Utilisation de concepts visuels et de la diversité visuelle pour améliorer la recherche d'images

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

Université Pierre et Marie Curie - Paris 6, UMR CNRS 7606 - LIP6
104 avenue du président Kennedy, 75016 Paris, France, prénom.nom@lip6.fr

RÉSUMÉ. *Dans cet article, nous étudions (i) comment extraire et exploiter des concepts visuels pour améliorer la recherche d'images basée sur le texte, et (ii) comment diversifier les résultats pertinents obtenus. Nous utilisons d'abord des forêts d'arbre de décisions flous (FFDTs) pour détecter les concepts dans les images, puis nous découvrons à l'aide de l'analyse des cooccurrences des relations d'exclusion mutuelle et d'implication entre les concepts. Ensuite, nous utilisons ces concepts pour améliorer la pertinence des résultats obtenus par un système de recherche d'images par le texte. Enfin, nous appliquons une méthode de diversité visuelle basée sur le partitionnement de l'espace visuel. Ce travail se place dans le cadre de la campagne d'évaluation CLEF. Il montre une nette amélioration des résultats lorsque l'on utilise les concepts apparaissant explicitement dans la requête textuelle, ainsi que l'efficacité du clustering spatial.*

ABSTRACT. *In this article, we study (i) how to automatically extract and exploit visual concepts and (ii) fast visual diversity. First, in the Visual Concept Detection Task (VCDT), we look at the mutual exclusion and implication relations between VCDT concepts in order to improve the automatic image annotation by Forest of Fuzzy Decision Trees (FFDTs). Second, in the ImageCLEFphoto task, we use the FFDTs learnt in VCDT task and WordNet to improve image retrieval. Third, we apply a fast visual diversity method based on space clustering to improve the cluster recall score. This study shows that there is a clear improvement, in terms of precision or cluster recall at 20, when using the visual concepts explicitly appearing in the query and that space clustering can be efficiently used to improve cluster recall.*

MOTS-CLÉS : *recherche d'images basée sur le texte, détection de concepts visuels, arbre de décisions flous, diversification*

KEYWORDS: *text-based images retrieval, visual concepts detection, Fuzzy Decision Trees*

1. Introduction

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 le contenu visuel de l'image. De plus, l'utilisateur préfèrant souvent exprimer son besoin d'information à l'aide de quelques mots-clés, il est difficile de trouver des liens avec l'information visuelle contenue dans les images. Il est donc intéressant de trouver des méthodes qui permettent de vérifier l'adéquation visuelle de l'image avec le texte de la requête posée par l'utilisateur. Dans (Yavlinsky *et al.*, 2006), des concepts visuels sont utilisés pour raffiner visuellement les résultats obtenus, cependant, l'utilisateur doit choisir manuellement le concept visuel à appliquer. Nous proposons dans cet article d'étudier une méthode qui permet de choisir automatiquement le concept visuel à appliquer.

D'un autre côté, quand un utilisateur pose une requête, ce qui l'intéresse c'est d'avoir des documents qui soient, certes tous pertinents, mais aussi qui soient les plus dissimilaires les uns des autres (Song *et al.*, 2006, Chen *et al.*, 2006, Zhai *et al.*, 2003). Par exemple, si l'utilisateur cherche des photographies d'animaux en train de nager pour illustrer un article sur la faculté de nager des animaux, toutes les images d'animaux en train de nager seront certes pertinentes, mais afin qu'il ait un aperçu, dès le début de sa recherche, de toute la diversité des images pertinentes, il serait intéressant de lui fournir dans les premiers résultats, des images qui contiennent toutes des animaux différents. Cette diversification des résultats selon le critère *animal* peut permettre à l'utilisateur de trouver plus rapidement ce qu'il recherche. Dans cet article, nous proposons une méthode de diversification basée sur les informations visuelles des images, et nous la comparons avec les résultats obtenus par une diversification aléatoire.

Notre travail se place dans le cadre de deux tâches de la campagne internationale CLEF 2008. La première tâche appelée *Visual Concept Detection Task (VCDT)* (Deselaers *et al.*, 2008) est une tâche de détection de concepts visuels. Lors de cette campagne, notre système de détection de concepts visuels est arrivé 3ième sur 11 équipes. La deuxième appelée ImageCLEFphoto (Arni *et al.*, 2008) est une tâche de recherche d'images basée sur les informations textuelles et visuelles, et propose d'étudier les problèmes soulevés par la diversité.

Dans la première partie, nous détaillons la méthode de détection des concepts visuels proposée, puis nous utilisons l'analyse des cooccurrences pour détecter des relations d'exclusion et d'implication, et nous discutons les résultats obtenus par notre méthode dans la tâche VCDT. Dans la deuxième partie, nous expliquons comment utiliser les concepts visuels de VCDT pour améliorer la recherche d'images basée sur le texte, puis nous présentons notre méthode de diversité visuelle basée sur le partitionnement de l'espace et enfin nous discutons les résultats obtenus. Dans la dernière partie, nous concluons.

2. Détection de concepts visuels

2.1. *Détection de concepts visuels à l'aide de forêts d'arbres de décision*

L'annotation automatique d'images est un problème typique d'apprentissage automatique inductif. Une des méthodes classiques dans ce domaine utilise les arbres de décisions (*Decision Trees (DT)*). Cependant, les arbres de décisions classiques rencontrent des difficultés pour traiter des données numériques ou imprécises. L'introduction de la logique floue a permis de réduire ces difficultés. L'apprentissage inductif consiste à passer du spécifique vers le général. Un arbre est construit, de la racine vers les feuilles, par partitionnements successifs de l'ensemble d'apprentissage en sous-ensembles. Chaque partition est réalisée au moyen d'un test sur un des attributs et amène à la définition d'un noeud de l'arbre (Marsala *et al.*, 1997). (Marsala *et al.*, 2006) montre que quand on considère de grands ensembles (en terme de dimension et de taille) de données non-équilibrés, il est intéressant de combiner plusieurs arbres de décisions, et ainsi d'obtenir une forêt d'arbres de décision (*Forest of Fuzzy Decision Trees (FFDT)*). De plus, la combinaison des résultats de plusieurs arbres de décision permet d'obtenir un degré de confiance dans la classification.

