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

Sabrina Tollari*, Hervé Glotin**

** Université Pierre et Marie Curie-Paris6, UMR CNRS 7606-LIP6*
*** Université du Sud Toulon-Var, UMR CNRS 6168-LSIS*

RÉSUMÉ. *Dans cet article, nous proposons une méthode pour améliorer la recherche d'images sur le web dans le cas de requêtes bimodales composées de quelques mots et de quelques images. Pour chaque page web et chaque requête, une moyenne pondérée fusionne les distances textuelles basées sur tfidf et les distances visuelles. Nous montrons alors que cette recherche bimodale d'images peut être optimisée en analysant simplement des images récupérées en ligne par des requêtes purement textuelle sur un moteur classique de recherche d'images sur le web. Nous approximons alors une Analyse Linéaire Discriminante (ALDA) sur ces images de développement pour estimer le sous-ensemble de traits optimaux de chaque requête traitée. Nous testons notre méthode sur la campagne Techno-Vision ImagEVAL (notre équipe s'y est classée 2nde sur 4), avec 700 URLs (700 pages web et 10k images). Nous discutons le comportement des résultats des requêtes en fonction du taux de texte dans la fusion. Les résultats montrent alors que nous pouvons automatiquement réduire le nombre de dimensions afin d'obtenir une réduction du temps de calcul de 35% sans dégradation des scores de Mean Average Precison.*

ABSTRACT. *In this article we propose a method to improve web image retrieval in the case of bimodal query composed of few words and few images. For each web image and multimodal query, a simple weighted distance merges the tfidf web page analysis and the visual features distance. We then apply an approximate linear discriminante analysis (ALDA) using web sampling to select the most discriminant visual feature. We check our method on the ImagEVAL official test set (our team was 2nd on 4 teams). This test set contains 700 URLs (700 web pages and 10k images). We study query behavior in fonction of the text rate int the fusion of the visual and textual distances, and we discuss of the "visualness" of each query. We show that we can reduce the number of dimensions to obtain a time reduction of 35%, without Mean Average Precison (MAP) degradation.*

MOTS-CLÉS : *recherche d'images, sélection de la dimension, ALDA, visualness*

KEYWORDS: *images retrieval, feature selection, ALDA, visualness*

1. Introduction

La recherche d'images par le contenu est toujours considérée comme une tâche difficile, c'est l'une des raisons pour lesquelles les moteurs de recherches d'images sur le web utilisent principalement des informations textuelles, telles que le titre de la page web, le nom de l'image, le texte adjacent, pour tenter de "comprendre" le sens de l'image. Cependant, le texte d'une page web n'est pas toujours en rapport avec l'image. Pour améliorer la recherche d'images combiner les informations textuelles et visuelles semble être une méthode prometteuse. De précédents travaux (Srihari, 1995, La Cascia *et al.*, 1998, Sclaroff *et al.*, 1997, Zhou *et al.*, 2002, Li *et al.*, 2003, Barnard *et al.*, 2001, Lavrenko *et al.*, 2003, Tollari *et al.*, 2005) montrent des expériences intéressantes de combinaison du texte et du visuel, mais aucun d'entre eux : (1) n'optimise le modèle avec une base d'images automatiquement extraites du web, (2) ni n'analyse objectivement la pertinence des modèle avec une vérité terrain déterminée par des experts pour des recherches d'images sur des URLs.

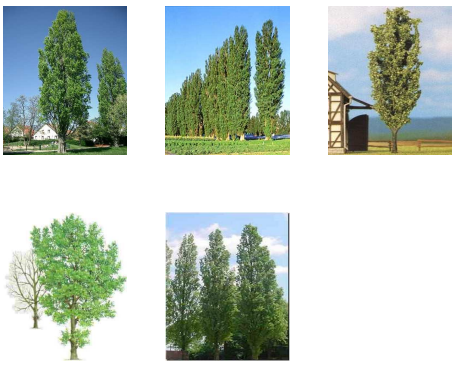
La nouvelle campagne ImagEVAL (ImagEVAL, 2006) du programme Techno-Vision propose un tel cadre pour ce genre d'étude. La seconde tâche d'ImagEVAL a pour objectif d'évaluer les techniques impliquant du texte et des images pour améliorer la recherche d'images dans le cadre de données multimédia. Cette tâche est fortement orientée web, car les données sont de vraies pages web contenant du texte et des images.

Dans la partie suivante, nous détaillons la tâche et le corpus de la campagne d'évaluation ImagEVAL. Ensuite, nous décrivons notre méthode d'estimation des traits visuels les plus discriminants sur des données mal annotées issues du web. Puis, nous expliquons comment utiliser la méthode de sélection des traits dans notre modèle de recherche d'images en ligne. Ensuite, nous présentons les résultats expérimentaux obtenus.

2. ImagEVAL

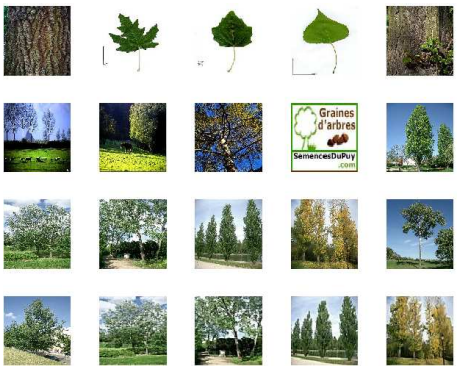
ImagEVAL (ImagEVAL, 2006) est une nouvelle campagne d'évaluation financée par le programme Techno-Vision. La seconde tâche de cette campagne consiste à retrouver des images du web en utilisant des informations textuelles et visuelles. La base de données proposée pour le test officiel est composée de 700 URLs. Nous avons récupéré les 700 pages web et les 10k images associées. Parmi ces 10k images, seules 5k images n'étaient pas de trop "petites" images, des images blanches, des images sans contenu...

