Vers une indexation personnalisée de photographies par apprentissage non supervisé de régularités

Stéphane Bissol, Philippe Mulhem, Yves Chiaramella

LIG / MRIM
BP 53 X
38041 Grenoble Cedex
Stephane.bissol@gmail.com, {Philippe.Mulhem, Yves.Chiaramella}@imag.fr

RÉSUMÉ. *Cet article propose une approche pour indexer des images photographiques avec pour objectif de permettre une bonne qualité d'annotation des images et aussi un moyen de visualiser, pour un utilisateur non expert, ce que le système apprend pour éventuellement corriger un apprentissage défectueux. Notre approche repose sur la génération de régularités dans l'espace des caractéristiques extraites en se basant sur un apprentissage non supervisé, puis sur un apprentissage supervisé afin d'associer ces régularités à des termes d'indexation. Des expérimentations préliminaires sur des photographies ont été menées et sont décrites.*

ABSTRACT. *This article proposes an approach to index photographic images, allowing a good quality of annotation and means for a non expert user to visualize the important features of the learning parameters. Our approach relies on the generation of regularities over an initial set of extracted features. We use an unsupervised learning approach followed by a supervised one which links the regularities extracted to indexing terms. Preliminary experiments on photographs were carried out and are described.*

MOTS-CLÉS : *Indexation d'images symbolique.*

KEYWORDS: *Symbolic Indexing, Photographs.*

CORIA 2007, pp. 181–196
Saint-Étienne, 28–30 mars 2007

1. Introduction

L'objectif de cet article est de décrire une approche permettant à des utilisateurs d'indexer des photographies personnelles par des symboles. Il a été établi dans (Mulhem *et al* 2004) que dans le cas de photographies personnelles, les utilisateurs organisent et recherchent leurs photographies entre autres sur le contenu des images. Il est donc nécessaire de prendre cet aspect en compte lors de l'indexation des images personnelles. De plus, étant donné la diversité des utilisateurs, il est illusoire de proposer un vocabulaire d'indexation figé.

D'après (Smeulders et al 2000), les problèmes inhérents à l'indexation d'images par termes dans le cas de domaines non restreints sont les suivants:

- Le **fossé sensoriel** est la disparité qui existe entre l'unicité d'un objet physique et la multiplicité des apparences qu'il peut revêtir.
- Le **fossé sémantique** est le manque de concordance entre l'information que l'on peut extraire des données visuelles et l'interprétation des mêmes données qu'en fait un utilisateur dans un contexte donné.

Le travail proposé ici tente de réduire le fossé sensoriel en établissant des régularités dans les caractéristiques visuelles extraites, et de réduire le fossé sémantique en associant les régularités à des termes. Un aspect important de notre proposition est que nous voulons garantir un apprentissage de bonne qualité sans pour autant nécessiter une grande taille de base d'apprentissage, de manière à permettre à un utilisateur de personnaliser le vocabulaire d'indexation des images. Pour atteindre ces objectifs, nous définissons un processus en deux étapes : l'une est non supervisée et permet d'extraire des régularités sur les caractéristiques de bas niveaux extraites, la seconde est supervisée et associe des régularités à des termes suivant une approche statistique classique.

Le plan proposé dans cet article est le suivant. Dans la section 2 nous décrivons les travaux relatifs à notre proposition. La partie 3 est le cœur de ce travail, et est dédiée à la description de notre proposition sur l'extraction des régularités de manière non supervisée, puis la description de l'apprentissage supervisé qui associe régularités et termes. La partie 4 présente la classification mise en place à partir de l'apprentissage. Les expérimentations menées sur notre proposition sont décrites en partie 5, avec une étude de chacune des deux étapes de la proposition et un exemple de visualisation de régularités sur un exemple. Nous concluons en section 6 en donnant les futures directions de ce travail.

2. Etat de l'art

Des travaux tels que ceux de (Town et Sinclair 2000), comme ceux de (Lim 01), proposent un processus d'apprentissage basé sur des réseaux de neurone ou des SVM, pour passer des caractéristiques signal aux descriptions par des termes de

parties d'images. Ces approches nécessitent des apprentissages lourds pour un utilisateur (des milliers d'exemples dans (Town et Sinclair 2000)), à cause de la variabilité des objets visuels rencontrés, ce qui n'est pas adapté à une personnalisation de l'apprentissage par un utilisateur spécifique. De plus, il est difficile pour un utilisateur de vérifier pourquoi un apprentissage ne donne pas de bons résultats car on ne peut pas présenter visuellement les caractéristiques de l'apprentissage. Ces approches ne nous semblent donc pas adaptés à notre problématique. D'autres travaux dans le cadre de TrecVid (Snoek et al. 2006) sur les vidéos souffrent des mêmes critiques.

