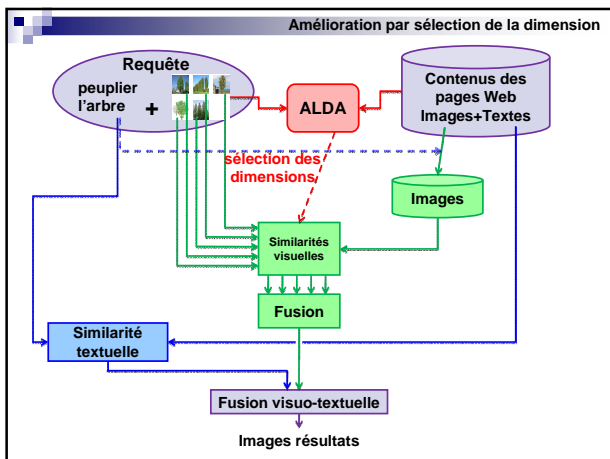
49

Rappel : LDA versus ACP

- L'ACP recherche l'axe qui représente le mieux les données
- La LDA recherche l'axe qui sépare le mieux les classes



50



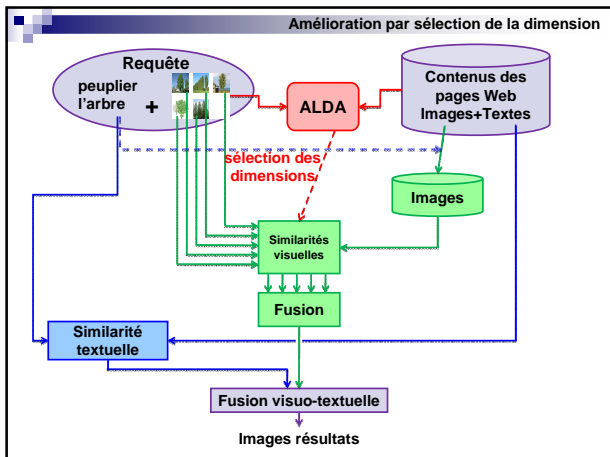
Sélection par Approximation de l'Analyse Linéaire Discriminante (ALDA)

- Construire un ensemble Ω d'images du web suffisamment grand
- Pour chaque requête Q
 - Construire l'ensemble $\Psi(Q)$ des images positives pour la requête
 - Pour chaque dimension X de l'espace visuel :
 - Calculer les variances interclasse $\hat{B}(X; Q)$ et intraclasse $\hat{W}(X; Q)$ entre les ensembles $\Psi(Q)$ et Ω
 - Calculer ensuite le pouvoir discriminant :

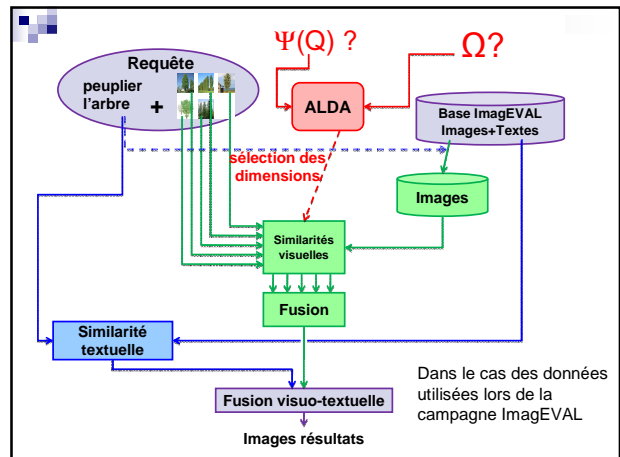
$$\hat{J}(X; Q) = \frac{\hat{B}(X; Q)}{\hat{B}(X; Q) + \hat{W}(X; Q)}$$

- Sélectionner les N dimensions qui ont le plus fort pouvoir discriminant
- Calculer les distances visuelles à partir des dimensions sélectionnées

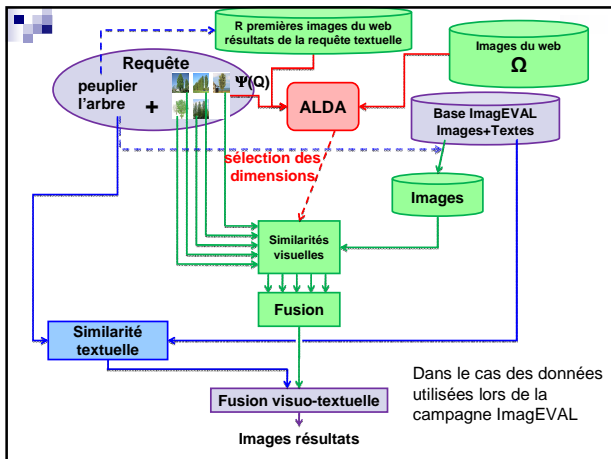
52



$\Psi(Q)$? Ω ?