Le but de la tâche est de trouver les images (parmi les 10k images du web) qui répondent à une requête $Q = \{\mathcal{K}(Q), \mathcal{I}(Q)\}$ composée d'un ensemble de mots $\mathcal{K}(Q)$ (par exemple : "la Tour Eiffel") et d'un ensemble d'images $\mathcal{I}(Q)$ (qui ne sont pas dans la base de test). Par exemple, la figure 1(a) montre les 5 images requêtes de la requête 24 dont les mots sont "peuplier, l'arbre". La figure 1(b) donne les 19 images pertinentes correspondantes. Le tableau 1 décrit les 25 requêtes du test officiel.

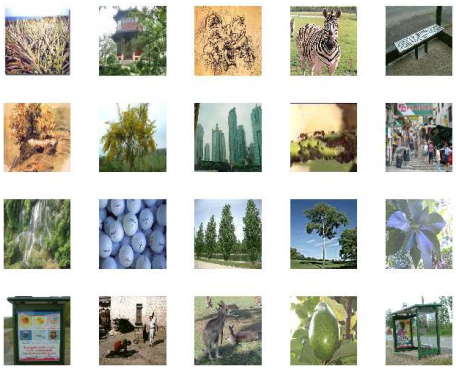


(a) Les 5 images requêtes de la requête 24

(b) Les 19 images pertinentes de la requête 24 qu'il faut retrouver parmi les 10k images

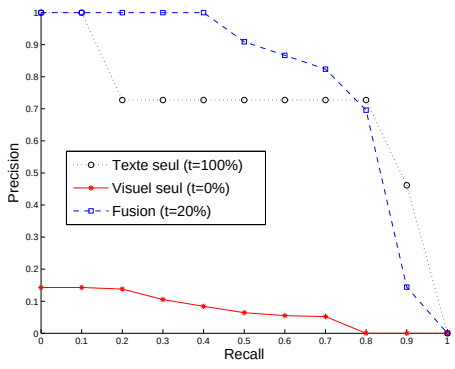


(c) 20 premiers résultats de la requête "peuplier, l'arbre" (texte seul) (MAP=0.63)



(d) 20 premiers résultats de la requête composée uniquement des images de la figure 1(a) (visuel seul) (MAP=0.07)

(e) 20 premiers résultats de la requête visuo-textuelle (fusion $t = 20\%$) (MAP=0.81)



(f) Courbes de rappel/précision correspondantes

Figure 1. *Exemples de résultats obtenus pour la requête 24 composée des mots* "peuplier, l'arbre" *et des 5 images de la figure 1(a). Les scores présentés sont ceux obtenus sans sélection de la dimension ($\gamma = gm$).*

Tableau 1. *Informations sur les requêtes. Chaque requête Q est composée d'un ensemble de mots $\mathcal{K}(Q)$ et d'un ensemble d'images requêtes $\mathcal{I}(Q)$*

Numéro de la requête	Mots de la requête	Nombre d'images requêtes	Nombre d'images pertinentes
1	abeille	7	39
2	avocat, le fruit	7	39
3	balle de tennis	4	20
4	citron	6	94
5	coccinelle, l'insecte	6	19
6	le drapeau éthiopien	1	13
7	le drapeau européen	1	31
8	Guernica de Picasso	3	19
9	la Joconde	2	14
10	coulée de lave	7	66
11	la Liberté guidant le peuple de Delacroix	3	11
12	la grande muraille de Chine	7	88
13	le Rocher Percé	6	33
14	poisson clown	7	51
15	chat siamois	6	33
16	court de tennis	9	40
17	Uluru, Ayers Rock	6	41
18	zèbre	6	30
19	la Tour Eiffel	5	53
20	la Statue de la Liberté	4	18
21	les chutes du Niagara	6	51
22	ours en peluche	6	9
23	tournevis	5	20
24	peuplier, l'arbre	5	19
25	carte de la Norvège	6	8

3. Sélection de la dimension sur des données mal-annotées

Lorsqu'un utilisateur effectue une requête, il faut qu'il obtienne des résultats pertinents en un temps raisonnable. Pour obtenir des résultats plus pertinents que ceux obtenus par les méthodes utilisées en recherche d'information conventionnelle, il peut être intéressant d'utiliser des techniques supervisées, mais cela nécessite d'avoir des données d'apprentissage. Or les grandes masses de données bien étiquetées sont trop coûteuses et donc rares, voire inexistantes. Le web permet d'accéder à beaucoup d'images, mais celles-ci sont majoritairement mal étiquetées, dans le sens où le web n'offre pas des données caractérisées par une bijection entre un mot (ou concept) et une région (ou objet) d'une image. Nous montrons néanmoins dans cet article que ces images issues du web peuvent servir à optimiser un modèle de recherche visuo-

textuelle d'images. Pour cela, pour chaque requête Q du test officiel, nous utilisons les mots associés $\mathcal{K}(Q)$ pour interroger un moteur de recherche d'images du web, et nous récupérons les images renvoyées. Ces images ne correspondront pas toutes visuellement à la requête, car les images du web sont annotées par rapport au texte de la page web, et leur contenu visuel est ignoré par le moteur de recherche. Cependant, nous les utilisons comme exemples positifs pour la requête Q. L'ordre des résultats obtenus est important car généralement les images sont triées de la plus pertinente à la moins pertinente. Par la suite, nous notons $\mathcal{A}_{1:R}(Q)$ les R premières images retournées par le moteur de recherche d'images du web pour la requête composée des mots $\mathcal{K}(Q)$.

Pour obtenir des résultats en un temps raisonnable, il est intéressant d'utiliser des techniques qui gardent ou augmentent la pertinence des résultats obtenus tout en réduisant le temps de calcul, comme par exemple, les techniques de sélection de la dimension. De plus, réduire le nombre de dimensions de l'espace visuel permet de passer outre le problème de la malédiction de la dimension (Beyer *et al.*, 1999, Amsaleg *et al.*, 2004). L'analyse linéaire discriminante (LDA) est l'une de ces techniques. Elle permet de sélectionner les dimensions les plus discriminantes, ce qui réduit le nombre de dimensions de l'espace et par conséquent le temps de calcul de la similarité entre deux images, tout en éliminant les dimensions non-discriminantes. Cette technique nécessite normalement un corpus d'apprentissage bien étiqueté. Dans (Glotin *et al.*, 2006), nous avons développé une approximation de la LDA (ALDA) pour découvrir les traits visuels les plus discriminants pour une requête donnée à partir d'images mal-étiquetées.

