Rehaussement de la classification textuelle d'images par leur contenu visuel

Enhancement of textual images classification using their visual content

Sabrina Tollari, Hervé Glotin, Jacques Le Maitre
Laboratoire SIS - Equipe informatique
Université de Toulon et du Var
Bâtiment R, BP 20132
F-83957 La Garde cedex
{tollari,glotin,lemaitre}@univ-tln.fr

Résumé

Cet article décrit une expérience ayant pour objectif de tester l'existence d'une cohérence entre l'indexation textuelle (un ensemble de mots-clés) d'une image et son indexation visuelle (attributs de couleurs et de formes). Cette expérience a été menée sur un corpus de photos de presse indexées manuellement par un ensemble de mots-clés extraits d'un thésaurus structuré hiérarchiquement. Elle a consisté à établir une classification de référence de ces photos à partir de leur indexation textuelle, considérée comme pertinente, puis à construire des indices textuels et visuels caractérisant ces classes et enfin à utiliser ces indices pour évaluer les performances obtenues par une recherche d'images combinant description textuelle et description visuelle. Nous obtenons par cette fusion 54% de gain de classification par rapport à l'information textuelle seule. Enfin, nous discutons d'une application sur un moteur de recherche d'images.

Mots clés

Recherche d'information, recherche d'images par le contenu, classification, modèle vectoriel, indexation, multimédia, distance de Kullback-Leibler.

Abstract

This paper deals with the existence of a dependance between the textual indexation (a set of keywords) of an image and its visual indexation (color and shape attributes). This experience has been realized on a corpus of news photos manually indexed by keywords extracted from a hierarchically structured thesaurus. First, a reference classification of these photos has been constructed from their textual indexation (regarded as relevant), then textual and visual features characterizing these classes have been constructed. Finally, they have been used to evaluate performances of a content-based image retrieval combining textual and visual description. Results of the visual-textual classification show an improvement of 54% against classification of textual information. Finally, we discuss on an application to an image search engine.

Keywords

Information retrieval, content-based image retrieval, classification, vectorial model, indexation, multimedia, Kullback-Leibler distance.

1 Introduction

La recherche d'information dans les textes a maintenant atteint une certaine maturité. Plusieurs modèles sont disponibles dont les performances et les limites sont bien connues [2]. Parmi ceux-ci le modèle vectoriel [16] est l'un des plus utilisés car il permet une interrogation souple basée sur une mesure de similarité qui permet de classer les réponses par ordre de pertinence. La recherche d'information dans les images est une discipline plus jeune. De nombreux systèmes ([7], [4], [14], [10], [13]) ont été développés dont la plupart sont basés sur une mesure de similarité entre une image requête et une image du corpus interrogé : similarité de couleurs (la plus utilisé car la plus simple à mettre en oeuvre), de formes, de texture, etc. D'autre part, les images peuvent être indéxées textuellement à partir de la légende de l'image ou du texte qui l'entoure, si cette image est insérée dans un document (« Google ») . Cependant, les performances obtenues ne sont pas vraiment satisfaisantes, sauf dans le cas de corpus très ciblés.

Pour améliorer ces performances une solution consiste à combiner l'indexation visuelle des images avec leur indexation textuelle. Dans cette optique, un découpage

par région des images peut être étiqueté par mots clés, mais ces systèmes de fusions visuo-textuelles en sont à leur début [3]. Cet article décrit une expérience montrant la cohérence entre l'indexation textuelle d'une image et son indexation visuelle. Ce système pourrait être utilisé dans le cas d'un filtrage visuel de requêtes d'images par mots clés comme nous le discuterons en dernière partie. Par exemple, une requête textuelle 'femme' et 'ouvrière' pourrait donner des images de femmes travaillant, mais aussi des logos d'usines, des graphiques sur la population ouvrière, alors que l'utilisateur désire seulement ces premières.

Le corpus sur lequel nous avons travaillé est constitué d'un ensemble de 665 photos de presse, mises à notre disposition par la société Editing et indexées manuellement par les documentalistes de cette société, par un ensemble de mot-clés extraits d'un thésaurus structuré hiérarchiquement. Notons que dans le cadre du projet RNTL Muse, nous avions élaboré une interface d'interrogation de ce corpus [5], dont l'expérimentation nous a convaincu de l'intérêt de chercher à combiner indexation textuelle et visuelle. L'indexation textuelle des images est réalisée suivant le modèle vectoriel et l'indexation visuelle est basée sur le découpage d'une image en plusieurs zones d'intérêt et sur la prise en compte, globale ou locale, de la luminance, des couleurs et des contours.

Cet article est organisé de la façon suivante : le paragraphe 2 présente le protocole expérimental, le paragraphe 3 décrit la construction de la base de référence, le paragraphe 4 présente les résultats d'une classification supervisée sur les indices textuels seuls, le paragraphe 5 présente les résultats d'une classification supervisée sur les indices visuels seuls, le paragraphe 6 décrit les résultats de fusion tardive textuelle-visuelle et enfin le paragraphe 7 montre une application sur un moteur de recherche et dresse des perspectives.