Dans le cas des données utilisées lors de la campagne ImagEVAL



Expérimentations

Extraction des descripteurs



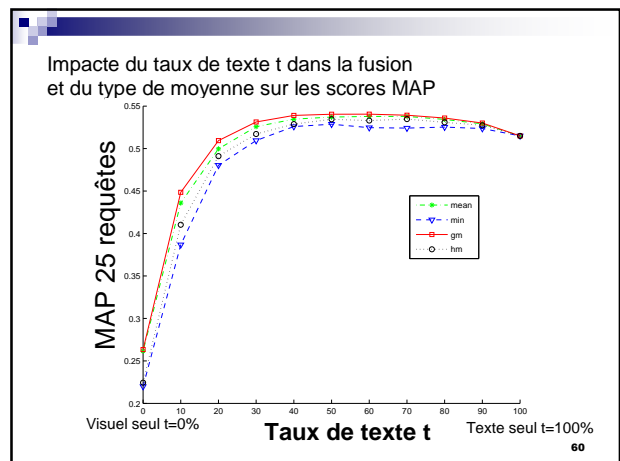
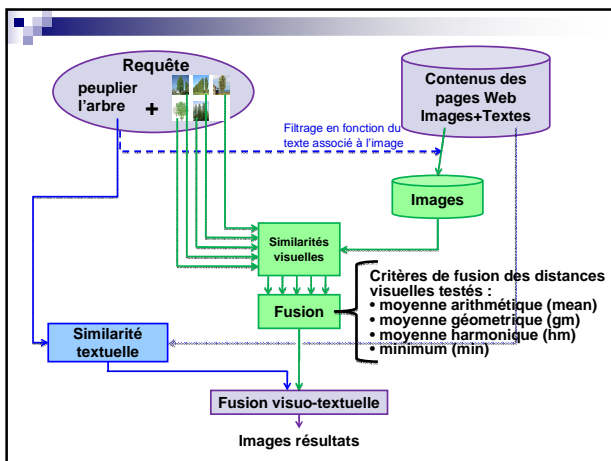
- Descripteurs visuels**
 - Les images sont segmentées en 3 bandes horizontales
 - Pour chaque bande et pour chaque couleur $L=R+G+B$, $r=R/L$, $g=G/L$
 - Calculer la moyenne et l'écart-type des valeurs des pixels
 - Calculer les profils horizontaux, verticaux et globaux
 - L'espace visuel est composé de $3 \times 3 \times 5 = 45$ dimensions visuelles
- Descripteurs textuels**
 - les balises HTML (<H1>, ...) sont supprimées, mais le contenu des balises (URL, nom de l'image...) est conservé
 - Le texte est normalisé (majuscules->minuscules, éêêê->e...)
 - Les caractères spéciaux et les « stops words » sont supprimés
- But :** avoir un traitement rapide des pages web