Durant la phase d'apprentissage, une forêt de n arbres est apprise pour chaque concept. Chaque arbre F_j de la forêt est construit en utilisant un sous-ensemble d'apprentissage T_j. Chaque sous-ensemble est un ensemble équilibré constitué d'images de l'ensemble d'apprentissage choisies aléatoirement.

Durant la phase de classification, chaque image I est classée par chaque arbre de la forêt. Le degré $d_j \in [0, 1]$ obtenu pour l'image I représente la présence du concept C pour l'arbre F_j de la forêt. Ainsi, pour chaque image I, n degrés $d_j,\ j = 1 \ldots n$ sont obtenus. Puis, tous les degrés sont agrégés par un vote : $d = \sum_{j=1}^{n} d_j$. Finalement, pour décider si une image contient un concept ou non, nous utilisons un seuil t tel que $t \leq n$: l'image I contient le concept C si $d \geq t$.

2.2. *Analyse des cooccurrences*

Les arbres de décisions apprennent chaque concept de manière indépendante. Cependant, les concepts sont reliés entre eux. Par exemple, une scène ne peut pas être simultanément à l'intérieur (*indoor*) et à l'extérieur (*outdoor*) ; si l'on observe qu'il y a des nuages (*cloudy*), on peut en déduire qu'il y le concept ciel (*sky*). Dans cette partie, nous proposons d'utiliser l'analyse des cooccurences pour déterminer automatiquement les relations entre les concepts. Une fois que nous avons découvert une relation, nous avons besoin d'une règle pour résoudre les conflits d'annotation. Cette règle doit prendre en compte les degrés de confiance donnés par les FFDTs. Par exemple, chaque image sera annotée par *outdoor* avec un certain degré et par *indoor* avec un autre degré. Cependant, les concepts *outdoor* et *indoor* ne peuvent apparaître simultanément. Pour trouver les règles à appliquer, nous étudions deux types de relations entre les concepts : les exclusions et les implications.

Exclusions Pour découvrir automatiquement les *exclusions* entre concepts, nous étudions les concepts qui n'apparaissent jamais ensemble. Pour cela, nous calculons la matrice de cooccurrences COOC entre les concepts. Comme il peut y avoir du bruit (erreurs d'annotation), nous utilisons un seuil α pour décider quels couples de concepts n'apparaissent jamais ensemble. Quand nous savons quels concepts sont reliés, nous appliquons une règle de résolution en fonction des degrés de confiance fournis par les FFDTs. Nous avons choisi une règle qui, pour les concepts mutuellement exclusifs, éliminent (c'est-à-dire donnent un degré de confiance de zéros) aux étiquettes ayant le plus faible degré de confiance. Par exemple, si *outdoor* a un degré de confiance de 42/50 et *indoor* a un degré de 20/50, alors le degré de confiance de *indoor* sera mis à zéro. Pour chaque image de test, soit d(I,C) le degré de l'image I pour le concept C, nous appliquons l'algorithme suivant :

Pour chaque couple (A,B) tel que $COOC(A, B) \leq \alpha$ (*découverte*)

Si d(I,A) $<$ d(I,B) alors d(I,A)=0 sinon d(I,B)=0 (*règle de résolution*)

où COOC est la matrice de cooccurrences des concepts.

Implications Pour découvrir les *implications*, nous étudions, par définition de l'implication, les cooccurrences entre l'absence d'un concept et la présence d'un autre concept. La matrice de cooccurrences résultante COOCNEG est asymétrique, ce qui reflète le fait qu'un concept implique un autre concept, mais cela n'est pas réciproque. La règle de résolution utilisée suppose que si un concept implique un autre concept, alors le degré de confiance de ce dernier doit être au moins égal au premier. Comme il peut y avoir du bruit, nous utilisons un seuil β pour décider quel concept implique un autre concept.

Pour chaque image de test I, soit d(I,C) le degré de l'image I pour le concept C, nous appliquons l'algorithme suivant :

Pour chaque couple (A,B) tel que $COOCNEG(A, B) \leq \beta$ (*découverte*)

d(I,B)=max(d(I,A),d(I,B)) (*règle de résolution*)

où COOCNEG est la matrice asymétrique de cooccurrences entre un concept et la négation d'un autre concept.

2.3. *Expérimentations et résultats*

2.3.1. *Corpus*

Nous appliquons notre méthode de détections de concepts visuels au corpus de la tâche *Visual Concept Detection Task (VCDT)* (Deselaers *et al.*, 2008) de la campagne internationale d'évaluation CLEF 2008. Cette tâche correspond à un problème de classification multi-classes multi-étiquettes. Le corpus de VCDT contient 1827 images d'apprentissage et 1000 images de test. Il y a 17 concepts visuels. Une image d'apprentissage est annotée en moyenne par 5.4 concepts (entre 0 (2 images) et 11 concepts

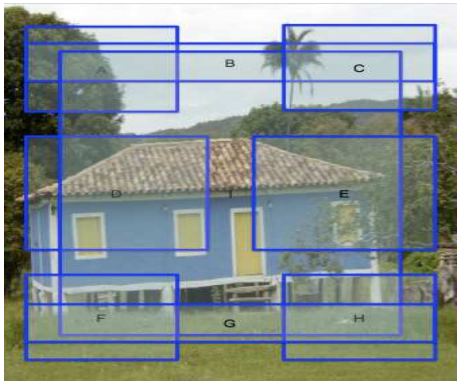


Figure 1. *Les images sont segmentées en 9 régions*

par image). Un concept annote en moyenne 584 images d'apprentissage (entre 68 et 1607 images d'apprentissage par concept).