2 Protocole expérimental

Il s'agit de construire un système de classification visuo-textuelle permettant d'améliorer la qualité des résultats d'une recherche d'images exprimées par un ensemble de mots clés, en exploitant le contenu visuel de ces images. A chaque image, on associe des descripteurs (ou indices) textuels et visuels. Puis, on les classe par classification ascendante hiérarchique afin d'obtenir un classement par rapport aux indices textuels seulement. La construction de la base de référence B_{Ref} est expliquée à la section 3. Ensuite, on sépare la base obtenue en deux parties : une base d'exemples classés (sous-base de référence) B_{Ex} et une base de test B_{Test}. Pour cela, on choisit aléatoirement 50% des images de chaque classe de B_{Ref} pour constituer B_{Test}, les autres images constituant la sous-base de référence B_{Ex} dont on connait la classe. On cherche à retrouver la classe de chaque image de B_{Test} par simple similarité[1] au sens DKL[2] avec les images de la base B_{Ex}. La performance de cette classification est évaluée par rapport au nombre d'images de B_{Test} qui sont bien classées. Notre objectif étant de mesurer les apports de l'indexation textuelle et de l'indexation visuelle, trois types de classifications ont été réalisées et comparées : textuelle qui ne tient compte que des mot-clés, visuelle qui ne tient compte que des indices visuels et visuo-textuelle qui est une fusion des deux.

3 Construction d'une base de référence par classification ascendante hiérarchique

Dans cette première étape, il s'agit de construire une classification des images à partir de leur indexation textuelle uniquement. Cette classification constituera une base de référence pour valider les indices de similarité textuels et visuels qui seront proposés par la suite. Pour réaliser cette classification, nous avons choisi d'utiliser la méthode classique de Classification Ascendante Hiérarchique (CAH) [9].

Pour mettre en oeuvre cette classification, il faut disposer de trois composantes : (i) une représentation textuelle des images, nous avons choisi le modèle vectoriel, (ii) une mesure de similarité qui permet de comparer les images, (iii) un critère d'agrégation qui permet de fusionner les classes.

Soit $D = \{d_1, d_2, \ldots, d_m\}$ un ensemble de documents et $T = \{t_1, t_2, \ldots, t_n\}$ un ensemble de mots clés indexant ces documents, dans le modèle vectoriel ([15],[17], [16], [2]) un document d_i est décrit par un vecteur :

$$\vec{d_i} = (\omega_{1,i},\ \omega_{2,i},\ \ldots,\ \omega_{j,i},\ \ldots,\ \omega_{n,i})$$

où $\omega_{j,i}$ est le poids du terme t_j dans le document d_i. La formule la plus classique pour calculer le poids est la suivante) :

$$\omega_{j,i} = tf_{j,i} \times \log \frac{m}{m_j}$$

où $tf_{j,i}$ est la fréquence du mot clé t_j dans le document d_i et m_j le nombre de documents du corpus indexés par le mot clé t_j.

Dans notre application, les documents sont des images (des photos) et les mots clés appartiennent à un thésaurus. Chaque image est décrite (indexée) par un ensemble de mots clés. Un mot clé est donc soit présent une seule fois, dans la description d'une image, soit en est absent. On a donc $\omega_{j,i} \in \{0, 1\}$.

[1] Plusieurs métriques sont envisageables et seraient plus ou moins équivalentes.

[2] La divergence entre deux distributions de probabilité d et g est donnée par l'entropie relative de Kullback-Leibler : $KL(d, g) = \sum_{y \in \chi} d(y) \log \frac{d(y)}{g(y)}$. La distance de Kullback-Leibler est $DKL(d, g) = KL(d, g) + KL(g, d)$.

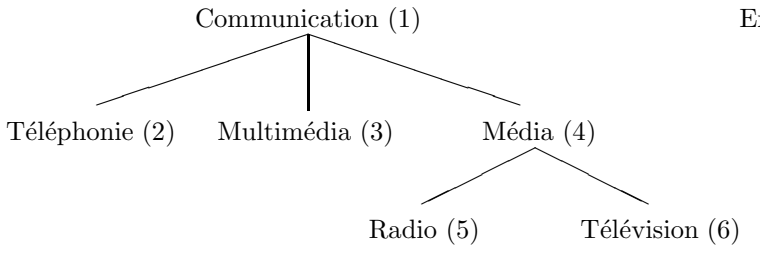

FIG. 1 – Extension d'un vecteur relativement à un thésaurus

De plus, le thésaurus est structuré hiérarchiquement par une relation de généricité ($\prec$) qui implique que si une image est indexée par un mot-clé t_j et que $t_j \prec t_k$ alors cette image est aussi indexée par le mot clé t_k. Il faut donc, comme dans [11], étendre le vecteur $\vec{d_i} = (\omega_{1,i},\ \omega_{2,i},\ \ldots,\ \omega_{j,i},\ \ldots,\ \omega_{n,i})$ d'une image de façon à ce que $\forall j,k \in [1,n]$, $\omega_{k,i} = 1$ si $\omega_{j,i} = 1$ et $t_j \prec t_k$.
Considérons le thésaurus et l'indexation de l'image d_3 présenté dans la figure 1. Le vecteur $\vec{d_3}$ initial est $(0,1,0,0,1,0)$. Puisque *Téléphonie* $\prec$ *Communication*, on a $\omega_{1,3} = 1$, et puisque *Radio* $\prec$ *Média* $\prec$ *Communication*, on a $\omega_{4,3} = 1$ et $\omega_{1,3} = 1$. Le vecteur étendu de l'image d_3 est donc $(1,1,0,1,1,0)$. Les vecteurs étendus des images d_1 et d_2 sont obtenus de façon similaire.
Dans le modèle vectoriel, une mesure classique de similarité entre un document d et une requête q est le cosinus de l'angle de leurs vecteurs. Nous avons adopté une formule analogue pour mesurer la similarité de deux images d_k et d_l :

$$cos(\vec{d_k}, \vec{d_l}) = \frac{\sum_{j=1}^{n} \omega_{j,k} \times \omega_{j,l}}{\sqrt{\sum_{j=1}^{n} \omega_{j,k}^2} \times \sqrt{\sum_{j=1}^{n} \omega_{j,l}^2}}$$