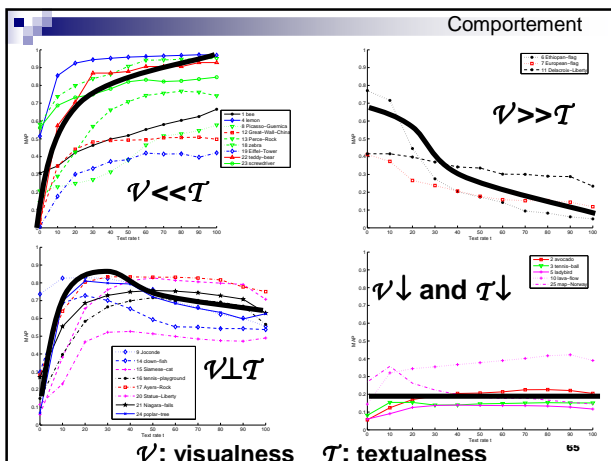
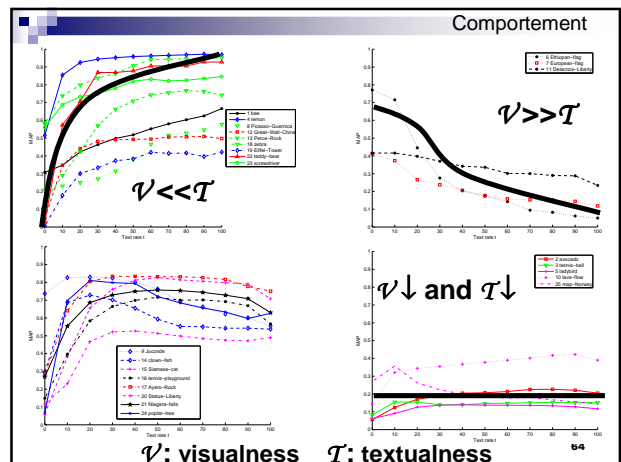
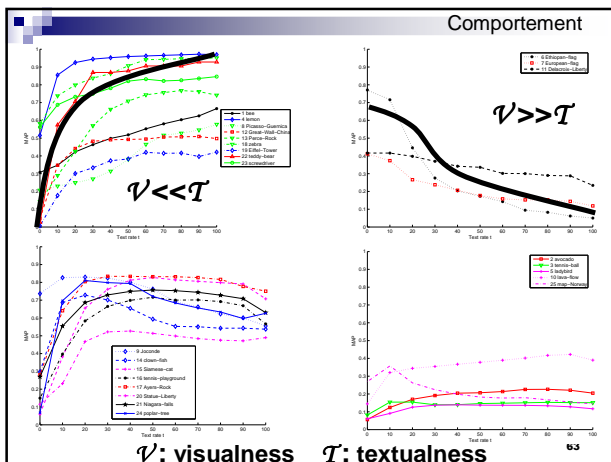
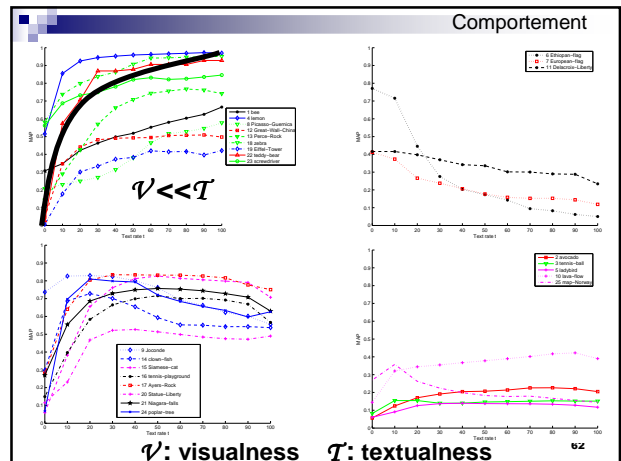
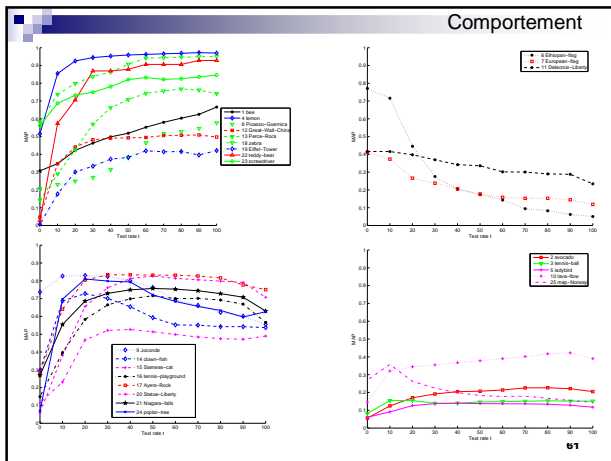
Résultats officiels de la campagne ImagEVAL

- 4 équipes
- 22 runs parmi lesquels 11 runs fusion, 7 runs texte seul et 5 runs visuel seul
- Nous avons proposé 3 runs et sommes arrivés 2nd / 4

Rang équipe	Texte seul	Visuel seul	Fusion visuo-textuelle
1	0.559 (Run 5)	0.271 (Run 16)	0.613 (Run 1)
2	0.513* (Run 9)	0.261* (Run 17)	0.536* (Run 7)
3	0.455 (Run 12)	0.181 (Run 20)	0.517 (Run 8)

* Score MAP LSIS. Pour la fusion visuo-textuelle t=50% ; le critère de fusion visuelle est la moyenne arithmétique ; pas de sélection de la dimension ! (les runs sont ordonnés du plus fort (run1) au plus faible MAP (run22))