2.3.2. *Descripteurs visuels*

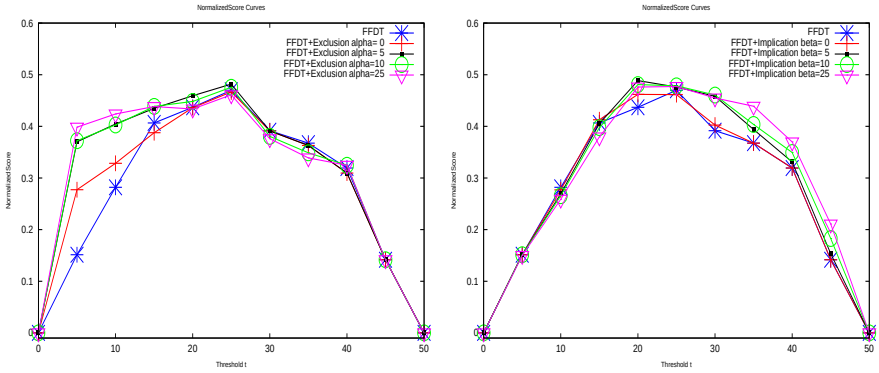
Les descripteurs visuels utilisés sont exclusivement basés sur la couleur. Afin d'obtenir une information liée à la disposition spatiale des objets dans les images, nous segmentons les images en 9 régions qui se chevauchent (voir figure 1). Pour chaque région, nous calculons un histogramme HSV. Le nombre de dimensions de l'histogramme reflète l'importance de la région. La région centrale représente le thème de l'image. La région en haut et celle du bas sont intéressantes pour les concepts visuels généraux, tels que le ciel, le soleil, la végétation, la mer... Les autres régions sont décrites en termes de différences de couleurs entre la gauche et la droite. L'idée est de rendre explicite les symétries. En effet, les objets peuvent apparaître d'un côté ou de l'autre de l'image. Or étant donné que les arbres de décisions ne sont pas capables de découvrir automatiquement ce genre de relations, l'utilisation de ces différences permet de leur donner la possibilité de tenir compte de ces symétries. Au final, chaque image est représentée par un vecteur de valeurs numériques.

2.3.3. *Mesure de performances*

Pour mesurer les performances de notre système de détection de concepts visuels, nous avons choisi d'utiliser le *Normalized Score (NS)* qui a déjà été utilisé par de nombreux modèles d'annotation automatique (Barnard *et al.*, 2003). Pour un concept C donné, le *Normalized Score (NS)* peut être défini ainsi :

$$NS = \frac{r}{w} - \frac{w}{N-n}$$

où N est le nombre d'images de l'ensemble de test, n est le nombre d'images de l'ensemble de test initialement annotées par le concept C, r est le nombre d'images



(a) α (b) β

Figure 2. *Influence des paramètres α et β sur le* Normalized Score (NS) *en fonction du seuil t de décision (les degrés de confiance fournis par les arbres varient entre 0 et 50)*

annotées par le système avec le concept C et qui étaient initialement annotées par ce concept, et w est le nombre d'images annotées par le système avec le concept C et qui n'étaient pas initialement annotées par ce concept C. Ce score correspond donc à la somme de la sensibilité et de la spécificité moins 1. Il varie entre -1 et 1. Le score vaut 1 quand le système ne commet aucune erreur, -1 quand toutes les images sont mal annotées, 0 quand toutes les images sont annotées par le concept C.

2.3.4. *Utilisation des relations d'exclusion et d'implication*

Une étape préliminaire avant d'extraire des concepts visuels est d'étudier les valeurs de cooccurrence entre les concepts pour découvrir les relations d'exclusion et d'implication. Pour les 17 concepts, il y a 136 valeurs de cooccurence. Ces valeurs varient de 0 (non-cooccurrences) à 1443 (sur les 1827 images d'apprentissage). Comme il peut y avoir du bruit, nous avions fixé, lors de notre participation à la tâche VCDT, le seuil α à la valeur 5. Ce seuil avait été déterminé grâce à la distribution des valeurs de cooccurrence de l'ensemble d'apprentissage. Si $\alpha = 5$, cela signifie que deux concepts sont considérés exclusifs si moins de 5 images sont annotées par les deux concepts. La figure 2(a) montre que cette valeur du seuil maximise en effet les résultats pour $t = 25$, mais que pour un seuil de décision $t \leq 15$, il peut être intéressant de prendre une valeur de α plus grande. De même, nous avions fixé $\beta = 5$ (un concept implique un autre concept si au maximum 5 images d'apprentissage ne sont pas annotées par le premier concept, et en même temps annotées par le second concept). La figure 2(b) confirme que les meilleurs scores à $t = 25$ sont obtenus pour $\beta = 5$, mais que pour $t \geq 30$, il peut être intéressant de prendre une valeur de β plus grande. Enfin, ces deux figures montrent que prendre $\alpha = 0$ ou $\beta = 0$ ne donne jamais de meilleurs résultats