Le pouvoir discriminant de l'ALDA est calculée comme suit. Pour chaque requête, nous séparons les données d'apprentissage en deux ensembles : l'ensemble $\Psi(Q)$ composé des images considérées positives pour la requête Q (par exemple, les images requêtes $\mathcal{I}(Q)$ et/ou les images de $\mathcal{A}_{1:R}(Q)$) et l'ensemble Ω contenant l'ensemble de toutes les images d'apprentissage. Pour chaque trait visuel X, nous calculons la variance interclasse $\hat{B}(X;Q)$ (moyenne des variances de chaque classe) et la variance intraclasse $\hat{W}(X;Q)$ (variance des moyennes de chaque classe). Finalement, nous estimons pour chaque requête Q et pour chaque trait visuel X le pouvoir discriminant :

$$\hat{J}(X;Q) = \frac{\hat{B}(X;Q)}{\hat{B}(X;Q) + \hat{W}(X;Q)}. \quad [1]$$

Une validation théorique de l'ALDA est démontrée dans (Glotin *et al.*, 2006), et une expérimentale sur la base COREL dans (Glotin *et al.*, 2006, Tollari *et al.*, 2006). Nous montrons dans (Glotin *et al.*, 2006) que l'erreur d'ordonnancement des traits visuels (du trait le plus discriminant au moins discriminant en fonction de la valeur de $\hat{J}(X;Q)$) est petite tant que le cardinal de Ω est grand devant celui de $\Psi(Q)$.

4. Recherche combinée texte et images

Pour effectuer une recherche d'images en utilisant seulement le contenu visuel (méthode dite "Visuel seul"), nous procédons comme suit. Premièrement, le système

détermine, hors ligne, une fois pour toute l'ensemble Ω et extrait les descripteurs visuels. Deuxièmement, il effectue en ligne l'algorithme suivant :

pour chaque requête bi-modale $Q = \{\mathcal{K}(Q), \mathcal{I}(Q)\}$ **faire**
  Poser la requête textuelle $\mathcal{K}(Q)$ dans un moteur de recherche d'images du web
  Utiliser les R premiers résultats $\mathcal{A}_{1:R}(Q)$ et $\mathcal{I}(Q)$ pour construire $\Psi(Q)$
  Sélectionner les N traits les plus discriminants en fonction de $\Psi(Q)$ et Ω
  pour chaque image I_i du test officiel **faire**
    pour chaque requête image q_j de $\mathcal{I}(Q)$ **faire**
      Calculer la distance euclidienne δ entre le vecteur visuel réduit de q_j et le vecteur visuel réduit de I_i : $\delta_{i,j} = L2norm(q_j, I_i)$
    fin pour
    Calculer δ_i^* la distance entre I_i et l'ensemble des images requêtes $\mathcal{I}(Q)$ en fonction de l'opérateur de moyenne γ : $\delta_i^* = \gamma_{j=1}^{n_{\mathcal{I}(Q)}}(\delta_{i,j})$
  fin pour
  Retourner les 300 premières images qui ont la plus petite distance δ^*
fin pour

L'opérateur de moyenne γ peut être la moyenne arithmétique (*mean*), le minimum (*min*), la moyenne géométrique (*gm*) ou la moyenne harmonique (*hm*). Notons que si N égale le nombre total de dimensions ($N = 45$) alors il n'y a pas de sélection des dimensions.

Pour réaliser une recherche d'images visuo-textuelle, nous fusionnons les informations visuelles et textuelles en utilisant une moyenne pondérée des distances textuelles et visuelles (une bonne alternative pourrait être l'ordonnancement moyenné). Un poids t, qui indique le taux de texte dans la fusion, est appliqué aux distances textuelles normalisées D_T (basées sur tfidf ; voir partie 5.1) tandis que le poids $1 - t$ est appliqué aux distances visuelles normalisées D_V. Nous obtenons finalement la distance visuo-textuelle par : $D = t \times D_T + (1 - t) \times D_V$. Dans la partie expérimentation, nous discutons de l'impact du poids t dans les résultats globaux, mais aussi pour chaque requête. (Yanai *et al.*, 2005) propose de déterminer dans quelle mesure un concept a des caractéristiques visuelles discriminantes pour l'annotation d'images, c'est ce qu'il appelle le *visualness* d'un concept. Par exemple, le concept de véhicules ne peut être directement relié à des caractéristiques visuelles, car des véhicules peuvent apparaître sous différentes formes, textures ou couleurs. Dans notre modèle, le *visualness* $\mathcal{V}$ peut être interprété comme le complémentaire du taux de texte t optimal. Dualement, nous définissons la notion de *textualness* $\mathcal{T}$ en fonction de l'efficacité du texte. Nous discutons de ces notions dans la partie 5.3.

5. Expérimentations

Dans cette partie, nous décrivons premièrement les descripteurs visuels (ou "traits") que nous utilisons pour décrire le contenu des images. Nous détaillons ensuite les résultats du test officiel. Puis nous discutons de l'importance du caractère visuel (*visualness*) ou textuel (*textualness*) de chaque requête. Ensuite, nous montrons

Tableau 2. *Résultats officiels de la tâche 2 de la campagne ImagEVAL. Dans la campagne, 22* runs *ont été proposés par les participants parmi lesquels 11* runs *de fusions, 7 Texte seul et 5 Visuel seul). *LSIS **XEROX*

Rang	Texte seul	Visuel seul	Fusion
1	0.559(Run5)**	0.271(Run 16)	0.613(Run1)**
2	0.513(Run9)*	0.261(Run17)*	0.536(Run7)*
3	0.455(Run12)	0.181(Run20)	0.517(Run8)

comment nous pouvons améliorer nos résultats par sélection de la dimension. Finalement, nous discutons de l'importance du nombre de données nécessaires pour calculer l'ALDA.

