

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/UMRCNRS 7606LIP6

**Université du Sud Toulon-Var/UMRCNRS6168LSIS

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

Trégastel, le 12 mars 2008, CORIA 2008

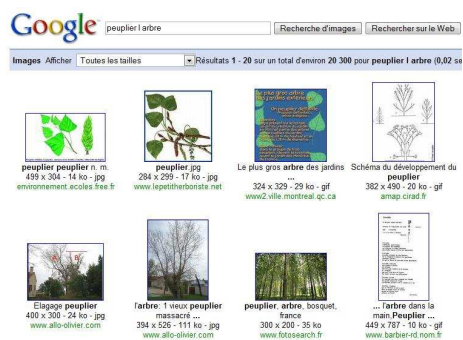
1

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)
- Expérimentation 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

2

Motivation: exemple de recherche textuelle



3

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/>

4

ImagEVAL

Description de la tâche 2 d'ImagEVAL

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

5

ImagEVAL

Exemples de requêtes visuo-textuelles de la campagne ImagEVAL

« poisson clown »



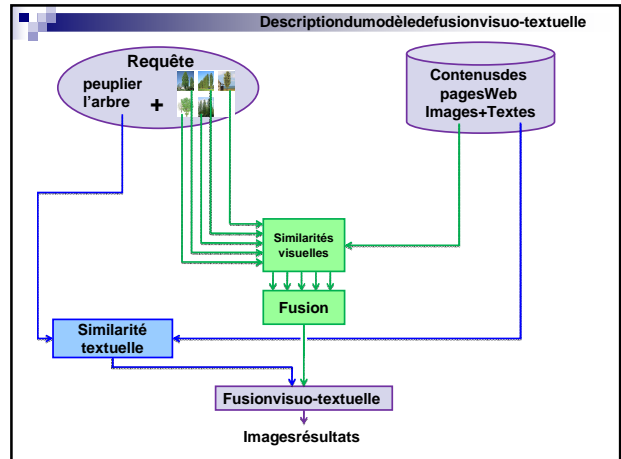
« tournevis »



6

Description du modèle de fusion visuo-textuelle

7



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:

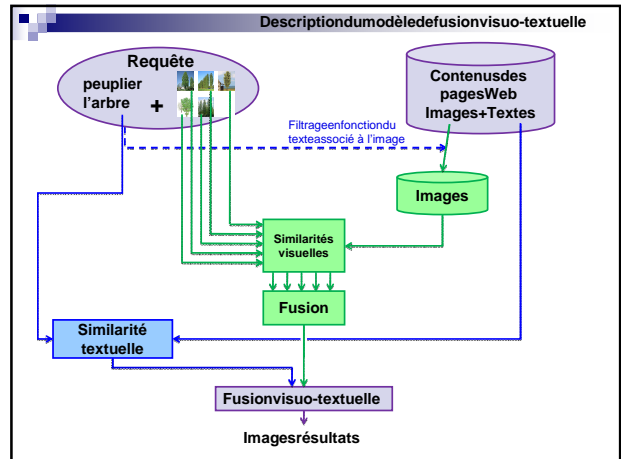
$$D(Q, I, t) = t \times D_T(Q, I) + (1 - t) \times D_V(Q, I)$$

où

D_T est la distance normalisée basée sur l'IDF, D_V est la distance visuelle entre les images de la requête et une image I et t représente le poids de texte dans la fusion.

Hypothèse: si 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)

9



Amélioration par sélection de la dimension

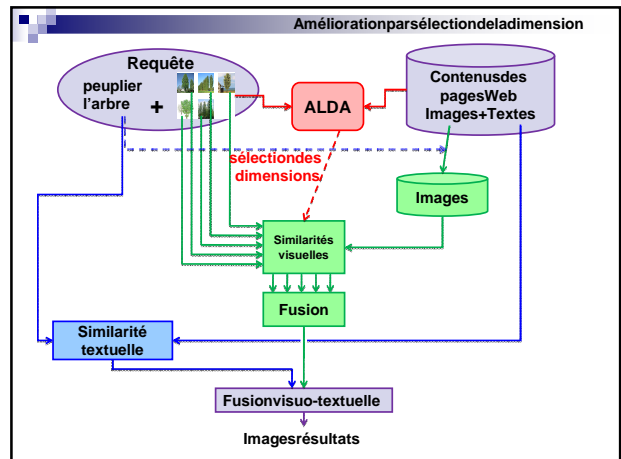
Problèmes:

- Réaliser des distances visuelles est très coûteuse 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?