où $\omega_{j,k}$ et $\omega_{j,l} \in \{0,1\}$.
Pour la classification, c'est la distance entre deux images que l'on a besoin de connaître. Nous la calculons par la formule :

$$dist(d_k, d_l) = 1 - |cos(\vec{d_k}, \vec{d_l})|.$$

Deux images entièrement similaires ont une distance égale à 0 et deux images entièrement dissimilaires ont une distance égale à 1. Par exemple, si l'on considère les images d_1, d_2 et d_3 de la figure 1, on a $sim(d_1, d_2) = 0.33$, $sim(d_2, d_3) = 0.25$ et $sim(d_1, d_3) = 0.25$.
A chaque étape d'une classification ascendante hiérarchique, on agrège les deux classes C_p et C_q qui ont une distance $D(C_p, C_q)$ minimum. Il existe plusieurs formules pour calculer cette distance D, nous avons tout d'abord expérimenté les trois plus classiques et les avons calculées relativement à l'hétérogénéité numérique et sémantique des classes obtenues. La première est la distance du plus proche voisin :

$$D(C_p, C_q) = \min\{dist(i,j); i \in C_p, j \in C_q\}.$$

L'inconvénient est que les classes les plus peuplées sont les plus attractives. On obtient donc quelques classes contenant beaucoup d'images, et beaucoup de classes n'en contenant que très peu. La deuxième est la distance du diamètre maximum (ou du voisin le plus éloigné) :

$$D(C_p, C_q) = \max\{dist(i,j); i \in C_p, j \in C_q\}.$$

Les classes obtenues sont numériquement plus homogènes, mais sémantiquement très hétérogènes. La troisième est la distance moyenne :

$$D(C_p, C_q) = \frac{\sum_{i,j}\{dist(i,j); i \in C_p, j \in C_q\}}{Card(C_p) \times Card(C_q)}.$$

Cette distance donne des résultats équivalents à celle du plus proche voisin.

Ces résultats étant peu satisfaisant, nous avons cherché un compromis entre la méthode du plus proche voisin et celle du diamètre maximum. Ce compromis a été d'utiliser le diamètre maximum, mais au lieu de prendre la distance maximale, nous avons pris la plus grande distance inférieure à un certain seuil que nous avons déterminé pour nos images de manière empirique[3] à 0.7, et que nous appelons diamètre maximum contraint.
Il restait à fixer la condition d'arrêt A de la classification. Nous avons déterminé de manière empirique que pour obtenir des classes représentatives, il fallait arrêter la classification lorsque la distance d'agrégation obtenue était de 0.55. On a ainsi obtenu 57 classes, contenant en tout 665 images.
Une dernière opération a consisté à supprimer les classes dont toutes les images étaient indexées par les mêmes mots clés, ainsi que les classes contenant moins de 8 images. Au final, la base de référence obtenue

[3]Le paramètre est resté valable sur une expérience menée sur une autre base [18].

Classe	T_{f_1}	T_{f_2}	T_{f_3}
1	Mexique	Politique	Portrait
2	Israël	Judaïsme	Patrimoine
3	Constructeurs	Transport	Automobile
4	Contemporaine	Portrait	Rhône
5	Portrait	Armée de l'air	Aéronautique
6	Société	Famille	Enfant
7	Cameroun	Agriculture	Géographie physique
8	Municipalité	Portrait	Les Verts
9	Elevages	Santé	Police national
10	Portrait	Média	Administrations
11	Femme	Ouvriers	Industrie de précision
12	Région	Municipalité	Conseil régionaux
13	Communication	Télécommunications	Multimédia
14	Production	Travail	Alimentation
15	Israël	Liban	Urbanisation
16	Parti socialiste	Portrait	Municipalité
17	Multimédia	Star'up	Ouvriers
18	Jeux de société	Humain	Librairies
19	Problèmes sociaux	Conflits sociaux	Europe
20	Politique	Paris	Bourse
21	Bars et Café	Restauration rapide	Etats-Unis
22	Infrastructures routières	Innondation	Véhicules
23	Portrait	Municipalité	RPR-UMP
24	Justice	Portrait	Scandales politiques

TAB. 1 – Liste des 3 premiers mot-clés les plus fréquents ($f_1 > f_2 > f_3$) de 12 classes

contient 517 images réparties dans 24 classes. Le tableau 1 donne les termes les plus fréquents de chaque classe.

Algorithme Classification ascendante hiérarchique
Données :
E : ensemble de n éléments à classer
Tableau $n \times n$ des distances entre éléments
Variables :
C : ensemble des c classes
Début
Pour chaque individu e de E **faire**
Créer une classe dans C contenant e
fin pour
Tant que non(A) **faire**
Pour chaque couple (C_p, C_q) de classes de C
Calculer la distance entre C_p et C_q
pour le critère d'agrégation considéré
fin pour
Agréger les deux classes C_a et C_b
de distance minimale
fin tant que
Fin

4 Classification textuelle

La base de référence pour notre corpus d'images étant construite, nous allons maintenant tester un système de classification automatique travaillant avec les indices visuels et/ou les indices textuels. Notons tout d'abord que le score d'un système aléatoire[4] est de 91.6% (en prenant en compte la fréquence de chaque classe).
Une première expérience consiste à tester la base de référence obtenue par CAH. Chaque classe C_k de B_{Ex} est représentée par un vecteur moyen textuel $\vec{C_k^t}^*$ normalisé obtenu en faisant la somme des vecteurs textuels des images qu'elle contient. La classe textuelle d'une image d_T de B_{Test} de vecteur textuel normalisé $\vec{d_T^t}^*$ est calculée par :

$$C^t(d_T) = \text{argmin}_{k\in\{1,2,\ldots,c\}} DKL(\vec{d_T^t}^*, \vec{C_k^t}^*)$$

où c est le nombre de classes de la base de référence. Deux tests ont été réalisés : le premier en étendant les