5.1. *Extraction des descripteurs visuels et textuels*

Afin que la recherche d'images par le contenu visuel s'effectue rapidement, nous proposons des descripteurs visuels compactes et légers en calculs. Nous segmentons donc grossièrement chaque image en trois bandes horizontales égales desquelles sont extraient les descripteurs visuels. Pour des raisons d'efficacité, nous n'utilisons pas de conversion de couleurs, nous faisons une simple normalisation ($r = R/(R+G+B)$, $g = G/(R+G+B)$, et $L = R+G+B$). Nous calculons alors les descripteurs classiques : moyenne ($Mean_C$) et écart-type (STD_C) de chaque couleur pour chacune des trois sous bandes de l'image. Ces traits serviront de référence par rapport aux traits moins conventionnels mais rapidement calculables décrits ci-dessous. Nous proposons de nouveaux descripteurs à partir des entropies des profils des sous-bandes d'images qui peuvent être assimilés à la quantité d'information visuelle (ou *visualness*). Ces profils sont les projetés (sommes) horizontaux ($NHhor$) ou verticaux ($NHvert$). Pour chacune des 3 couleurs nous en calculons l'entropie d'après l'histogramme du profil. Nous considérons aussi pour chaque bande l'entropie de l'histogramme des valeurs des pixels ($NHsurf$). Pour chaque bande et pour chaque couleur, nous extrayons donc 5 types de traits, soit un total de 45 traits (ou dimensions) par image.

Pour extraire les descripteurs textuels, les pages web sont d'abord, pour plus de rapidité, converties en texte en enlevant les balises HTML[1], tout en gardant le contenu des balises, notamment le nom de l'image et son URL. Les caractères spéciaux et les mots-vides sont supprimés. Le même traitement est appliqué au texte des requêtes. Puis, nous comptons le nombre d'occurrences de chaque mot-clé dans chaque page web. Nous supposons ensuite que chaque mot-clé de la requête a la même importance et nous recherchons les pages web qui ont au moins l'un des mots-clés. Nous calculons les poids standards du tfidf (Salton *et al.,* 1988) et les associons à chaque image de

1. De précédents travaux (Coelho *et al.,* 2004) montrent que l'utilisation des informations de balisage du texte (mots en gras, mots-clés du titre, distance entre mots et images...) permettent d'améliorer la recherche d'images. Cependant, nous n'étudions pas cet aspect dans cet article.

chaque page web. Pour rechercher les images pertinentes pour chaque requête à partir seulement des informations textuelles, nous trions les images dans l'ordre décroissant de leur tfidf et nous gardons seulement les 300 premières images comme demandé dans la campagne ImagEVAL.

5.2. *Résultats du test officiel de la tâche 2 d'ImagEVAL*

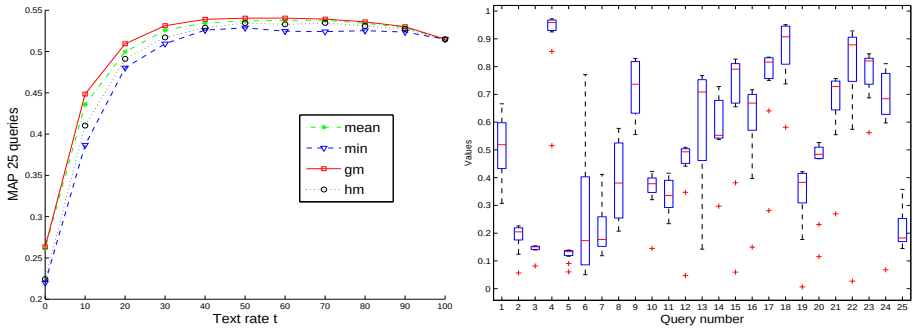
Nous présentons dans le tableau 2 les meilleurs résultats du test officiel par catégorie : Texte seul, Visuel seul et Fusion (voir le site officiel d'ImagEVAL (ImagEVAL, 2006) pour les détails). Notre modèle, en dépit de sa simplicité, est arrivé second dans chacune des catégories, derrière XEROX qui utilise son modèle de langue performant. Les descripteurs visuels des trois premiers modèles sont composés d'environs 50 dimensions. Ils contiennent des traits visuels plus ou moins standard, mais sont tous choisis pour leur faible coût CPU. En ce qui concerne le temps de traitement, le consortium ImagEVAL a proposé trois principaux temps : (1) extraction des descripteurs, (2) apprentissage et modélisation et (3) recherche. La moyenne de ces trois temps montre que notre modèle obtient un temps similaire aux autres modèles, avec 1,5 pages analysées par seconde. Pour le test officiel, nous n'avons utilisé aucun ensemble d'apprentissage. Nous avons mis *a priori* le taux de texte t à 50% et avons utilisé la moyenne arithmétique. Dans la prochaine partie, nous discutons du choix de t.

5.3. *Résultats globaux et analyse des comportements de fusion*

Nous avons testés quatre moyennes : le minimum *(min)*, la moyenne arithmétique *(mean)*, la moyenne géométrique *(gm)* et la moyenne harmonique *(hm)*. La figure 2(a) montre que le minimum donne les plus mauvais résultats, tandis que les moyennes arithmétique et géométrique donnent les meilleurs résultats. Les courbes MAP atteignent leur maximum entre $t = 40\%$ et $t = 70\%$. Par la suite, nous considérerons que la moyenne géométrique. La figure 2(b) illustre les variations du MAP de chaque requête en fonction de t.