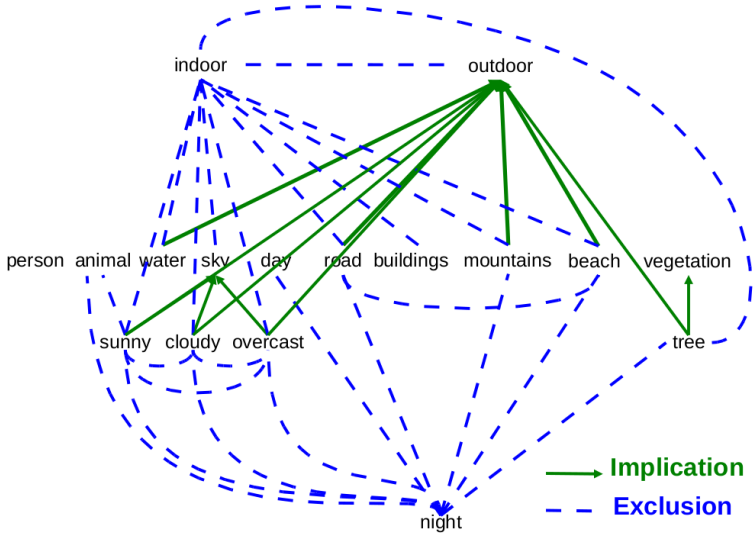


Figure 3. *Schéma montrant les relations d'exclusion et d'implication entre les concepts automatiquement découvertes*

que pour $\alpha = 5$ ou $\beta = 5$, et que donc prendre en compte les erreurs d'annotations permet d'améliorer sensiblement les résultats.

Pour $\alpha = 5$ et $\beta = 5$, notre système a automatiquement découvert 25 relations d'exclusion et 12 relations d'implication (voir figure 3). Nous avons trouvé toutes les relations évidentes (par exemple, les concepts *indoor* et *outdoor* sont exclusifs ; un arbre implique de la végétation), ainsi que quelques relations moins triviales. Par exemple, les concepts *sunny* et *animal* sont trouvés exclusifs. Cette relation peut être expliquée par le fait que, lorsqu'une personne annote une image, son attention peut se focaliser sur un objet (ou un animal) et ne pas préter attention au fait que le ciel soit ensoleillé, cette information étant jugée secondaire ou inintéressante.

Finalement, la figure 4 compare les scores NS obtenus par les FFDTs seuls ou avec les régles d'implication et d'exclusion. Nous notons que les meilleurs résultats sont obtenus pour $t = 20$ par l'application des règles d'implication aux degrés de confiance des FFDT. Pour un seuil de décision $t \leq 15$, il vaut mieux utiliser les exclusions ; pour $t \geq 30$, il vaut mieux utiliser les implications. Finalement, la méthode FFDT+Implication+Exclusion donne globalement les meilleurs résultats. Cependant, nous remarquons que pour $t = 25$, toutes les méthodes donnent quasiment les mêmes résultats. Pour ce seuil que nous utiliserions classiquement pour prendre une décision, l'intérêt d'utiliser les règles d'implication et d'exclusion n'est donc pas totalement concluante.

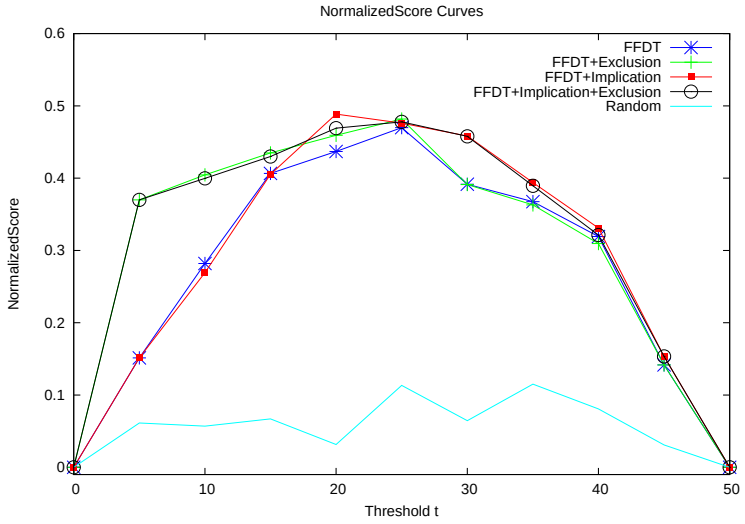


Figure 4. *Courbes de scores NS en fonction du seuil de décision (les degrés de confiance fournis par les arbres varient entre 0 et 50) avec* $\alpha = 5$ *et* $\beta = 5$

	EER	Gain	AUC	Gain
Moyenne des 53 runs	33.92	-	63.64	-
Classifieurs aléatoires	50.17	-48%	49.68	-22%
FFDT	24.55	+28%	82.74	+30%

Tableau 1. *Résultats de la tâche VCDT 2008 (EER : Equal Error Rate - AUC : Area under ROC curve). La moyenne des 53 runs correspond à la moyenne des scores EER et AUC des 53 runs soumis par les 11 équipes participantes à la tâche VCDT en 2008.*

Le tableau 1 montre les résultats obtenus lors de la compagne d'évaluation ImageCLEF en 2008. Notre méthode basée sur les FFDTs est arrivée quatrième sur 53 résultats soumis (troisème équipe sur 11 équipes internationales).

3. Recherche d'images

3.1. *Utilisation de concepts visuels pour améliorer la recherche d'images basée sur le texte*

De nombreux travaux (Barnard *et al.*, 2003, Datta *et al.*, 2008, Tollari *et al.*, 2007) montrent que combiner des informations visuelles et textuelles améliorent la recherche

d'images, mais la plupart de ces travaux se concentrent sur la fusion précoce ou tardive de ces informations ou sur l'annotation des images. Nous proposons d'utiliser les concepts visuels appris par les FFDTs pour améliorer la recherche d'images basée uniquement sur le texte. La difficulté est de déterminer comment utiliser les concepts visuels dans le cas où les seules informations que l'on peut utiliser sont le nom du concept, les descripteurs visuels, et la requête composée de quelques mots-clés.

A l'aide des FFDTs apprises pour chaque concept (voir partie 2) et des descripteurs visuels de chaque image, nous pouvons donner un degré de confiance qu'un certain concept apparaisse dans une nouvelle image. Il reste donc à trouver un moyen de faire la correspondance entre la requête et le (ou les) concept(s) que l'on veut détecter dans les images.