Exemple de requête $\mathcal{V} \mathcal{T}$

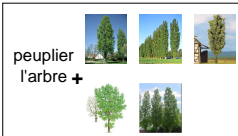
Requête: peuplier l'arbre +

Images pertinentes que nous souhaitons retrouver

Exemple de requête \mathcal{VLT}

Requête

peuplier
l'arbre +




67

Exemple de requête \mathcal{VLT}

Requête

peuplier
l'arbre +



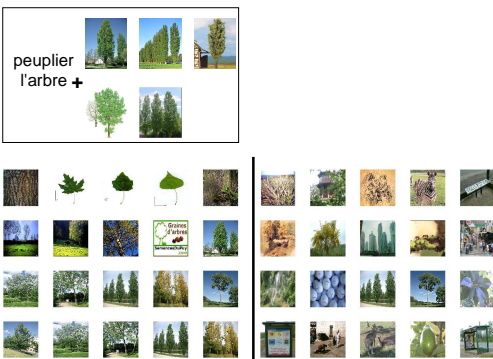
Texte seul (MAP=0.63)

68

Exemple de requête \mathcal{VLT}

Requête

peuplier
l'arbre +



Texte seul (MAP=0.63)

Visuel seul (MAP=0.07)

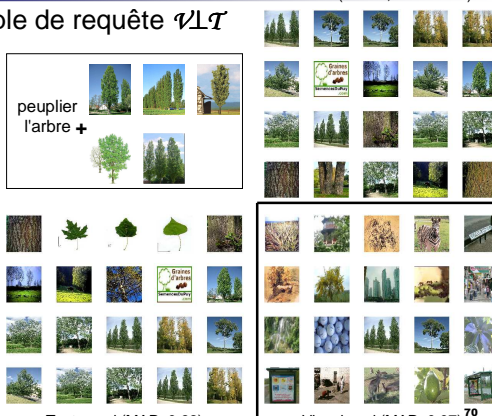
69

Exemple de requête \mathcal{VLT}

Fusion (t=20%, MAP=0.81)

Requête

peuplier
l'arbre +



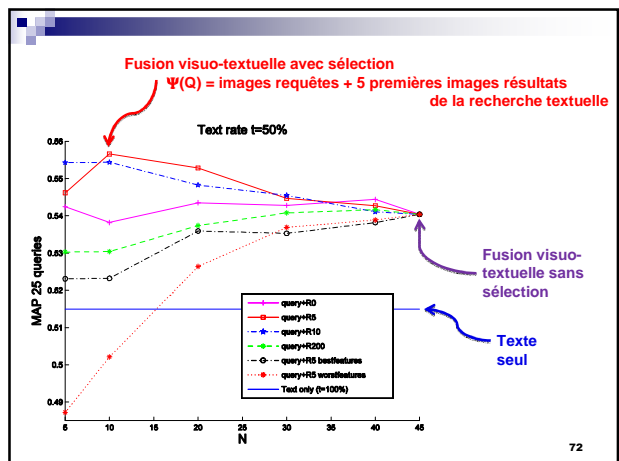
Texte seul (MAP=0.63)

Visuel seul (MAP=0.07)

70

Amélioration par sélection de la dimension

71



Résultats avec sélection de la dimension

	Nombre de dimensions visuelles	Score MAP	Temps de calcul des distances visuelles en secondes
Texte seul	-	0.513	-
Visuel seul	45	0.261	309
Fusion	45	0.539	309
Fusion	10	0.557	202

$\Psi(Q)$ = « Query+R5 » (images requêtes + 5 premières images résultats) ; Ω composé de 17k images ; taux de fusion $t=50\%$; le critère de fusion visuelle est la moyenne géométrique ; temps pour 25 requêtes (distance entre 131 vecteurs visuels des images requêtes et 100k vecteurs visuels synthétiques)

73

Conclusion

- La fusion visuo-textuelle permet d'améliorer la recherche d'images au moins pour certaines requêtes
- Le comportement des courbes de fusions visuo-textuelles dépend de la nature intrinsèque des requêtes
- La sélection de la dimension permet d'améliorer les scores tout en réduisant les temps de calcul
- Certaines techniques supervisées qui ne peuvent généralement pas être appliquées en RI à cause de l'absence de données d'apprentissage bien étiquetées peuvent être utilisées dans le cas de données mal étiquetées
- Notre système peut être utilisé dans le cas de requêtes « en ligne » (temps estimé par requête sur la base ImagEVAL ≈ 1 seconde sans compter le temps d'extraction des descripteurs visuels 0.17s/image)

