Recherche visuo-textuelle d'images sur le Web améliorée par sélection de la dimension

Sabrina TOLLARI* et Hervé GLOTIN**

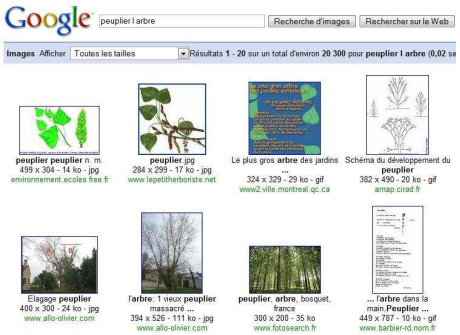
* Université Pierre et Marie Curie-Paris6 / UMR CNRS 7606 LIP6
** Université du Sud Toulon-Var / UMR CNRS 6168 LSIS
sabrina.tollari@lip6.fr, glotin@univ-tln.fr

Trégastel, le 12 mars 2008, CORIA2008

Plan

- Description de la tâche 2 de la campagne ImagEVAL
- Description du modèle 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

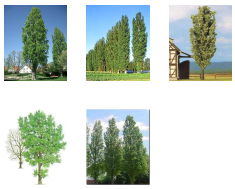
Motivation : exemple de recherche textuelle



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/

Description de la tâche 2 d'ImagEVAL

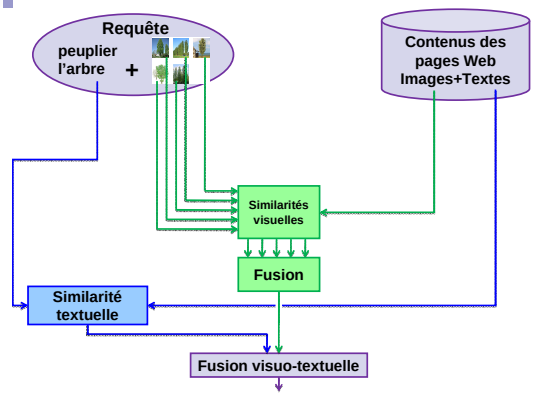
- Corpus : 700 urls
  - 700 pages Web
  - 10k images Web
- 25 requêtes : chaque requête est composée de mots-clés et d'images
- But : trouver parmi les 10k images celles qui sont pertinentes pour chaque requête
- Pour le test officiel :
  - 300 images doivent être rendues
  - Les MAP (Mean Average Precision) 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

Exemples de requêtes visuo-textuelles de la campagne ImagEVAL

« poisson clown » +

« tournevis » +

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 :

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

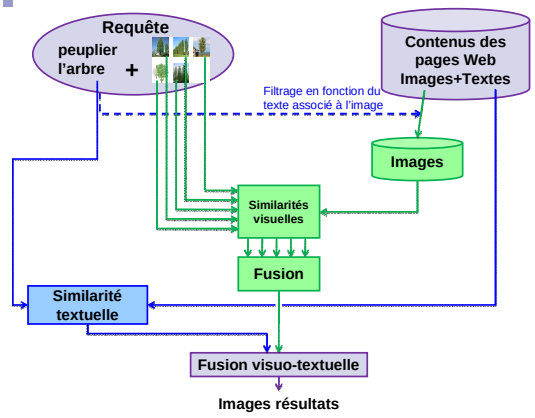
où

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

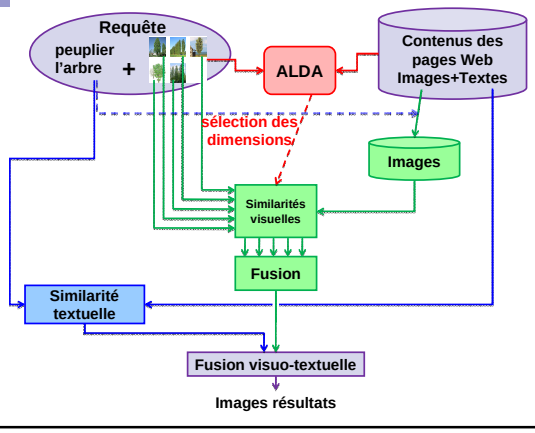
t représente le poids de texte dans la fusion.

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)



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 ?