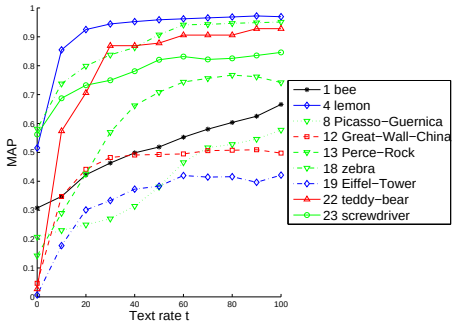
Pour analyser en détails les raisons de ces différences de variations nous séparons les requêtes en deux classes : les requêtes qui ont des courbes MAP en fonction de t strictement monotones et celles qui ont des courbes non-monotones. La figure 2(c) montre des courbes strictement croissantes. Nous pouvons dire que pour ces courbes $\mathcal{V} << \mathcal{T}$. D'un autre côté, les courbes des requêtes 6 (*le drapeau éthiopien*), 7 (*le drapeau européen*) et 11 (*la Liberté guidant le peuple de Delacroix*) sont strictement décroissantes (figure 2(d)). Ces trois requêtes sont mieux discriminées par leurs traits visuels que par le texte ($\mathcal{V} >> \mathcal{T}$).

La fusion la plus intéressante est illustrée figure 2(e). Les courbes des requêtes 9, 14, 15, 16, 21 et 24 atteignent leur maximum pour une valeur de t intermédiaire. Pour les requêtes 24 (*peuplier, l'arbre*), 9 (*la Joconde*) et 14 (*poisson clown*), le t optimal

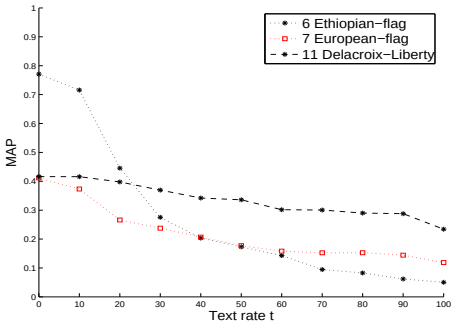


(a) MAP des 25 requêtes variant le taux de texte t dans la fusion pour différents opérateurs de moyenne γ

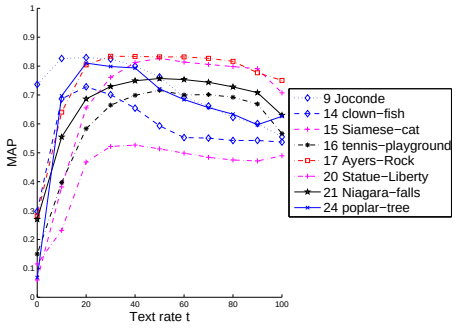
(b) Boxplot des valeurs de MAP pour chaque requête (t variant de $t = 0\%$ à $t = 100\%$ par pas de 10%, γ=gm)



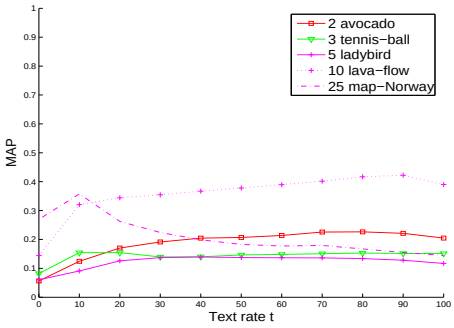
(c) Courbes MAP croissantes ($\mathcal{V} << \mathcal{T}$)



(d) Courbes MAP décroissantes ($\mathcal{V} >> \mathcal{T}$)



(e) Courbes MAP non monotoniques ($\mathcal{V} \perp \mathcal{T}$)



(f) Courbes MAP plates (faibles $\mathcal{V}$ et $\mathcal{T}$)

Figure 2. *Études du comportement des courbes MAP en fonction du taux de texte t utilisé dans la fusion ($\mathcal{V}$:* visualness *et $\mathcal{T}$:* textualness*). Ces courbes sont obtenues sans sélection de la dimension*

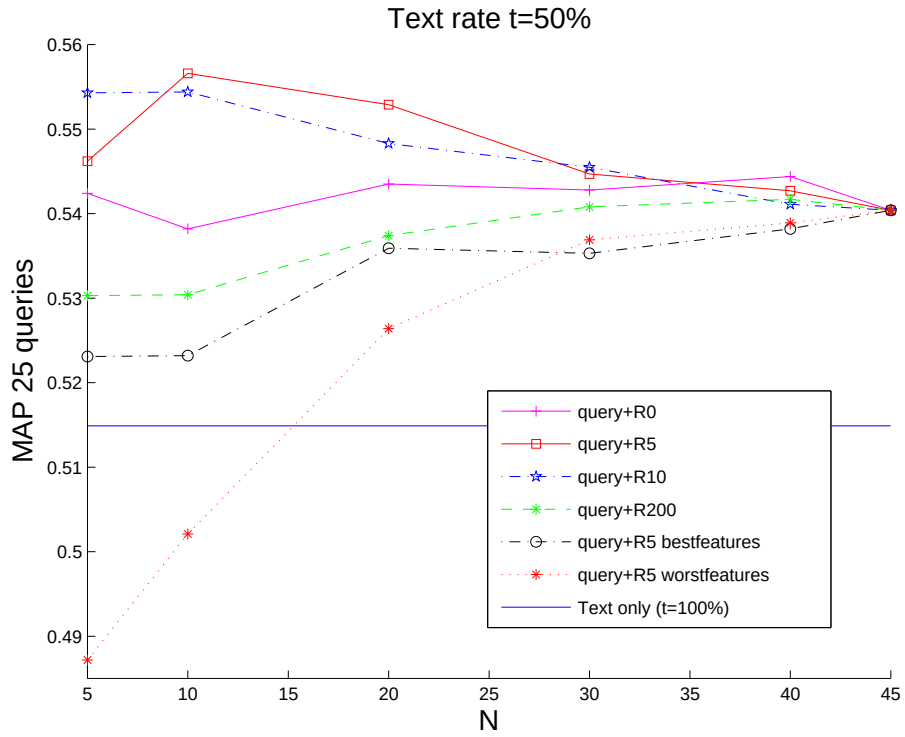


Figure 3. *Courbes MAP Fusion t=50% pour différents* $\Psi(Q)$

est 20%, tandis que pour les requêtes 3 (*balle de tennis*), 21 (*les chutes du Niagara*) et 15 (*chat siamois*) le t optimal est 50%. Les figures 1(c), 1(d) et 1(e) comparent les 20 premiers résultats pour les méthodes basées sur le Texte seul, le Visuel seul ou la Fusion des deux pour la requête 24 (*peuplier, l'arbre*). Nous pouvons voir que lorsque l'on utilise le texte seulement, les images retournées ont bien un rapport avec les peupliers (feuille, tronc...) mais que la plupart ne correspondent pas visuellement à ce que l'utilisateur attendait, information qu'il donne avec les images requêtes (figure 1(a)). Lorsque l'on utilise seulement les informations visuelles, les résultats obtenus sont très mauvais (MAP=0.07), car les images visuellement similaires ne sont pas forcement sémantiquement similaires. Pour ces requêtes, l'utilisation du texte seulement ou bien de l'information visuelle seule n'est pas optimale, les informations textuelles et visuelles sont complémentaires : $\mathcal{V} \perp \mathcal{T}$. Ce type de requêtes montre le grand intérêt des modèles de fusion visuo-textuelle.