Premièrement, si le nom du concept apparaît directement dans les mots de la requête (méthode DIRECTE), nous proposons de filtrer les images ordonnées par la recherche textuelle en fonction du degré qu'elles obtiennent pour ce concept.

Deuxièmement, si le nom du concept apparaît dans les mots de la requête ou dans une liste de synonymes des mots de la requête donnés par WordNet (Fellbaum, 1998) (méthode WN), nous proposons également de filtrer les images ordonnées par la recherche textuelle en fonction du degré qu'elles obtiennent pour ce concept. Par exemple, la requête 5 d'ImageCLEFphoto 2008 est «*animal swimming*». En utilisant la méthode DIRECTE, le système détermine automatiquement qu'il doit utiliser la FFDT du concept *animal*. Si, de plus nous utilisons WordNet (méthode WN), le système détermine automatiquement qu'il doit utiliser les FFDTs des concepts *animal* et *water* (car d'après WordNet, un synonyme pour *swimming* est «*water sport, aquatics*».

Pour chaque requête, nous déterminons la liste ordonnée des images pertinentes selon le modèle de langue (LM) ou selon le modèle TD-IDF utilisé sur le texte. Puis, à l'aide des FFDTs, nous réordonnons les 50 premières images de chaque requête ainsi : le système parcourt les images retrouvées du rang 1 au rang 50. Si le degré d'une image est inférieur à un seuil t, alors cette image est réordonnée à la fin de la liste des 50 premières images. De cette façon, les images pertinentes se trouvent toujours dans les 50 premiers résultats.

3.2. *Promouvoir la diversité en utilisant le clustering spatial*

Pour une requête donnée, les documents similaires sont naturellement ordonnées à des rangs proches. Quand un utilisateur pose une requête, ce qui l'intéresse c'est d'avoir des documents qui soient certes tous pertinents, mais aussi qui soient les plus dissimilaires les uns des autres.

Les techniques de *clustering* visuel sont étudiées depuis de nombreuses années. Deux approches sont généralement proposées : le partitionnement des données et le partitionnement de l'espace. La première approche nécessite beaucoup de temps de calcul et doit être adaptée à la distribution des premières images résultats d'une re-

quête donnée comme dans (Inoue *et al.*, 2008). La seconde approche, comme elle est faite indépendamment des données, est souvent moins efficace, mais peut-être appliquée de manière très rapide. Nous avons choisi de réaliser un partitionnement de l'espace visuel fondé sur l'histogramme Hue de l'espace HSV. Pour chaque image, nous binarisons l'histogramme Hue. Chaque vecteur binaire correspond à un cluster. En fonction du nombre de dimensions nh de l'histogramme, nous obtiendrons 2^{nh} clusters possibles (les clusters ne seront pas forcement tous instantiés, car certains pourront correspondre à aucune donnée).

Pour chaque requête, les images sont classées en deux listes. Le système parcours les images dans l'ordre : des plus pertinentes vers les moins pertinentes. Si une image appartient à aucun des clusters des images plus pertinentes qu'elle, alors cette image est ajoutée à la fin de la première liste. Si une image appartient au même cluster qu'une image plus pertinente qu'elle, alors cette image est ajoutée à la fin de la deuxième liste. Au final, nous obtenons dans la première liste uniquement des images avec des clusters différents. Nous concaténons ensuite la première liste et la deuxième liste. L'image de rang 1 est toujours au rang 1 ; l'image de rang 2 se retrouve soit au rang 2 si son cluster est différent du cluster de l'image de rang 1, soit à la position $nbcv + 1$ si son cluster est identique (avec $nbcv$ le nombre de clusters visuels), et ainsi de suite. Dans la pratique, comme nous nous intéressons seulement aux 20 premiers documents pertinents, il suffit de s'arrêter de parcourir les images lorsque nous avons trouvé 20 images dans 20 clusters différents. Plus nous aurons de clusters, et moins il y aura de changement dans l'ordre des images. Nous appelons cette méthode : DIVVISU.

Pour avoir un point de comparaison, nous proposons également une méthode de diversification "naïve" qui consiste à permuter aléatoirement les a premiers résultats. Nous appelons cette méthode : DIVALEA.

3.3. *Expérimentations et résultats*

3.3.1. *Corpus*

Nous utilisons le corpus la tâche ImageCLEFphoto (Arni *et al.*, 2008) de la campagne dévaluation CLEF 2008. Ce corpus contient 20k images généralistes et 39 *topics*. Chaque *topic* est composé d'un titre, d'une partie narrative, de 3 images correspondant au *topic*, ainsi que d'un élément indiquant sur quel critère doit être appliqué la diversité (<CLUSTER>). Par exemple, le premier *topic* est :

```
<TITLE>church with more than two towers</TITLE>
<CLUSTER>city</CLUSTER>
<NARR>Relevant images will show a church, cathedral or a mosque with
three or more towers. Churches with only one or two towers are not
relevant. Buildings that are not churches, cathedrals or mosques are
not relevant even if they have more than two towers.</NARR>
```

Les 39 requêtes doivent être dérivées de chacun des *topics*. Il y a 17 critères (parfois appelés sous-thèmes (*subtopics*) (Zhai *et al.*, 2003)) de diversités différents : *animal*,

bird, city, city/nationalpark, composition, country, group composition, landmark location, sport, state, statue, tourist attraction, vehicle type, venue, volcano, weather condition. La plupart de ces critères correspondent à des lieux. Par exemple, pour la première requête, le critère de diversité est *city*. Pour cette requête, ce critère contient 5 clusters ($n_c = 5$) : *Moscow, Saint Petersburg, Melbourne, Sydney, Bolshaya Reka*. Chaque image du corpus est associée à une légende contenant le titre de l'image, sa date de création, sa localisation, le nom du photographe, une description sémantique du contenu de l'image (déterminée par le photographe) ainsi que de notes additionnelles.