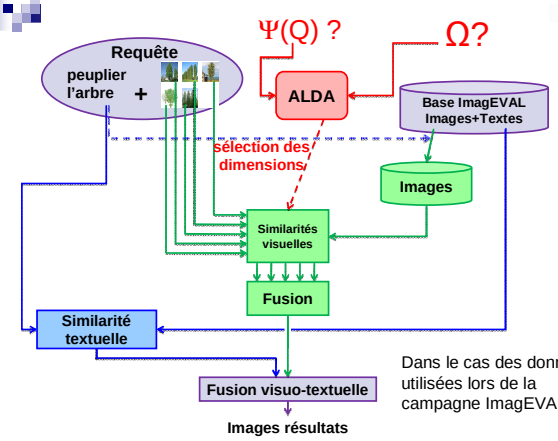


Sélection par Approximation de l'Analyse Linéaire Discriminante (ALDA)

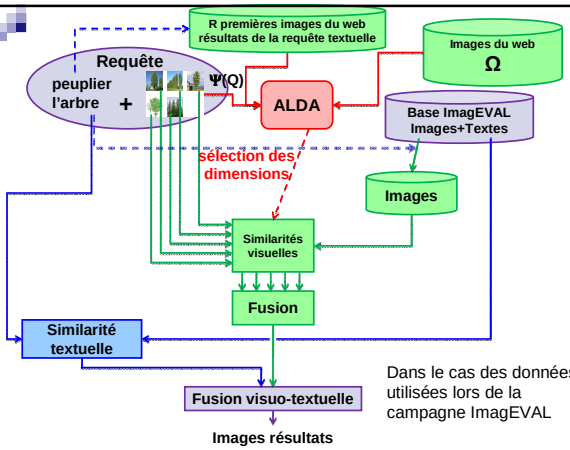
- Construire un ensemble **Ω** d'images du web suffisamment grand
- Pour chaque requête Q
  - Construire l'ensemble **Ψ(Q)** des images positives pour la requête
  - Pour chaque dimension X de l'espace visuel :
    - Calculer les variances interclasse $\hat{B}(X;Q)$ et intraclasse $\hat{W}(X;Q)$ entre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



Dans le cas des données utilisées lors de la campagne ImagEVAL



Dans le cas des données utilisées lors de la campagne ImagEVAL

Expérimentations

Extraction des descripteurs

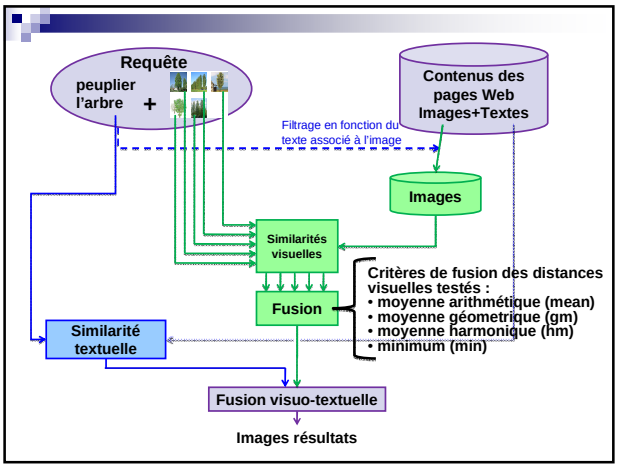
- Descripteurs visuels
  - Les images sont segmentées en 3 bandes horizontales
  - Pour chaque bande et pour chaque couleur L=R+G+B, r=R/L, g=G/L
    - Calculer la moyenne et l'écart-type des valeurs des pixels
    - Calculer les profils horizontaux, verticaux et globaux
  - L'espace visuel est composé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

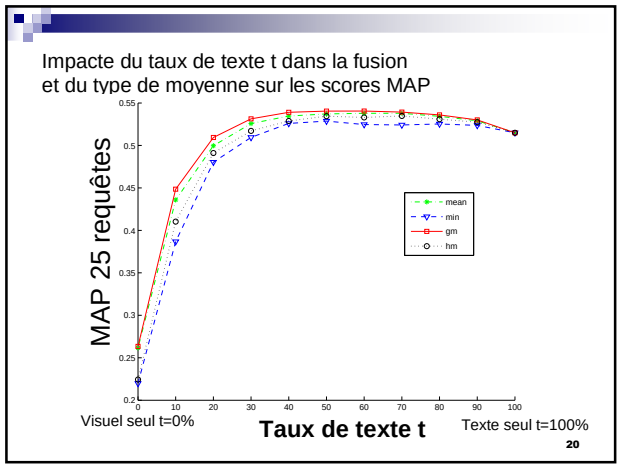
Résultats officiels de la campagne ImagEVAL