74

Sabrina Tollari,
Université Pierre et Marie CURIE – Paris 6
Laboratoire LIP6
sabrina.tollari@lip6.fr
Nancy, le 26 juin 2009

Partie 3 / 3

75

Plan

- Problématique :
 - améliorer la recherche d'images en utilisant le texte associé à l'image en combinaison avec le visuel
- Problèmes à prendre en compte :
 - fossé sémantique, passage à l'échelle, aspect « en ligne »
- Méthodes proposées :
 - Un modèle rapide d'annotation automatique d'images
 - Un système de recherche d'images combinant textes et images amélioré par sélection de la dimension
 - L'utilisation de concepts visuels pour améliorer la recherche d'images

76

Utilisation de concepts visuels pour améliorer la recherche d'images

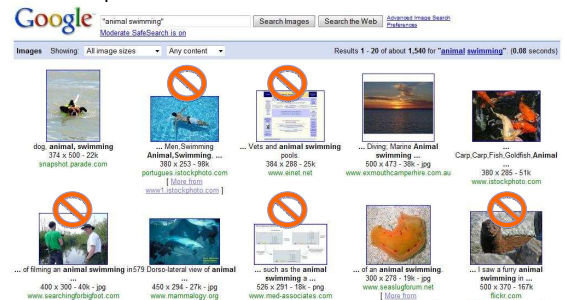
Sabrina Tollari, Marcin Detyniecki, Ali Fakeri Tabrizi,
Christophe Marsala, Massih-Reza Amini, Patrick Gallinari

Université Pierre et Marie Curie – Paris 6
UMR CNRS 7606 - LIP6

77

Problématique :

améliorer la recherche d'images basée sur le texte en utilisant des concepts visuels



78

VCDT

Détection de concepts visuels dans des images généralistes

- Proposition :
 - Apprendre des forêts d'arbres de décision flous pour détecter les concepts visuels
 - Pour chaque image et pour chaque concept, on obtient un score indiquant la présence du concept dans l'image
 - Analyser les cooccurrences des concepts pour déterminer les relations entre les concepts
 - Découvertes de relations d'exclusion et d'implication entre concepts à partir de la matrice de cooccurrences
 - Utiliser ces relations pour améliorer la détection des concepts
 - Applications de règles pour filtrer les scores des arbres

79

VCDT

Tâche « Visual Concept Detection Task » (VCDT) de ImageCLEF 2008

outdoor person day vegetation animal

concepts ? 80

- 17 concepts visuels en partie hiérarchisés
- 2k images d'apprentissage :
 - chaque image d'apprentissage est associée à plusieurs concepts
- 1k images de test
 - ⇒ Problème multi-classes multi-étiquettes avec classes hiérarchisées
 - ⇒ Annotation automatique d'images

VCDT

Forêt d'arbres de décision flou

- Pour chaque concept :
 - une forêt de 50 arbres flous* (critère entropique flou et agrégation par somme) est apprise à partir des images d'apprentissage
 - Un seuil t est fixé pour prendre la décision d'annoter ou pas une image de test par ce concept

* Logiciel Salammbô, Christophe Marsala, Apprentissage inductif en présence de données imprécises : Construction et utilisation d'arbres de décision flous, thèse de doctorat de l'université Paris 6, 1998

81

VCDT

Comment découvrir les relations ?

- Les arbres de décisions apprennent chacun des concepts indépendamment, or les concepts sont dépendants entre eux
- Relation d'exclusion :
 - A partir de la matrice de cooccurrence, nous déterminons les concepts qui ne sont jamais co-occurents
- Relation d'implication (ou de nécessité) :
 - Nous construisons une matrice : présence de A versus absence de B, pour tous couples de concepts (A,B)
 - Quand dans la matrice, on n'a jamais : A et non(B) alors A implique B
 - car $\text{non}(A \text{ et non}(B)) \rightarrow \text{non}(A) \text{ ou } B \rightarrow A \Rightarrow B$
 - Exemple : on n'a jamais *Tree* et non(*Vegetation*)