3.3.2. *Mesures de performances*

Les mesures classiquement utilisées en recherche d'information sont généralement la précision et le rappel. De plus, afin de combiner ces deux mesures, la F1-mesure est généralement utilisée. Pour ImageCLEFphoto 2008, le but est de retrouver non seulement les documents pertinents, mais aussi de retrouver, dans les premiers résultats, les documents pertinents qui sont les plus différents les uns des autres en fonction du critère de diversité choisi. C'est pourquoi les mesures de performances utilisées sont : la précision à 20 (P20), le *cluster recall* à 20 (CR20) (Zhai *et al.*, 2003) et la F1-mesure appliquée au P20 et au CR20. Soit $nbpr(n)$ le nombre de documents pertinents retrouvés parmi les n premiers documents retrouvés, la précision à 20 peut être définie ainsi :

$$P20 = \frac{nbpr(20)}{20}.$$

Le *cluster recall* à 20 (CR20) (appelé aussi *S-recall*) (Zhai *et al.*, 2003) a pour but de mesurer le nombre de clusters différents présents dans les 20 premiers résultats. Soit n_c le nombre de clusters différents pour une requête donnée, soit $nbcpr(n)$ le nombre de clusters différents couverts par les documents pertinents retrouvés parmi les n premiers documents retrouvés pour cette requête, alors le CR20 est définie ainsi :

$$CR20 = \frac{nbcpr(20)}{n_c}.$$

La dernière mesure utilisée est la F1-mesure définit ainsi dans notre cas :

$$F1 - mesure = 2 \times \frac{P20 \times CR20}{P20 + CR20}.$$

3.3.3. *Correspondance directe et par WordNet*

Nous réalisons d'abord une recherche d'images basée uniquement sur le texte. Pour cela, nous construisons les requêtes en utilisant les éléments du titre ainsi que les phrases de la balise `<NARR>` qui ne contiennent pas le mot *not*. Pour décrire chaque image, nous prenons en compte le texte contenu dans tous les éléments de la légende. Nous appliquons ensuite un modèle classique de langue (LM) ainsi que le modèle TF-IDF.

Pour déterminer si une image contient un concept visuel, nous choisissons de fixer le seuil t à la médianne de tous les degrés obtenus par un concept donné (cette valeur

Tableau 2. *Comparaison des méthodes DIRECTE et WN. Par la méthode DIRECTE, seulement 11 requêtes sont modifiées. Par la méthode WN, 25 requêtes sont modifiées*

		Moyenne sur 39 requêtes		Moyenne des requêtes modifiées		
Texte	Méthode	P20 (gain %)	CR20 (gain %)	Nb topics	P20 (gain %)	CR20 (gain %)
LM	-	0.185(-)	0.247(-)	11	0.041(-)	0.090(-)
				25	0.148 (-)	0.254 (-)
	DIRECTE	0.195(+6)	0.257(+4)	11	0.077(+88)	0.126(+40)
	WN	0.176(-5)	0.248(+1)	25	0.134(-9)	0.257(+1)
TF-IDF	-	0.250(-)	0.300(-)	11	0.155(-)	0.161(-)
				25	0.210(-)	0.305(-)
	DIRECTE	0.269(+8)	0.313(+5)	11	0.223(+44)	0.209(+30)
	WN	0.260(+4)	0.293(-2)	25	0.226(+8)	0.294(-4)

varie de 7.3 (*overcast*) à 28.8 (*outdoor*). Nous n'avons pas utiliser dans cette partie les règles d'exclusion et d'implication.

Le tableau 2 montre que, en moyenne sur tous les topics, la méthode DIRECTE améliore la précision à 20 documents (P20) de +8% par rapport au TF-IDF et de +6% par rapport au LM, tandis que la méthode WN améliore le P20 du TF-IDF de +4%, mais diminue de -5% le P20 du LM. Comme les méthodes DITECTE et WN dépendent de la présence du nom du concept dans la requête textuelle et que certaines requêtes ne contiennent aucun des noms des 17 concepts, les résultats de certaines requêtes ne sont pas modifiés. La méthode DIRECTE modifie seulement 11 requêtes, tandis que la méthode WN modifie 25 requêtes. C'est pourquoi nous séparons, dans le tableau 2, les résultats en 3 groupes. Nous remarquons une amélioration des scores de P20 de +44% par rapport au TF-IDF (+30% par rapport au LM) pour les 11 requêtes modifiés par la méthode DIRECTE, mais une amélioration de seulement +8% et une diminution de -9% en utilisant WordNet. Nous en déduisons que la méthode DIRECTE permet d'améliorer sensiblement les résultats obtenus par le texte seul, mais que par contre l'utilisation de WordNet n'est pas adapté pour ce genre de tâche. Par exemple, d'après WordNet, le concept visuel *person* n'est pas dans la liste des synonymes du mot *people*. Nous ne pouvons donc trouver de lien entre les requêtes recherchant des personnes et le concept *person*. Nous avons également étudié d'autres types de relations (hyponymie, hypernymie...), mais les résultats obtenus étaient globalement inférieurs à ceux obtenus par synonymie.