Tableau 3. *Meilleurs MAP pour les expériences "query+R5". La sélection des dimensions par l'ALDA améliore légèrement les scores MAP, tout en réduisant le temps de calcul des distances visuelles par rapport au temps de calcul sans sélection des dimensions ($N = 45$). Temps : nombre de secondes nécessaires pour calculer la distance entre les 131 vecteurs visuels des images requêtes et 100k vecteurs visuels synthétiques*

	Sélection	t	N	MAP	Temps	Rapidité
Texte seul	-	100%	-	**0.513**	-	-
Visuel seul	sans	0%	45	0.261	309	1
Visuel seul	avec	0%	20	**0.271**	237	1.2
Visuel seul	avec	0%	10	0.269	**202**	1.5
Fusion	sans	50%	45	0.539	309	1
Fusion	avec	50%	20	0.553	237	1.2
Fusion	avec	50%	10	**0.557**	**202**	1.5

5.4. *Amélioration par sélection en ligne sur le web des dimensions visuelles*

Lors de la campagne officielle, nous n'avions pas utilisé de méthode de sélection de la dimension. Pour améliorer nos précédents résultats, nous proposons d'estimer et de sélectionner les traits visuels les plus discriminants pour chaque requête en utilisant l'ALDA pour différents ensembles $\Psi(Q)$. Pour toutes les expériences, l'ensemble Ω est composé de 17k images du web obtenus par l'union des ensembles $\mathcal{A}$ des 25 requêtes. La figure 3 montre que si $\Psi(Q) = \mathcal{I}(Q)$ ("query+R0") alors quand le nombre de dimensions N décroît, il n'y pas de baisse des scores MAP. Si maintenant nous utilisons en plus des images de l'ensemble d'apprentissage ($\Psi(Q) = \mathcal{I}(Q) \cup \mathcal{A}_{1:R}(Q)$ noté "query + Rx" avec pour valeur de $R : x \in$ 5, 10 ou 200), alors nous notons que nous obtenons les meilleurs MAP pour $N = 10$ et $R = 5$. L'expérience nommée Worstfeatures a consisté à choisir les N dimensions les moins discriminantes en moyenne sur toutes les requêtes, tandis que l'expérience nommée Bestfeatures a consisté à choisir les N dimensions les plus discriminantes en moyenne pour toutes les requêtes (et non pas pour chaque requête). Par exemple, pour $N = 10$, les 10 dimensions choisies pour toutes les requêtes sont les 10 dimensions les plus discriminantes d'après la figure 4 (en haut à gauche). La comparaison des courbes "query+R5" et "query+R5 Bestfeatures" montre bien l'importance d'adapter les traits visuels que l'on utilise en fonction de la requête.

Pour calculer les 25 pouvoirs discriminants sur un bi-Xeon 3Ghz 4Mb RAM, nous avons eu besoin de 4,6 secondes quelque soit la valeur de R. Le temps nécessaire pour calculer la fusion entre les distances textuelles et visuelles est d'environ 1 seconde pour les 25 requêtes et pour un t donné. Pour mettre en évidence la réduction du temps obtenue en utilisant l'ALDA, nous avons calculé le temps nécessaire pour réaliser la distance entre les 131 images requêtes et 100k vecteurs synthétiques de même taille

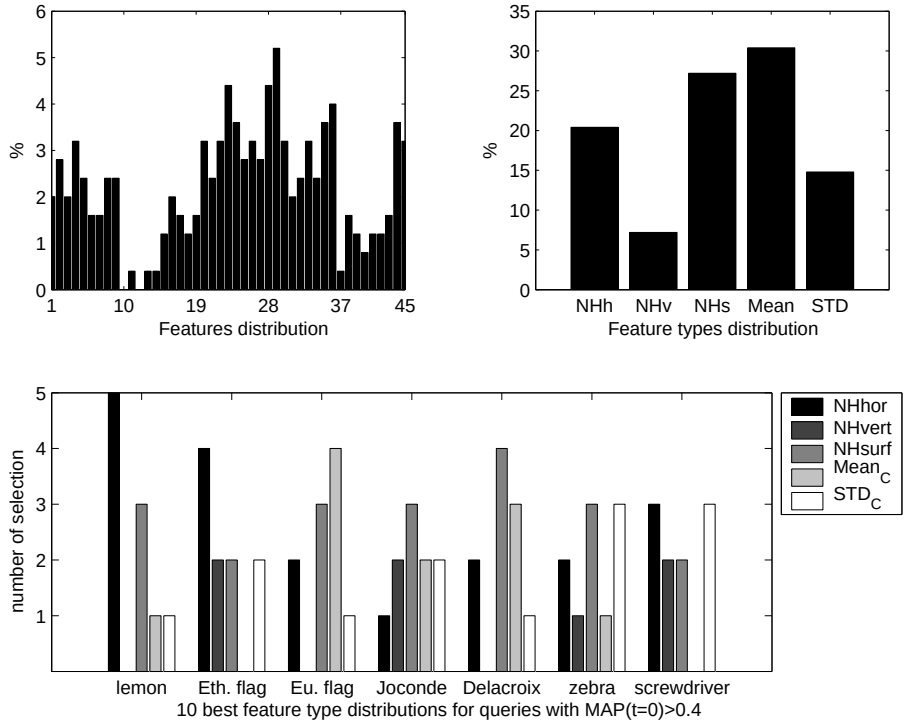
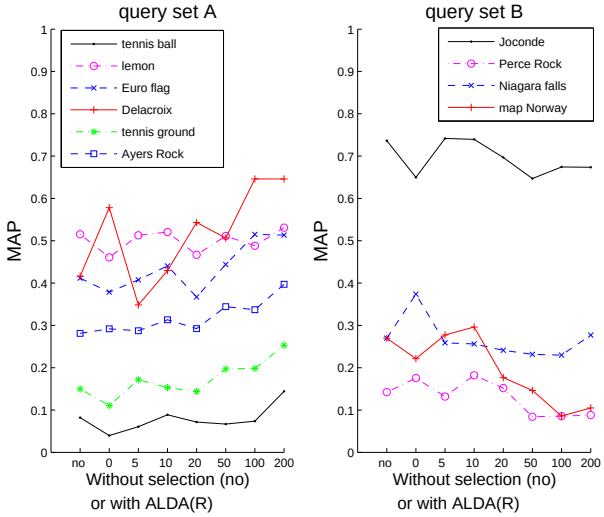