11



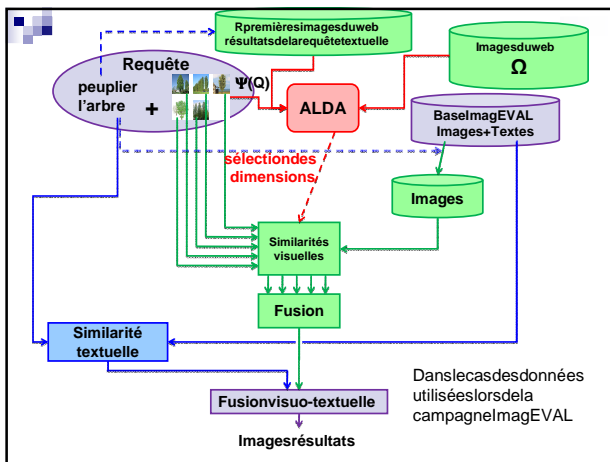
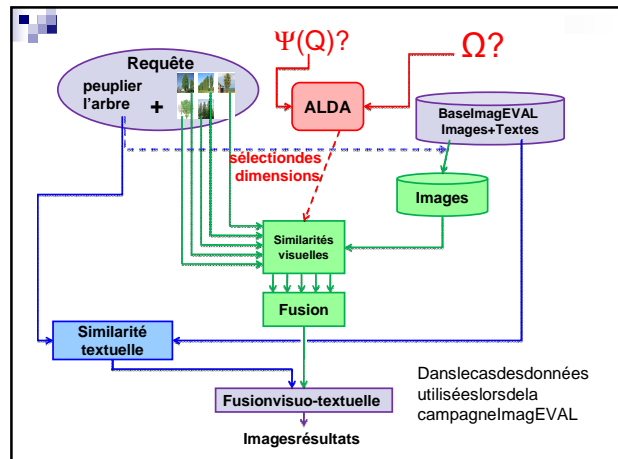
Sélection par Approximation de l'Analyse Linéaire Discriminante (ALDA)

- 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 inter-classe et $\hat{B}(X; Q)$ intra-classe $\hat{W}(X; Q)$ 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

13



Expérimentations

16

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, é, è, ê -> e...)
 - Les caractères spéciaux et les « stopwords » sont supprimés
- But:** avoir un traitement rapide des pages web

17

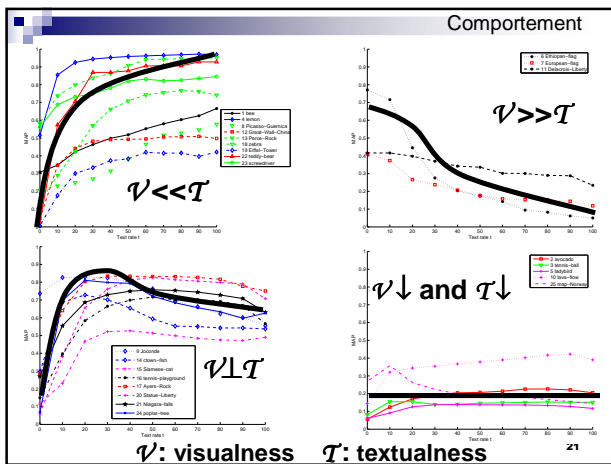
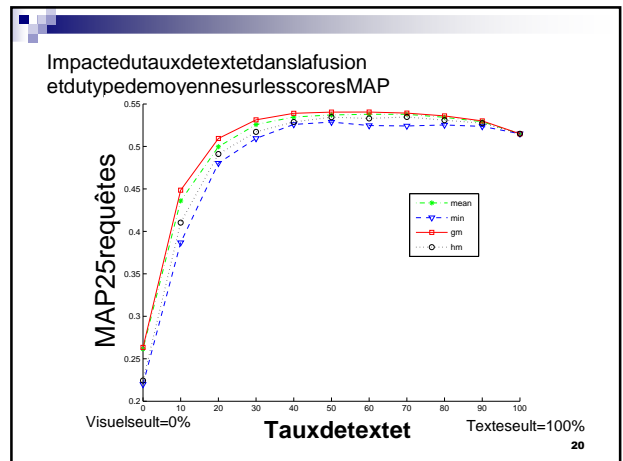
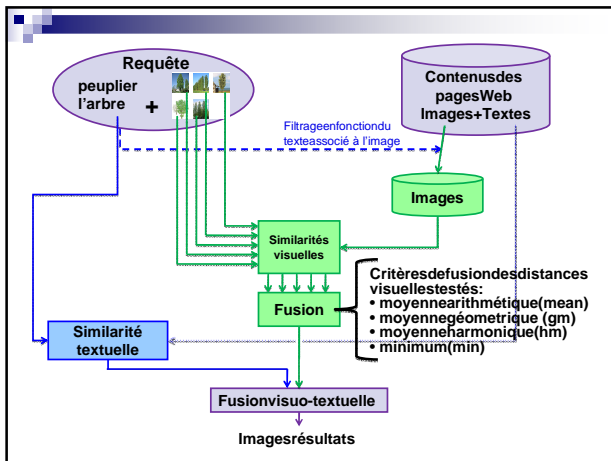
Résultats officiels de la campagne ImagEVAL