3.3.4. *Diversification*

Notre méthode de diversification est basée sur un partitionnement de l'espace visuel, et non pas sur des informations sémantiques qui pourraient être utiles comme, par exemple, une liste de villes pour le critère *city*. Nous savons que nos résultats seront sous-optimaux, car la diversité visuelle n'entraîne pas forcement la diversité

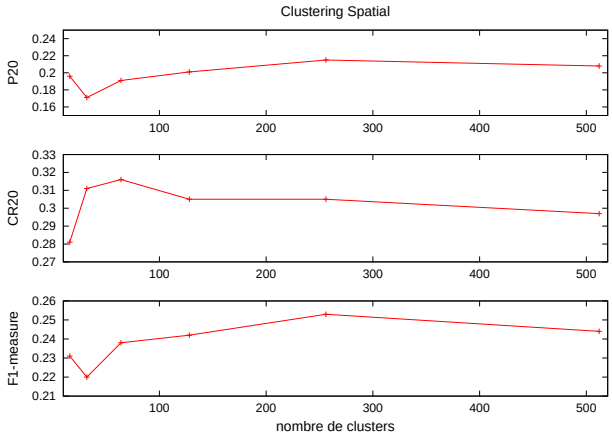


Figure 5. *Influence du nombre de clusters sur les scores P20, CR20 et F1-measure obtenus par diversification DIVVISU des résultats du modèle TF-IDF*

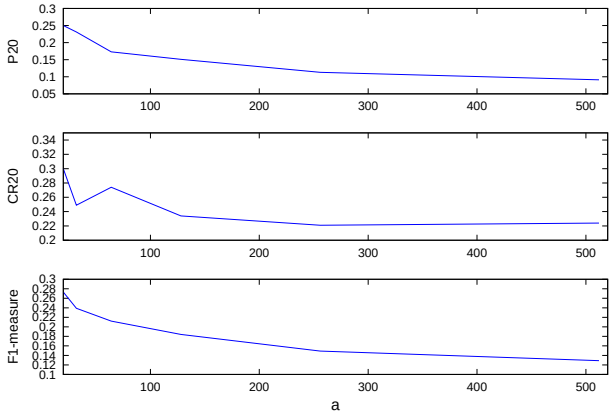


Figure 6. *Permutation aléatoire des a premiers documents obtenus par TF-IDF*

sémantique. Nous supposons cependant que des photographies de cathédrales prises dans une même ville seront assez visuellement similaires. Pour pouvoir étudier l'influence du nombre de clusters visuels, nous avons choisi de faire varier le nombre de dimensions de l'histogramme Hue de 4 à 9 dimensions. La quantité d'information initiale étant ainsi toujours la même. Nous obtenons un nombre théorique de clusters variant de 16 à 512, et un nombre de clusters instanciés légèrement inférieur. La figure 5 montre l'évolution des scores P20, CR20 et F1-measure en fonction du nombre de clusters de l'espace visuel. Pour un nombre de clusters égal à 16, la diversification

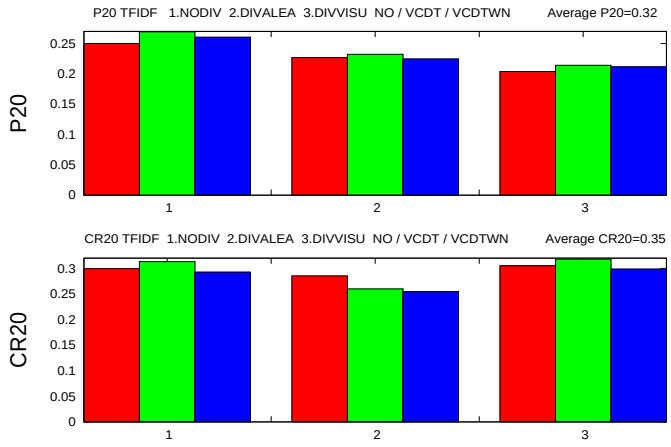


Figure 7. *Comparaison des méthodes de diversification (1. sans diversification, 2. diversification aléatoire (DIVALEA), 3. diversification par clustering spatial (DIVVISU avec $nh = 8$)). Pour chaque méthode de diversification, la première barre correspond au TF-IDF seul, la deuxième à TF-IDF+DIRECTE et la troisième à TF-IDF+WN*

n'est pas complètement effectuée, car certaines images ayant des clusters similaires aux images de rangs plus élevés se retrouvent toujours dans les 20 premiers résultats. Pour un nombre de clusters supérieur à 32, le P20 chute brutalement et puis remonte doucement vers le P20 du TF-IDF sans toutefois l'atteindre. A l'inverse, le CR20 augmente atteignant son maximum pour un nombre de clusters de 64, puis diminue pour atteindre le CR20 du TF-IDF. En moyenne, la meilleur valeur de F1-measure est obtenu pour un nombre de clusters de 256 (soit une dimension de l'espace de 8). Pour étudier la difficulté de garder des scores forts lorsque l'on effectue une diversification des résultats, la figure 6 montre la baisse des scores P20, CR20 et F1-mesure en fonction du nombre de documents permutés aléatoirement. La figure 7 compare les scores de diversification pour les méthodes DIVVISU (à 256 clusters) et DIVALEA (permutation aléatoire des 40 premiers documents) par rapport aux scores TF-IDF sans diversification. Nous remarquons que les deux méthodes proposées donnent des scores P20 largement inférieurs au P20 du TF-IDF, mais DIVALEA diminue le CR20, tandis que DIVVISU l'améliore légèrement (+2%). Nous en concluons que notre méthode DIVVISU améliore légèrement les résultats de diversification, mais que, comme de nombreuses autres méthodes (voir (Tollari *et al.*, 2008)), elle diminue le P20. La figure 8 montre les scores CR20 de TF-IDF versus TF-IDF+DIVVISU. Notre méthode DIVVISU, qui est basée uniquement sur de l'information visuelle, ne semble pas être adaptée spécialement pour un type de critère de diversification. En effet, il n'y a aucun critère de diversification qui donne de bons résultats seulement pour TF-IDF+DIVVISU.