[4]Soit P_k la fréquence de la classe C_k dans la classification, le score du système aléatoire est calculé par :

$$TE_a = 1 - \sum_{k=1}^{c} (P_k)^2 = 1 - \sum_{k=1}^{c} \left(\frac{card(C_k)}{\sum_{k=1}^{c} card(C)} \right)^2$$

où c est le nombre de classes et $card(C_k)$ est le nombre d'images de la classe C_k.

vecteurs textuels à l'aide du thésaurus, le deuxième en utilisant des vecteurs non-étendus. Le tableau 2 donne les taux d'erreurs obtenus. Nous remarquons

Textuelle avec thésaurus	Textuelle sans thésaurus	Système aléatoire
1.17	13.72	91.6

TAB. 2 – Comparaison des taux d'erreurs textuelles (en %)

que lorsque les vecteurs sont étendus, les résultats donnent un taux d'erreur très faible. Ceci démontre que la description des images et la procédure de classification utilisées sont efficaces. Mais en pratique la qualité du thésaurus influence nettement les résultats de classification (une expérience réalisée sur la base Corel [12] avec un thésaurus construit d'après Wordnet [8] montre que l'utilisation du thésaurus n'améliore pas les scores de classification textuelle [18] : 17% de taux d'erreurs conter 18% avec le thesaurus). Nous nous plaçons donc dans ce cas réel en n'étendant pas les vecteurs avec l'information du thésaurus. Une autre limitation de l'usage du thésaurus est donnée lors du couplage de noter système avec un moteur de recherche d'image dont on ne possède pas le code (voir l'application section 7).

5 Classification visuelle

Une deuxième expérience consiste à établir, de même que pour la classification textuelle, une classification supervisée mais avec les indices visuels seuls. Nous discuterons tout d'abord du choix des indices visuels, puis nous présenterons les différents résultats des classifications visuelles.

5.1 Indices visuels

Parmi les nombreux attributs visuels envisageables (texture, forme, spectre, ...), nous choisissons ceux qui sont les moins coûteux en calculs. Nos indices visuels sont composés de 5 attributs :

- l'histogramme de la luminance (V_1),
- les 3 histogrammes des couleurs rouge (V_2), vert (V_3), bleu (V_4), normalisés par la luminance (indiquant donc les composantes absolues de chaque couleur),
- l'histogramme des directions des contours (V_5). Pour obtenir ce dernier histogramme, on commence par extraire les contours par la méthode des gradients maximum (méthode de Canny [6] ; des méthodes plus lourdes seraient envisageables [1]). L'image « edges » de la figure 2 donne un exemple de matrice binaire de contours. Ensuite, on « fait passer » (convolution) une à une 6 matrices carrées de 7 pixels de côté sur la matrice de contours. Ces matrices contiennent des coefficients qui codent un segment dont la pente varie de $-\pi/2$ pour la première matrice à $\pi/3$ pour la dernière, par pas de $\pi/6$ avec une tolérance de $\pm\pi/12$. Ce système permet de détecter, de classer et de compter les traits des contours selon leur pente. Par exemple, dans l'histogramme de direction de la figure 2, les segments de pente $\pi/2$ (trait vertical) et π (trait horizontal) sont les plus représentés. Ces pentes caractérisent classiquement les bâtiments.

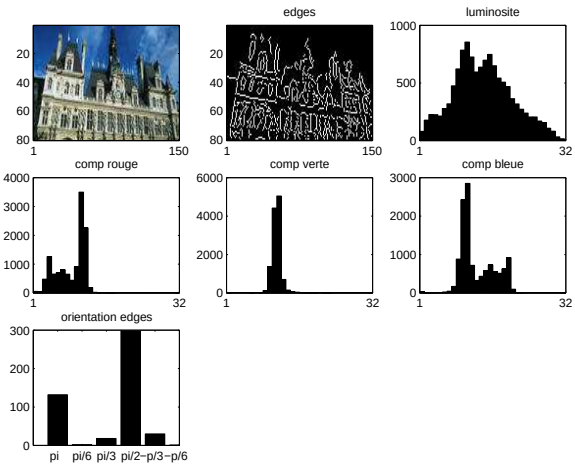

FIG. 2 – Indices visuels sous la forme d'histogrammes. Photo @Editing.

Ces attributs visuels sont extraits pour les images complètes (la région globale est notée r_0).

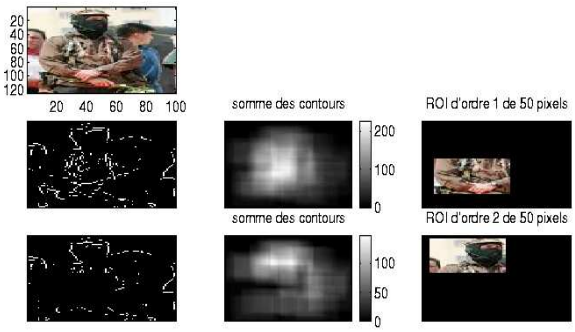

FIG. 3 – Sélection des 2 premières régions locales d'intérêts(ROI) d'une image par détection des contours par la méthode de Canny et par maximisation des sommes des contours par région. Photo ©Editing.