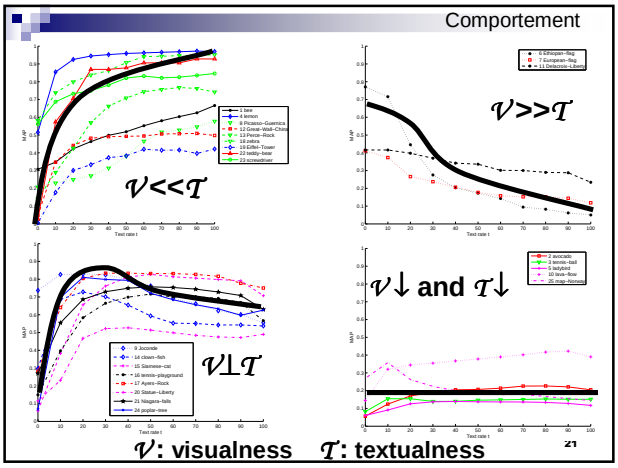
- 4 équipes
- 22 runs parmi lesquels 11 runs fusion, 7 runs texte seul et 5 runs visuel seul
- Nous avons proposé 3 runs et sommes arrivés 2nd / 4

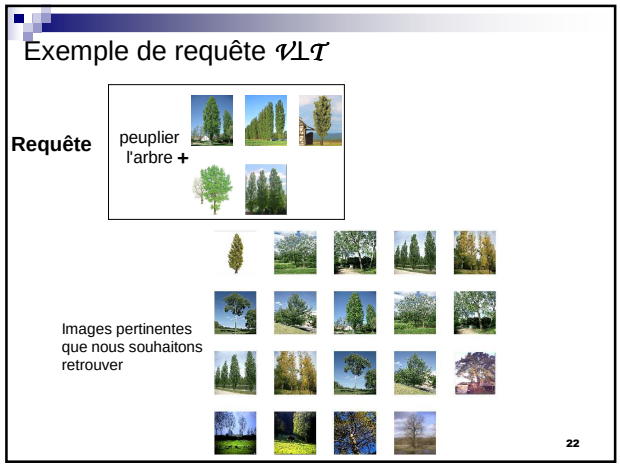
Rang équipe	Texte seul	Visuel seul	Fusion visuo-textuelle
1	0.559 (Run 5)	0.271 (Run 16)	**0.613 (Run 1)**
2	0.513* (Run 9)	0.261* (Run 17)	0.536* (Run 7)
3	0.455 (Run 12)	0.181 (Run 20)	0.517 (Run 8)

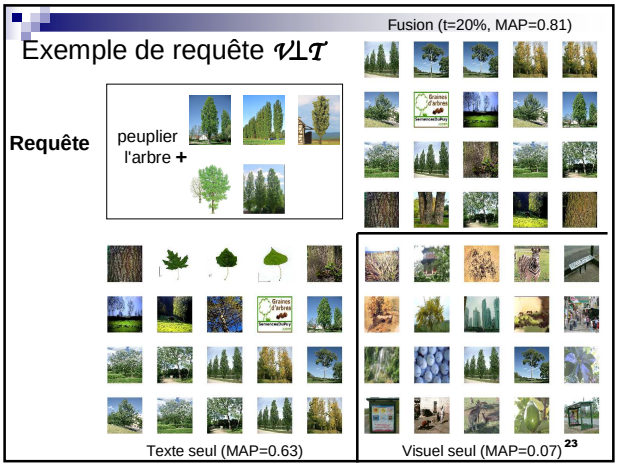
* Score MAP LSIS. Pour la fusion visuo-textuelle t=50% ; le critère de fusion visuelle est la moyenne arithmétique ; pas de sélection de la dimension ! (les runs sont ordonnés du plus fort (run1) au plus faible MAP (run22))