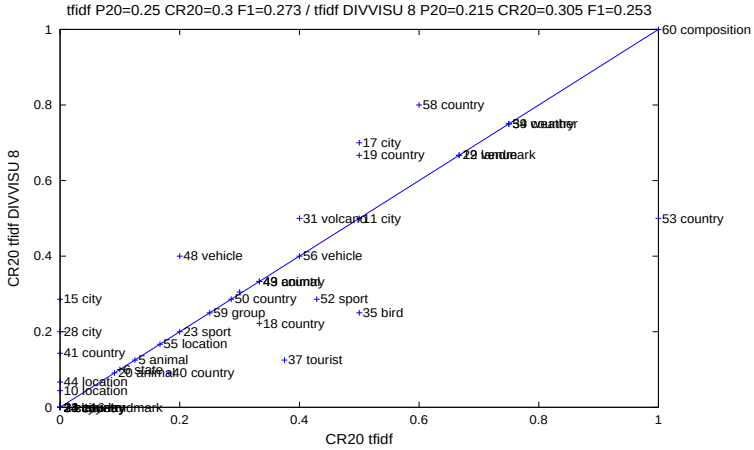


Figure 8. *Comparaison des scores CR20 du TF-IDF et du TF-IDF diversifié par DIV-VISU avec $nh = 8$ (256 clusters visuels). Chaque point correspond à une requête (le nombre correspond au numéro de la requête et le mot associé correspond au critère de diversification demandé dans le topic)*

4. Conclusion

Dans cet article, nous nous intéressons à deux difficultés. La première est l'exploitation de concepts visuels pour améliorer la recherche d'images basée uniquement sur le texte. Pour tenter de résoudre cette difficulté, nous utilisons des forêts d'arbres de décisions flous pour donner un degré de confiance que le concept soit dans une image, puis en fonction des termes de la requête, nous filtrons les images dont le degré de confiance correspond aux concepts visuels de la requête est trop faible. Les résultats montrent une nette amélioration des scores pour les requêtes qui contiennent explicitement le nom d'un concept. Nous en déduisons que la difficulté principale est de déterminer quel concept appliquer pour une requête qui ne contient pas explicitement de concept. La seconde est la diversification des résultats pertinents. Nous proposons d'utiliser le partitionnement de l'espace visuel afin d'obtenir très rapidement le cluster visuel d'une image, puis de garder dans les 20 premiers documents uniquement des images qui ont des clusters différents. Notre méthode augmente légèrement les scores de diversité, et a l'avantage de pouvoir être utilisée très simplement sans lourd calcul.

Dans nos futurs travaux, nous souhaitons améliorer nos règles de résolutions (exclusion et implication) pour obtenir de meilleurs résultats de classification, puis les utiliser dans la tâche de recherche d'images. En effet, nous avons fixé un seuil de décision t à la médiane de tous les degrés obtenus, or cette valeur varie de 7.3 à 28.8, l'utilisation de règles d'exclusion dans la tâche de recherche d'images devrait, d'après la figure 4, améliorer nos résultats.

Remerciements

Ce travail a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence ANR-06-MDCA-002 (projet AVEIR).

5. Bibliographie

Arni T., Clough P., Sanderson M., Grubinger M., « Overview of the ImageCLEFphoto 2008 Photographic Retrieval Task », *Evaluating Systems for Multilingual and Multimodal Information Access – 9th Workshop of the Cross-Language Evaluation Forum*, Lecture Notes in Computer Science, Aarhus, Denmark, 2008.

Barnard K., Duygulu P., de Freitas N., Forsyth D., Blei D., Jordan M. I., « Matching Words and Pictures », *Journal of Machine Learning Research*, vol. 3, p. 1107-1135, 2003.

Chen H., Karger D. R., « Less is more : probabilistic models for retrieving fewer relevant documents », *ACM SIGIR*, p. 429-436, 2006.

Datta R., Joshi D., Li J., Wang J. Z., « Image retrieval : Ideas, influences, and trends of the new age », *ACM Computing Surveys*, 2008.

Deselaers T., Deserno T. M., « The Visual Concept Detection Task in ImageCLEF 2008 », *Evaluating Systems for Multilingual and Multimodal Information Access – 9th Workshop of the Cross-Language Evaluation Forum*, Lecture Notes in Computer Science, Aarhus, Denmark, 2008.

Fellbaum C., *WordNet - An Electronic Lexical Database*, Bradford books, 1998.

Inoue M., Grover P., « Effects of Visual Concept-based Post-retrieval Clustering in ImageCLEFphoto 2008 », *Working Notes for the CLEF 2008 workshop*, 2008.

Marsala C., Bouchon-Meunier B., « Forest of fuzzy decision trees », *Proceedings of the Seventh International Fuzzy Systems Association World Congress*, vol. 1, p. 369-374, 1997.

Marsala C., Detyniecki M., « TRECVID 2006 : Forests of fuzzy decision trees for high-level feature extraction », *TREC Video Retrieval Evaluation Online Proceedings*, 2006.

Song K., Tian Y., Gao W., Huang T., « Diversifying the image retrieval results », *ACM Multimedia*, ACM, New York, NY, USA, p. 707-710, 2006.

Tollari S., Glotin H., « Web Image Retrieval on ImagEVAL : Evidences on visualness and textualness concept dependency in fusion model », *ACM Conference on Image and Video Retrieval (CIVR)*, p. 65-72, 2007.

Tollari S., Mulhem P., Ferecatu M., Glotin H., Detyniecki M., Gallinari P., Sahbi H., Zhao Z.-Q., « A comparative study of diversity methods for different text and image retrieval approaches », *Evaluating Systems for Multilingual and Multimodal Information Access – 9th Workshop of the Cross-Language Evaluation Forum*, Lecture Notes in Computer Science, Aarhus, Denmark, 2008.

Yavlinsky A., Heesch D., Rüger S. M., « A Large Scale System for Searching and Browsing Images from the World Wide Web », *CIVR 2006*, p. 537-540, 2006.

Zhai C. X., Cohen W. W., Lafferty J., « Beyond independent relevance : methods and evaluation metrics for subtopic retrieval », *ACM SIGIR*, p. 10-17, 2003.