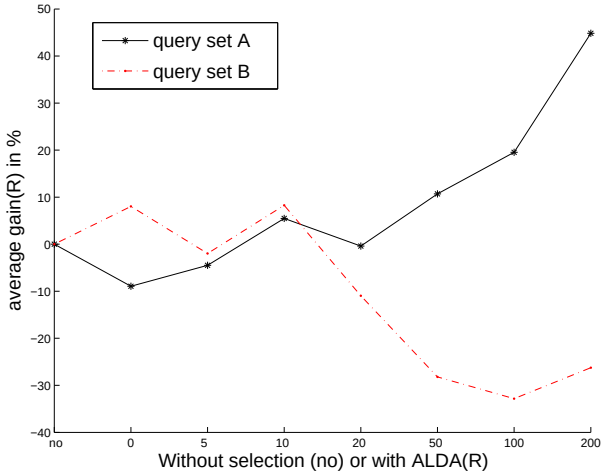
Figure 4. *Distribution des 10 traits visuels les plus discriminants pour chaque requête en exécutant l'ALDA dans le cas "query+R5". (en haut à gauche) 1-9 :* $NHhor$*, 10-18 :* $NHvert$*, 19-27 :* $NHsurf$*, 28-36 :* $Mean_C$*, 37-45 :* STD_C

que les vecteurs de notre base. Le tableau 3 donne les résultats obtenus. Il montre que nous pouvons augmenter les scores MAP (de 0.539 à 0.557) et en même temps réduire le temps de calcul de 35% (de 309 à 202). Nous voyons que les scores sont presque identiques pour $N = 10$ et $N = 20$, mais que le temps de calcul est plus faible pour $N = 10$ (voir aussi la figure 3). Pour le test officiel d'ImagEVAL, le temps total pour retrouver les 300 images demandées est estimé à 0,8 secondes pour chaque requête visuo-textuelle (sans compter le temps d'extraction des vecteurs visuels). Le temps d'extraction des vecteurs visuels est de 0.17 secondes par image. Notons que nos traits visuels sont en moyenne mieux adaptés (plus souvent choisis par l'ALDA) que les traits basiques r,g,L, leur moyenne ou STD, ce qui valide nos traits "entropie de profil".

Nous analysons en détails dans la figure 4 les traits visuels sélectionnés. Nous présentons d'abord (figure en haut à gauche) l'histogramme des 10 traits visuels les plus sélectionnés par l'ALDA pour chaque requête. Puis nous moyennons ces histogrammes par catégorie de traits visuels : $NHhor$, $NHvert$, $NHsurf$, $Mean_C$ et



(a) Comportement des requêtes des ensembles A et B en fonction de R



(b) Gain MAP moyen en %

Figure 5. *(a) Importance du nombre d'images (=R) utilisées pour calculer le pouvoir discriminant de l'ALDA ($N = 10$). (b) Gains MAP par rapport au score sans sélection pour les requêtes de l'ensemble A (3, 4, 7, 11, 16, 17) et de l'ensemble B (9, 13, 21, 25).*

STD_C (figure en haut à droite). Nous notons qu'en moyenne sur toutes les requêtes $NHsurf$ et $Mean_C$ sont les catégories de traits visuels les plus choisies. Ensuite, c'est $NHhor$. Il est intéressant de noter que les traits d'écarts-types STD_C sont peu

sélectionnés, et que $NHvert$ (l'analyse orthogonale à $NHhor$) semble être peu efficace, peut-être car elle intègre des bandes trop larges. Dans le bas de la figure 4, nous détaillons la distribution des catégories de traits visuels pour les 10 traits visuels les plus discriminants de chaque requête, pour les requêtes qui ont un fort MAP en Visuel seul ($t = 0, MAP > 0.4$: 4 citron, 6 drapeau éthiopien, 7 drapeau européen, 9 la Joconde, 11 la Liberté guidant le peuple de Delacroix, 18 zèbre, 23 tournevis). Nous voyons que les différentes catégories de traits sélectionnés varient fortement d'une requête à l'autre. Les différences de catégories de traits sélectionnés entre les requêtes 6 (drapeau éthiopien) et 7 (drapeau européen) peuvent venir de la différence d'orientation des drapeaux.