De plus, nous avons testé une méthode originale d'extraction de sous-images d'intérêt pour lesquelles nous calculons aussi les attributs visuels. En effet, pour chaque image, 4 sous-images sont détectées automatiquement. L'algorithme de détection commence par extraire les contours de l'image par la méthode de Canny

comme précédemment. Puis, il fait la sommation de ces contours par région de dimension fixée. Nous avons choisi comme dimension[5] une surface d'un quart de la surface de l'image globale.
Ensuite, on extrait la région qui contient le plus de contours et on la soustrait de la matrice des contours. Enfin, la détection d'une nouvelle région d'intérêt est relancée sur la nouvelle matrice des contours. La figure 3 montre la détection automatique des 2 premières régions d'intérêts (ROI). On numérote ces régions de r_1 à r_4 selon leur ordre de détection.
L'intérêt de l'étude des histogrammes de couleurs de sous-images est de classer ensemble des images similaires. Par exemple, détecter les images contenant un visage grâce à la couleur de la peau, sans être bruitée par le fond de l'image.
Au final, les indices visuels associés à l'image se présentent sous la forme de vecteurs de flottants (les histogrammes) qui permettent des calculs simples et rapides entre deux régions de l'image par simple mesure de similarité au sens DKL des vecteurs. Un grand nombre de combinaisons possibles a été expérimenté pour choisir les meilleurs distances visuelles, nous présentons celles qui donnent les meilleurs résultats.
On note $\mathrm{DKL}_{V_A}(r_i, r_j)$ la distance DKL entre la région r_i de l'image d_T de B_{Test} et la région r_j de l'image d_E de B_{Ex} pour l'attribut visuel V_A.

5.2 Distance entre régions de même ordre

Nous commençons par calculer la distance entre la région r_i de l'image de la base de test et la région d'ordre égal r_i de chacune des images de la base d'exemples (table 3).
On remarque que, en général, les distances sur les indices globaux sont meilleurs, sauf pour la direction où la région 1 donne de meilleurs résultats. En effet, la région 1 est celle qui contient le plus de contours, elle est donc la plus significative. Pour l'attribut vert, le bon résultat obtenu pour la région 2 s'explique par un artefact du aux données (une classe contenant plus de vert que les autres). L'hypothèse de départ supposant que les régions locales les plus descriptives sont celles qui contiennent le plus de contour est vérifiée, car les régions 1 et 2 ont les plus faibles taux d'erreur.

5.3 Distances par fusion précoce des indices visuels

Pour un attribut V_A donné, chaque image possède 5 histogrammes. Pour une image d_T de B_{Test} et pour une image d_E de B_{Ex}, il existe donc 5×5 distances entre régions de l'image possibles. Si l'on considère seulement les $L \in [1, 5]$ régions d'intérêt, il existe $L \times L$ distances entre régions de l'image possibles (si $L = 2$, $L^2 = 4$ et on ne considère que les distances $\mathrm{DKL}_{V_A}(r_1, r_1)$, $\mathrm{DKL}_{V_A}(r_1, r_2)$, $\mathrm{DKL}_{V_A}(r_2, r_1)$ et $\mathrm{DKL}_{V_A}(r_2, r_2)$). Nous allons définir une distance entre les indices visuels de deux images qui prend en compte les meilleurs scores parmi ces distances. Pour les besoins du calcul de ces distances, on note $\mathrm{moymin_Z}$ la fonction :

$$\mathrm{moymin_Z} : \{\alpha_1, \alpha_2, \ldots, \alpha_M\} \rightarrow (\alpha_{min1} + \alpha_{min2} + \ldots + \alpha_{minZ})/Z$$

qui fait la moyenne arithmétique des Z premières valeurs minimales. La fonction $moymin_Z$ permet de rejeter les comparaisons aux images de référence trop différentes de l'image de test et de rejeter les imposteurs.
Pour calculer la distance visuelle entre une image d_T de B_{Test} et une image d_E de B_{Ex}, on calcule les L^2 distances possibles entre 2 images, puis la moyenne des N plus petites valeurs ($N \in [1, L^2]$), on obtient la distance :

$$\gamma_{V_A}(d_T, d_E) = \mathrm{moymin_N}(\{DKL_{V_A}(i, j); \forall i, j \in [1, L]\}).$$

Maintenant, si on considère la distance entre une image d_T de B_{Test} et la classe C_k, on calcule les distances entre d_T et les images d_{E_k} de C_k et on garde les I minimums dont on calcule la moyenne pour obtenir la distance entre l'image d_T et la classe C_k :

$$\delta_{V_A}(d_T, C_k) = \mathrm{moymin_I}(\{\gamma_{V_A}(d_T, d_{E_k}); \forall d_{E_k} \in C_k\})$$

où d_{E_k} est un élément de la classe C_k de la base d'exemples et $I \in [1, \mathrm{card(C_k)}]$ est le nombre de valeurs minimales prises parmi les $\mathrm{card(C_k)}$ distances entre d_T et les éléments de la classe C_k possibles. La classe visuelle de d_T pour l'attribut V_A est obtenue par :

$$C^v_{V_A}(d_T) = \mathrm{argmin}_{k \in \{1,2,\ldots,c\}} \delta_{V_A}(d_T, C_k).$$

5.4 Résultats de la fusion précoce visuelle

Les tableaux 4, 5 et 6 donnent les taux d'erreur obtenus par cette méthode dite de « fusion précoce » des indices visuels en faisant varier les paramètres N, I et L. Le tableau 4 donne l'influence du paramètre N pour les valeurs de I et L donnant les meilleurs résulats. On remarque que le paramètre N a peu d'influence pour les attributs Rouge, Vert, Bleu et Luminance. Par contre, pour la direction, on observe une réelle amélioration du T.E. quand on prend N grand. Le tableau 5 montre qu'il vaut mieux regarder si l'image test est similaire à plusieurs images d'une même classe qu'à une seule. Enfin, dans le tableau 6, on remarque que la région d'intérêt 1, seule, n'est pas

[5] Le nombre et la surface des sous-images pourraient être optimisés suivant le critère de dispersion des contours dans l'image globale.