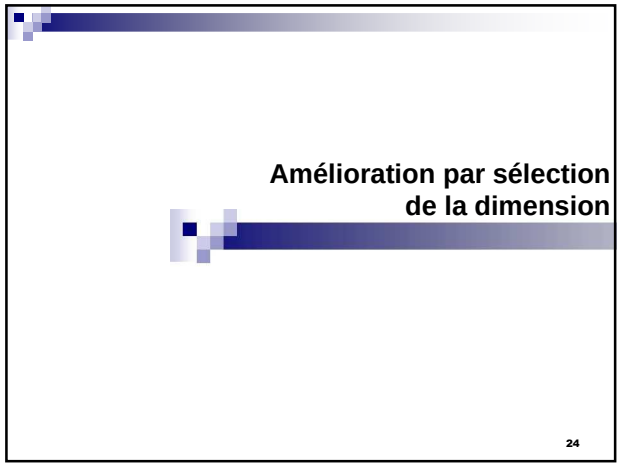

Requête
peuplier
l'arbre
+
Contenus des
pages Web
Images+Textes
Filtrage en fonction du
texte associé à l'image
Images
Similarités
visuelles
Critères de fusion des distances
visuelles testés :
• moyenne arithmétique (mean)
• moyenne géometrique (gm)
• moyenne harmonique (hm)
• minimum (min)
Fusion
Similarité
textuelle
Fusion visuo-textuelle
Images résultats

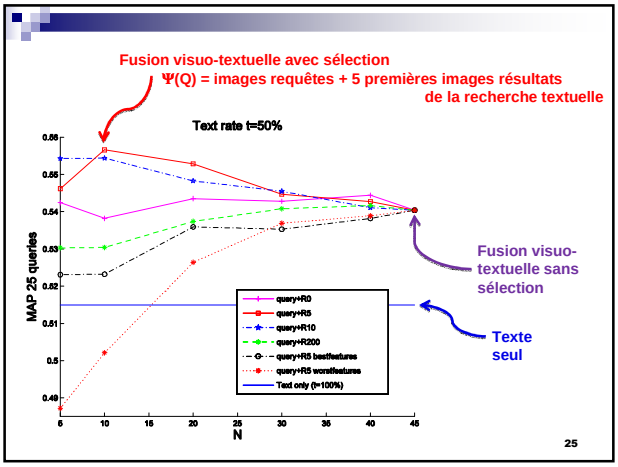

Impacte du taux de texte t dans la fusion
et du type de moyenne sur les scores MAP
MAP 25 requêtes
mean
min
gm
hm
Visuel seul t=0%
Taux de texte t
Texte seul t=100%
20


Comportement
$\mathcal{V}$<<$\mathcal{T}$
$\mathcal{V}$>>$\mathcal{T}$
$\mathcal{V}\perp\mathcal{T}$
$\mathcal{V}\downarrow$ and $\mathcal{T}\downarrow$
$\mathcal{V}$: visualness
$\mathcal{T}$: textualness
21


Exemple de requête $\mathcal{V}\perp\mathcal{T}$
Requête
peuplier
l'arbre +
Images pertinentes
que nous souhaitons
retrouver
22


Exemple de requête $\mathcal{V}\perp\mathcal{T}$
Fusion (t=20%, MAP=0.81)
Requête
peuplier
l'arbre +
Texte seul (MAP=0.63)
Visuel seul (MAP=0.07)
23


Amélioration par sélection
de la dimension
24



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

Ψ(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)

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)

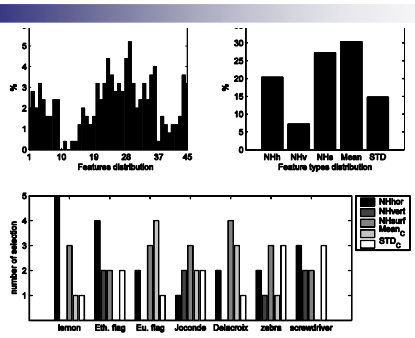
Merci pour votre attention

Informations sur les 25 requêtes

	Requête	Nombre d'images requête	Nombre d'images pertinentes
1	bee	7	39
2	avocado	7	39
3	tennis ball	4	20
4	lemon	6	94
5	Ladybird	6	19
6	Etiopian flag	1	13
7	European flag	1	31
8	Picasso Guernica	3	19
9	Joconde	2	14
10	Lava flow	7	66
11	Delacroix Liberty	3	11
12	Great Wall China	6	88
13	Perce Rock	7	33

	Requête	Nombre d'images requête	Nombre d'images pertinentes
14	clown fish	7	51
15	Siamese cat	6	33
16	tennis playground	9	40
17	Ayers Rock	6	41
18	zebra	6	30
19	Eiffel Tower	5	53
20	Statue Liberty	4	18
21	Niagara Falls	6	51
22	teddy bear	6	9
23	screwdriver	5	20
24	poplar tree	5	19
25	map Norway	6	8

http://www.imageval.org



Distributions des 10 traits les plus discriminants pour chaque requête en utilisant l'ALDA