82

VCDT

Relations découvertes automatiquement

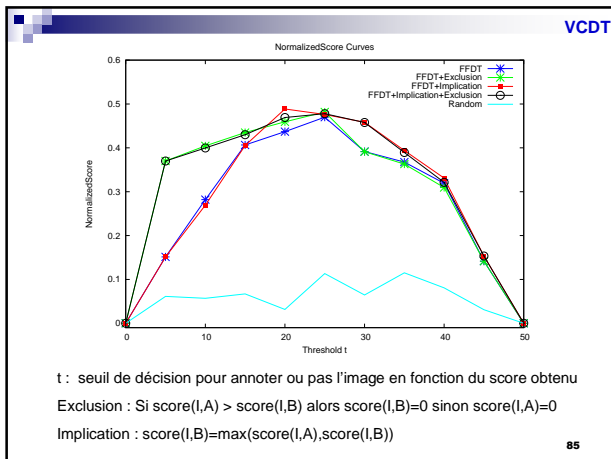
83

VCDT

Comment utiliser ces relations ?

- Exclusion :
 - $\text{COOC}(A,B) \neq 0$
 - Si $\text{score}(I,A) > \text{score}(I,B)$ alors $\text{score}(I,B) = 0$ sinon $\text{score}(I,A) = 0$
 - Exemple :
 - Avant : $\text{outdoor} = 0.8$ et $\text{indoor} = 0.4$
 - Après : $\text{outdoor} = 0.8$ et $\text{indoor} = 0$
- Implication :
 - $\text{COOC}(A, \text{non } B) \neq 0$
 - $\text{score}(I,B) = \max(\text{score}(I,A), \text{score}(I,B))$
 - Exemple :
 - Avant : $\text{cloudy} = 0.8$ et $\text{sky} = 0.4$
 - Après : $\text{cloudy} = 0.8$ et $\text{sky} = 0.8$

84



ImageCLEFphoto

Utilisation des concepts visuels pour améliorer la recherche d'images

- Comment utiliser les résultats de détection de concepts visuels pour améliorer la recherche textuelle d'images ?

Requête : "animal swimming"

- Proposition :
 - Filtrer à l'aide des scores des arbres de décision les résultats obtenus par la recherche textuelle en fonction de la présence des concepts dans :
 - Les mots de la requête (filtrage directe)
 - Les mots de la requête élargis à l'aide de la relation de synonymie de WordNet (filtrage WN)

86

ImageCLEFphoto

Exemple de filtrage directe

Requête : "seals near water"

Filtrage en fonction des scores des arbres de décision du concept « water »

87

ImageCLEFphoto

Exemple de filtrage en utilisant WordNet

- Requête : "animal swimming"
 - Animal: organism, plankton, mascot, fungus, ...
 - Swimming: bathe, diving, floating, surfing, water sport, ...
- Filtrage en fonction des scores des arbres de décision du concept « animal » et du concept « water »

88

ImageCLEFphoto

Résultats obtenus lors de ImageCLEFphoto 2008

- ImageCLEFphoto 2008 : 20 000 images avec du texte associé, 39 requêtes multimédia
- Par filtrage directe, 11 requêtes modifiées et 7 concepts utilisés
- Par le filtrage utilisant WordNet, 25 requêtes modifiées et 9 concepts utilisés

Texte	Méthode de filtrage		Les 39 requêtes		Les 11 requêtes filtrées	
	Directe	WN	P20 (gain %)	Nombre de requêtes	P20 (gain %)	
Modèle de langues			0.185 (-)	11	0.041 (-)	
	X		0.195(+6)	11	0.077 (+88)	
	X	X	0.176(-5)	25	0.134 (-9)	
TF-IDF			0.250 (-)	11	0.155 (-)	
	X		0.269(+8)	11	0.223 (+44)	
	X	X	0.258(+4)	25	0.226 (+8)	

89

ImageCLEFphoto

Conclusion

- Les méthodes d'annotation automatique d'images uniquement à partir du contenu visuel permettent d'obtenir de bons résultats seulement dans le cas où les « mots » sont des concepts visuels
 - Exemple : intérieur, extérieur, arbre, mer, portrait...
- Les méthodes de recherche d'images fusionnant tardivement le texte et les images nécessite de trouver les bons poids entre les deux modalités pour chaque requête ce qui n'est pas évident
- Utiliser les concepts visuels pour améliorer la recherche d'image est une piste intéressante, mais la difficulté est de trouver la correspondance entre la requête textuelle et les concepts visuels

90



MERCI DE VOTRE ATTENTION

QUESTIONS ?