	DKL(r_1, r_1)	DKL(r_2, r_2)	DKL(r_3, r_3)	DKL(r_4, r_4)	DKL(r_0, r_0)
T.E. Rouge	81.17	79.21	81.17	82.35	**73.33**
T.E. Vert	83.13	**78.03**	86.66	80.78	78.43
T.E. Bleu	82.35	80.39	83.92	84.70	**74.50**
T.E. Luminance	80.39	81.17	81.56	83.52	**76.40**
T.E. Direction	**79.60**	81.56	80.00	84.31	85.49

TAB. 3 – Influence du choix de la région d'intérêt sur le Taux d'Erreur(T.E. en %) pour les différents attributs de l'image

N	1	2	3	4	5	6	7	8
T.E. Rouge	**71.76**	72.54	72.54	73.72	76.47	77.64	77.64	76.07
T.E. Vert	**76.07**	77.64	77.64	76.86	76.86	76.47	78.82	78.82
T.E. Bleu	77.64	**77.25**	79.60	80,00	79.60	81.56	81.96	81.96
T.E. Luminance	**77.64**	79.21	77.64	77.64	79.21	79.21	78.82	78.03
T.E. Direction	83.52	80.39	80.39	80,00	79.21	78.82	78.43	**76.86**

TAB. 4 – Taux d'Erreur(T.E. en %) pour différentes valeurs de N et pour les différents attributs par fusion précoce des indices visuels ($I = 4, L = 5$)

I	1	2	3	4
T.E. Rouge	75.68	74.50	**71.76**	**71.76**
T.E. Vert	79.60	78.03	76.86	**76.07**
T.E. Bleu	78.03	77.64	78.03	**77.25**
T.E. Luminance	79.21	78.03	**76.07**	77.64
T.E. Direction	84.70	78.03	**76.86**	**76.86**

TAB. 5 – Taux d'Erreur(T.E. en %) pour différentes valeurs de I, et pour les valeurs de N pour lesquels le taux d'erreur est le plus faible par fusion précoce des indices visuels des différents attributs ($L = 5$)

L Dimension L^2	1 1	2 4	3 9	4 16	4+g 25
T.E. Rouge	81.17	78.82	76.07	76.07	**71.76**
T.E. Vert	83.13	78.82	**75.68**	79.60	76.07
T.E. Bleu	82.35	80.00	79.60	81.56	**77.25**
T.E. Luminance	80.39	79.60	78.03	**77.64**	**77.64**
T.E. Direction	79.60	78.03	**76.07**	76.47	76.86

TAB. 6 – Taux d'Erreur(T.E. en %) pour différentes valeurs de L, et pour les valeurs de N pour lesquels le taux d'erreur est le plus faible par fusion précoce des indices visuels des différents attributs ($I = 4$)

suffisante($L = 1$) et que la région d'intérêt numéro 4 n'apporte finalement pas d'information, car les T.E. pour $L = 4$ sont plus grand que pour $L = 3$. On remarque aussi que, pour $L = 5$ (les 4 ROI plus l'image globale), les indices globaux apportent une nette amélioration du T.E., sauf dans le cas de la direction, ce qui était prévisible.
Si on compare ces résultats à ceux du tableau 3, on remarque que le gain apporté par la fusion précoce des indices visuels et par l'utilisation de régions d'intérêts locales est négligeable, sauf pour la direction (gain 10%).

6 Classification visuo-textuelle

Nous allons maintenant fusionner les indices textuels et visuels afin d'améliorer les résultats obtenus pour la classification textuelle.
Pour chaque image d_T et pour chaque classe C_k, on calcule la distance textuelle $DKL(\vec{d_T^t}^*, \vec{C_k^t}^*)$ comme expliqué à la section 4. Puis, on la normalise et on la complète à 1 pour estimer la probabilité d'appartenance $P^t('d_T \in C'_k|t)$ de l'image d_T à la classe C_k par rapport aux indices textuels :

$$P^t('d_T \in C'_k|t) = 1 - \frac{DKL(\vec{d_T^t}^*, \vec{C_k^t}^*)}{\sum_k DKL(\vec{d_T^t}^*, \vec{C_k^t}^*)}.$$

De même, on estime la probabilité d'appartenance $P^v('d_T \in C'_k|V_A)$ de l'image d_T à la classe C_k par rapport à l'attribut visuel V_A :

$$P^v('d_T \in C'_k|V_A) = 1 - \frac{\delta_{V_A}(d_T, C_k)}{\sum_k \delta_{V_A}(d_T, C_k)}.$$

On numérote de 1 à 5 les attributs visuels et on donne le numéro 6 à l'indice textuel. La probabilité d'appartenance $P^{v\vee t}('d_T \in C'_k)$ de l'image d_T à la classe C_k par fusion tardive des indices textuels et visuels est :

$$P^{v\vee t}('d_T \in C'_k) = \sum_{j=1}^{5} \omega'(V_j) \cdot P^v('d_T \in C'_k|V_j)$$
$$+\omega'(V_6) \cdot P^t('d_T \in C'_k|t)$$

où $\omega'(V_j) = \frac{\omega(V_j)^p}{\sum_{i=1}^{6} \omega(V_i)^p}$, $\omega(V_j) = \frac{1-TE(j)}{\sum_{i=1}^{6} 1-TE(i)}$, $TE(j)$ est le taux d'erreur obtenu pour l'attribut V_j. Le paramètre p est déterminé de manière empirique.
La classe d'appartenance de chaque image d_T de B_{Test} est alors celle qui maximise cette probabilité :

$$C^{v\vee t}('d_T \in C'_k) = \text{argmax}_{k\in\{1,2,\ldots,c\}} P^{v\vee t}('d_T \in C'_k).$$