- 4 équipes
- 22 runs, parmi lesquels 11 runs fusion, 7 runs textuel 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 MAPLSIS. Pour la fusion visuo-textuelle = 50% le critère de fusion visuelle est la moyenne arithmétique; pas de sélection de la dimension! (les runs sont ordonnés du plus fort (run 1) au plus faible (run 22))

18



Exemple de requête $V \approx T$

Requête: peuplier l'arbre +

Images pertinentes que nous souhaitons retrouver

Exemple de requête $V \approx T$

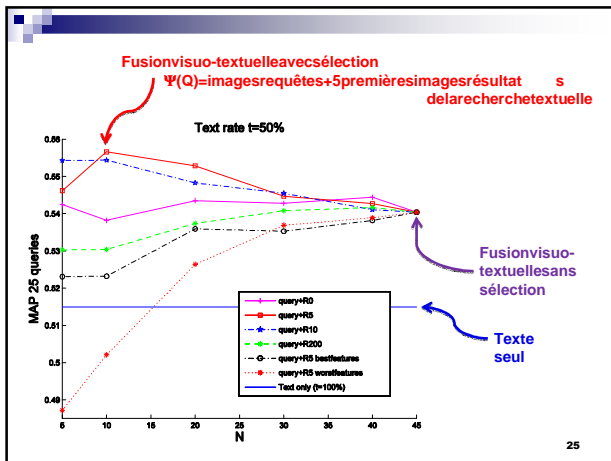
Fusion (t=20%, MAP=0.81)

Requête: peuplier l'arbre +

Texteseuil (MAP=0.63)

Visuelseuil (MAP=0.07)

Amélioration par sélection de la dimension



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

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

Conclusion

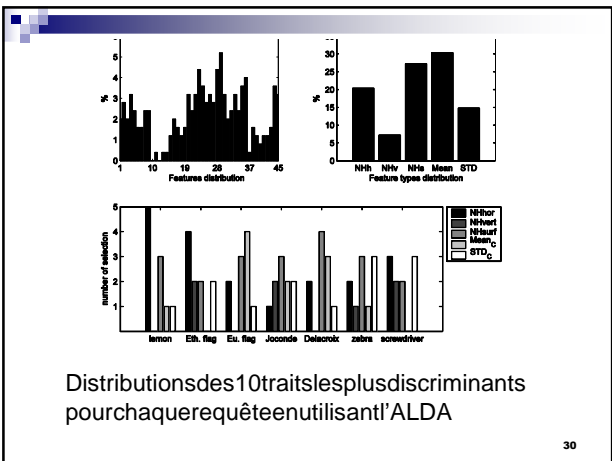
- La fusion visuo-textuelle permet d'améliorer la recherche d'images au moins pour certaines requêtes
- Le comportement des courbes de fusions visuo-textuelles dépend de la nature intrinsèque des requêtes
- La sélection de la dimension permet d'améliorer le score et tout en réduisant le temps de calcul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

Merci pour votre attention

Information sur les 25 requêtes

Requête	Nombre d'images requête	Nombre d'images pertinentes
1 bee	7	39
2 avocado	7	39
3 tennisball	4	20
4 lemon	6	94
5 Ladybird	6	19
6 Ethiopian flag	1	13
7 European flag	1	31
8 PicassoGuernica	3	19
9 Joconde	2	14
10 Lavalflow	7	66
11 DelacroixLiberty	3	11
12 GreatWallChina	6	88
13 PerceRock	7	33
14 clownfish	7	51
15 Siamese cat	6	33
16 tennisplayground	9	40
17 Ayers Rock	6	41
18 zebra	6	30
19 EiffelTower	5	53
20 StatueLiberty	4	18
21 NiagaraFalls	6	51
22 teddybear	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