5.5. *Relations entre les MAP obtenus et le nombre de données du web utilisées*

Nous montrons dans la figure 5(b) les requêtes qui ont une amélioration du MAP quand R augmente, en comparaison d'autres pour lesquelles l'effet est l'inverse. D'autres requêtes qui ne sont pas données ici n'ont pas de variation du MAP significative. Les courbes de la figure 5(b) montrent que les gains moyens augmentent ou diminuent en fonction de R, mettant en évidence deux différents types de comportement des requêtes. Les requêtes de l'ensemble A peuvent être considérées comme étant des requêtes visuellement dépendantes. Plusieurs explications peuvent être données pour le comportement des requêtes de l'ensemble B : ces requêtes ne sont pas visuellement dépendantes, les résultats des requêtes textuelles comportent plus d'erreurs (l'indexation textuelle effectuée par le moteur de recherche pour ces requêtes est plus mauvaise du fait de la difficulté de la requête), ces requêtes ont plusieurs sens (polysémie)... Pour améliorer les requêtes de l'ensemble B, on peut envisager d'utiliser d'autres sources d'information que l'information visuelle. Il pourrait être intéressant de prédire le R optimal pour chaque requête pour utiliser l'ALDA de la manière appropriée. Pour l'ensemble A, plus on donne d'exemples à l'ALDA, plus le système est capable de découvrir les traits visuels discriminants. Pour l'ensemble B, c'est l'inverse. Cela peut signifier soit que le moteur de recherche d'images utilisé n'est pas efficace pour ces requêtes ou bien qu'il n'existe pas une représentation visuelle simple de ces requêtes.

6. Discussion et conclusion

Cet article montre, premièrement, l'intérêt de notre méthode d'optimisation en ligne pour un système de recherche visuo-textuelle d'images. Notre méthode approxime l'analyse linéaire discriminante pour adapter les dimensions visuelles utilisées pour calculer les distances visuelles en fonction de la requête visuo-textuelle. Nous montrons que des statistiques d'approximation de la LDA, calculées sur des ensembles d'images chargées à la volée sur le web, permettent de réduire automatiquement et pertinemment les traits visuels, ce qui a un impact sur les temps de calculs

et les scores du modèle (réduction du temps de calcul de 35% sans dégradation des scores de Mean Average Precison).

Nous démontrons également que des modèles simples suffisent à révéler des performances très variables en fonction du taux de fusion des informations textuelles et visuelles : courbes de performances monotoniques ou en cloches. Ce dernier type de comportement démontre clairement la complémentarité des deux informations pour un nombre significatif de requêtes, et donc l'intérêt des requêtes bimodales et de l'optimisation de la fusion visuo-textuelle.

D'autre part, si nous optimisons les paramètres N, R and t pour chaque requête, nous obtenons un MAP de 0.67, ce qui est un score significativement meilleur que le meilleur modèle présenté à ImagEVAL. Nous en déduisons que pour obtenir un modèle efficace une bonne stratégie est d'estimer les paramètres optimaux, plutôt que de générer des traits visuels complexes ce qui serait long en calculs. Nos prochains travaux seront conduits dans cette direction afin de déterminer automatiquement les paramètres optimaux. Une piste pour cela est d'utiliser un ensemble de développement extrait de pages web comme nous l'avons fait pour l'utilisation de l'ALDA.

Remerciements

Ces travaux sont soutenus par l'ANR AVEIR (ANR-06-MDCA-002). Nous remercions Pierre-Alain Moëllic (CEA LIST) pour l'organisation de la campagne ImagEval.

7. Bibliographie

Amsaleg L., Gros P., Berrani S.-A., « Robust Object Recognition in Images and the Related Database Problems », *Multimedia Tools and applications*, vol. 23, n° 3, p. 221-235, 2004.

Barnard K., Forsyth D., « Learning the Semantics of Words and Pictures », *International Conference on Computer Vision (ICCV)*, vol. 2, p. 408-415, 2001.

Beyer K. S., Goldstein J., Ramakrishnan R., Shaft U., « When Is "Nearest Neighbor" Meaningful ? », *International Conference on Database Theory (ICDT)*, Springer-Verlag, p. 217-235, 1999.

Coelho T. A. S., Calado P. P., Souza L. V., Ribeiro-Neto B., Muntz R., « Image Retrieval Using Multiple Evidence Ranking », *IEEE Knowledge and Data Engineering*, p. 408-417, 2004.

Glotin H., Tollari S., Giraudet P., « Shape reasoning on mis-segmented and mis-labeled objects using approximated Fisher criterion », *Computers & Graphics*, vol. 30, n° 2, p. 177-184, April, 2006.

ImagEVAL, « Nicephore days : ImagEVAL International results symposium », 2006. http ://www.imageval.org.

La Cascia M., Sethi S., Sclaroff S., « Combining Textual and Visual Cues for Content-based Image Retrieval on the World Wide Web », *IEEE Workshop on Content-based Access of Image and Video Libraries*, 1998.

Lavrenko V., Manmatha R., Jeon J., « A Model for Learning the Semantics of Pictures », *Neural Information Processing Systems (NIPS)*, 2003.

Li J., Wang J. Z., « Automatic Linguistic Indexing of Pictures by a Statistical Modeling Approach », *IEEE Pattern Analysis and Machine Intelligence (PAMI)*, vol. 25, n° 9, p. 1075-1088, 2003.

Salton G., Buckley C., « Term-weighting approaches in automatic retrieval », *Information Processing and Management*, vol. 24, n° 5, p. 513-523, 1988.

Sclaroff S., Taycher L., Cascia M. L., « ImageRover : A Content-Based Image Browser for the World Wide Web », *IEEE Workshop on Content-based Access of Image and Video Libraries*, 1997.

Srihari R. K., « Automatic Indexing and Content-Based Retrieval of Captioned Images », *IEEE Computer*, vol. 28, n° 9, p. 49-56, 1995.

Tollari S., Glotin H., « LDA versus MMD Approximation on Mislabeled Images for Keyword Dependant Selection of Visual Features and their Heterogeneity », *IEEE ICASSP*, vol. 2, May, 2006.

Tollari S., Glotin H., Le Maitre J., « Enhancement of Textual Images Classification using Segmented Visual Contents for Image Search Engine », *Multimedia Tools and Applications*, vol. 25, n° 3, p. 405-417, 2005.

Yanai K., Barnard K., « Image region entropy : a measure of "visualness" of web images associated with one concept », *ACM Multimedia*, p. 419-422, 2005.

Zhou X. S., Huang T. S., « Unifying Keywords and Visual Contents in Image Retrieval », *IEEE Multimedia*, vol. 9, n° 2, p. 23-33, 2002.