La figure 4 décrit les résultats obtenus pour la fusion de la classification textuelle sans thésaurus (T.E. 13.72%) et de plusieurs classifications visuelles. Le premier résultat (T+Vis[Locaux]) est obtenu à partir des meilleures classifications par fusion précoce des locaux ($L \in [1, 4]$) uniquement. Le deuxième (T+Vis[Globaux]) considère les classifications sur les indices globaux uniquement. Le troisième (T+Vis[Locaux+Globaux]) utilise les meilleurs paramètres de fusion précoce des indices locaux et globaux ($L \in [1, 5]$). Le dernier (T+Vis[Dir+Globaux]) prend en compte les globaux pour les attributs rouge, vert, bleu et luminance, et la direction locale calculée par $DKL(r_1, r_1)$. Sur cette figure, on remarque que les locaux accélèrent le gain de classification par rapport à p, montrant donc que les poids $\omega(V_j)$ sont mieux adaptés que ceux des méthodes globales. On remarque aussi que les quatres méthodes tendent pour $p = 4$ vers le même résultat[6]. Le tableau 7 donne le gain final que l'on peut espérer du rehaussement de la classification textuelle par la classifiation visuelle.

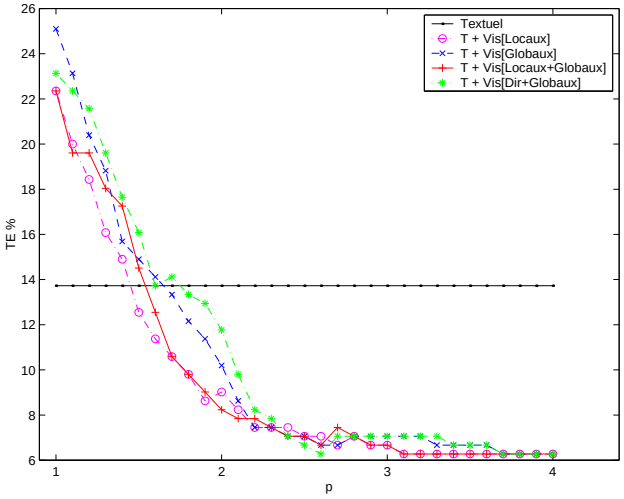

FIG. 4 – Influence de p sur le Taux d'Erreur(T.E. en %) pour la fusion tardive des probabilités textuelles (T) et visuelles (Vis) de différents indices visuels

Textuelle sans thésaurus	Fusion textuelle/visuelle	Gain
13.72	6.27	+54.3

TAB. 7 – Résultat en % du rehaussement de la classification textuelle par fusion tardive avec la classification visuelle

7 Discussions et conclusion

Nous avons présenté un système simple de mise en relation d'informations textuelles et visuelles par confrontation aux images de références de chaque

[6]Evidemment, pour p grand ($p > 8$), toutes les méthodes convergent vers le T.E. textuel.

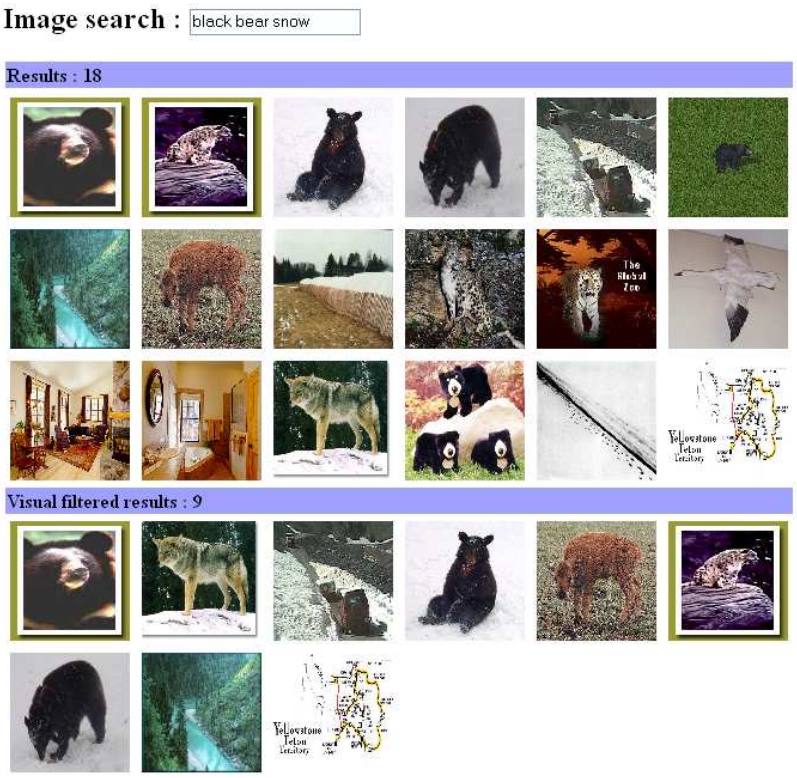

FIG. 5 – Résultats d'une recherche d'images sur Google avec les mots 'Black', 'Bear' et 'Snow'. Puis filtrage et reclassement des images en fonction des valeurs obtenues par la distance δ_{V_A} (voir figure 6).

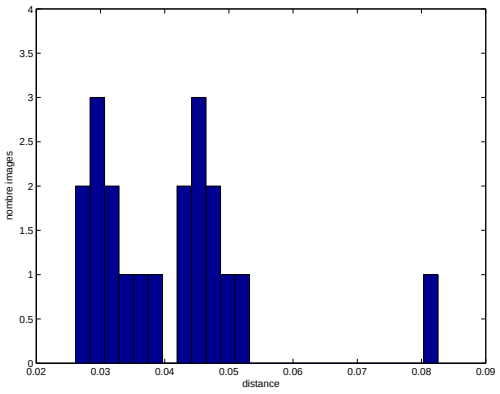

FIG. 6 – Distribution des distances δ_{V_A} pour chaqu'une des images trouvées par la requête sur Google. Cette distribution est bimodale, ce qui permet de considérer que les images du premier groupe (à gauche de 0.04) sont adéquates à la requête, à droite non.

classe textuelle. D'autres méthodes d'apprentissage automatique (réseau de neurone artificiel notamment), pourraient être mises en oeuvre pour déterminer les frontières entre tous les paramètres (visuels et textuels), mais seraient lourdes et demanderaient une base d'apprentissage beaucoup plus grande.

Nous avons montré que les informations visuelles réduisent les erreurs de classification textuelle privée d'un thésaurus de l'ordre de 50%, ce qui est très prometteur du fait de la simplicité de la méthode.

Le volume des données à noter disposition ne nous a pas permis d'optimiser certains paramètres (I, N, L, p) sur une base de développement. Cependant, nous avons relancé nos algorithmes sur une autre base avec d'autres segmentations et d'autres indices visuels afin de tester leur généralisation. Cette expérience [18] a montré un gain de l'ordre de 65%.

Notre système peut-être utilisé comme un filtre visuel rapide en s'appliquant directement sur le résultat d'une requête textuelle d'images d'un moteur de recherche (tel que « Google ») composée d'un petit nombre de mots clés, et sans usage de thésaurus dont la nature dans Google est inconnue (figure 5 et 6).

Une autre utilisation de notre système serait d'agrandir ou de créer une base de référence en associant une sémantique à une série de traits visuels. En-

	Système aléatoire	Meilleur visuelle	Textuelle sans thésaurus	Fusion visuo-textuelle
T.E. en %	91.60	71.16	13.72	6.27
Intervalles de confiance	±3.40	±5.56	±4.22	±2.98

TAB. 8 – Intervalles de confiance (95%) de quelques T.E.

fin, nous avons étudié les indices visuels par rapport à des classes textuelles. Nous pourrions inverser l'expérience en considérant les indices textuels par rapport à des classes visuelles. Cette méthode permettrait par exemple de corriger une mauvaise indexation textuelle à l'aide du contenu visuel. Ainsi, si l'image d'un graphique sur la population ouvrière a été étiqueté automatiquement par 'femme' et 'ouvrière', une comparaison avec des classes visuelles représentant des femmes montrerait l'erreur d'indexation et permettrait d'enlever le mot 'femme'.

Références

[1] O. Amadieu, E. Debreuve, M. Barlaud, and G. Aubert. Segmentation par contours actifs déformables approche bidirectionnelle. In *Actes du* 12^e *Congrès Francophone AFRIF-AFIA (RFIA)*, pages 237–244, 2000.

[2] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley, 1999.

[3] K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. Blei, and M. Jordan. Matching words and pictures. In *Journal of Machine Learning Research*, volume 3, pages 1107–1135, 2003.

[4] Marinette Bouet and Ali Khenchaf. Traitement de l'information multimédia : recherche de média image. *Ingénierie des systèmes d'information (RSTI série ISI-NIS)*, 7(5-6) : 65–90, 2002.

[5] E. Bruno, J. Le Maitre, and E. Murisasco. Indexation et interrogation de photos de presse décrites en MPEG-7 et stockées dans une base de données XML. *Ingénierie des systèmes d'information (RSTI série ISI-NIS)*, 7(5-6) : 169–186, 2002.

[6] J. Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8(6) : 679–698, 1986.

[7] V. Castelli and L. D. Bergman, editors. *Image Databases*. John Wiley & Sons, 2002.

[8] Christiane Fellbaum, editor. *WordNet - An Electronic Lexical Database*. Bradford books, 1998.

[9] G.N. Lance and W.T. Williams. A general theory of classificatory sorting strategies : I. hierarchical systems. *Computer Journal*, 9 : 373–380, 1967.

[10] Ying Li, C.C. Jay Kuo, and X. Wan. Introduction to content-based image retrieval overview of key techniques. In V. Castelli and L. D. Bergman, editors, *Image Databases*, chapter 10, pages 261–284. John Wiley & Sons, 2002.

[11] Jean Martinet, Yves Chiaramella, and Philippe Mulhem. Un modèle vectoriel étendu de recherche d'informations adapté aux images. *Actes du XXème Congrès INFORSID*, pages 337–348, 4-7 juin 2002.

[12] Henning Müller, Stéphane Marchand-Maillet, and Thierry Pun. The truth about corel – evaluation in image retrieval. In *The Challenge of Image and Video Retrieval (CIVR2002)*, 2002.

[13] C. Nastar. Indexation d'images par le contenu : un état de l'art. *Actes de CORESA'97*, 1997.

[14] W. Niblack. The QBIC project : querying images by content using color, texture and shape. *Proceedings SPIE : Storage and Retrieval for Image and Video Database*, pages 173–181, 1993.

[15] G. Salton. *The SMART Retrieval System ; Experiments in Automatic Document Processing*. Englenwood Cliffs, Prenctice-Hall, New Jersey, 1971.

[16] G. Salton and C. Buckley. Term-weighting approaches in automatic retrieval. *Information processing and management*, 24(5) : 513–523, 1988.

[17] G. Salton and M.J. Lesk. Computer evaluation of indexing and text-processing. *Journal of the ACM*, 15(1) : 8–36, 1968.

[18] S. Tollari, H. Glotin, and J. Le Maitre. Enhancement of textual images classification using segmented visual contents for image search engine. In *Multimedia Tools and Applications, to appear*. Kluwer Academic, 2004.