Des travaux sur l'annotation automatique d'images comme le *Cross-Media Relevance Model* (Jeon et al 2003), ou les *Multi-Media Hierarchical Models* (Barnard et al. 2003) proposent des approches probabilistes pour résoudre ce problème, en obtenant de bons résultats, mais qui nécessitent l'utilisation de processus coûteux (maximisation de l'espérance) qui nécessitent de nombreux exemples (plusieurs centaines d'images). De plus, il est difficile de savoir de manière simple pourquoi un apprentissage ne fonctionne pas correctement pour un utilisateur. Nous pouvons cependant souligner que dans (Barnard et al. 2003) l'utilisation d'un modèle hiérarchique pour prendre en compte les éléments plus ou moins récurrents dans les images annotées.

La notion de représentation hiérarchique se retrouve dans un travail de Sabrina Tollari (Tollari 2006), dans lequel des regroupements hiérarchiques de caractéristiques visuelles de régions potentiellement indexées par un terme. Un choix est effectué pour déterminer le meilleur choix des regroupements dans la hiérarchie. Dans ce cas, on peut considérer que les regroupements choisis décrivent en fait une certaine variabilité dans la représentation visuelle d'un objet correspondant à un terme. Dans ce travail, un nombre important d'images exemples ou de régions sont utilisées pour obtenir de bons résultats par cette approche.

Dans tous les cas étudiés ci-dessus, on ne va pas jusqu'à une description explicite claire pour un utilisateur des associations entre termes et éléments visuels. Rappelons qu'une telle visualisation pourrait faciliter un apprentissage par un utilisateur non spécialiste en explicitant pourquoi l'apprentissage ne fonctionne pas bien. Les travaux de Mooney (Mooney 1995) se positionnent sur les techniques d'apprentissage, en proposant une approche pour apprendre des concepts formés à partir de formes normales conjonctives. Les formes normales conjonctives (CNF) représentent les concepts sous la forme de conjonctions de disjonctions ($F_1 \wedge F_2 \wedge ... \wedge F_j$ où chaque F_i est une disjonction de littéraux). Sur 5 jeux de données, les représentations sous forme CNF produisent systématiquement des résultats meilleurs ou équivalents aux formes normales disjonctives et arbres de décisions, et ceci sous une forme plus compacte. Pour nous, des descriptions de termes sous ces formes permettraient d'interagir plus aisément avec un utilisateur sur l'apprentissage.

Notre proposition tente donc de permettre une indexation d'images photographiques personnelles en s'inspirant de (Mooney 1995), tout en visant un bons résultats sur un ensemble réduit d'exemples et contre-exemples d'apprentissage.

3. Apprentissage à base de régularités non supervisées

3.1. *Description générale*

La Figure 1 schématise l'apprentissage tel que nous l'envisageons : l'apprenant (système informatique) est soumis à un flux (1) de données, ces données n'étant pas générées aléatoirement, elles possèdent certaines régularités que l'apprenant découvre : d'abord, des régularités simples (2) impliquant un petit nombre de traits, puis ces régularités sont utilisées pour construire des traits plus complexes, de manière hiérarchique (3). Lorsque l'apprenant possède les représentations nécessaires, l'apprentissage supervisé peut avoir lieu : les régularités perçues influent positivement ou négativement sur la détection de certains concepts (4). Par exemple, une flèche '+' dans la figure 1 signifie qu'une régularité est fréquemment présente dans les instances de Terme_2 par exemple. Par conséquent, découvrir cette régularité dans un exemple inconnu renforcera la certitude que cet exemple est une instance de ce terme. A l'inverse, une flèche '-' signifie qu'une régularité n'est pas statistiquement liée à un terme.

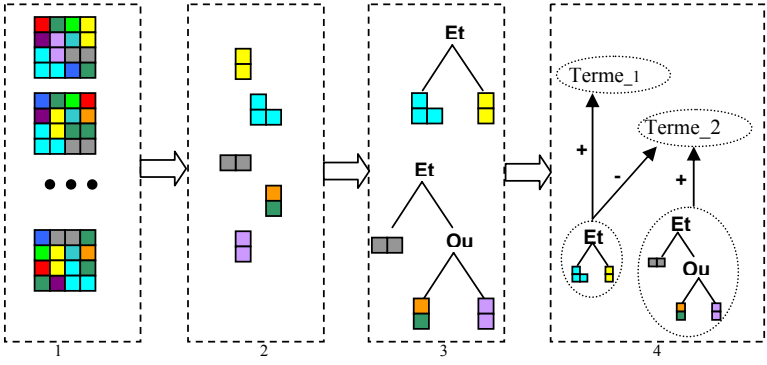

Figure 1. Processus d'annotation.

L'idée centrale de cette vision de l'apprentissage est qu'une régularité connue devient, lorsqu'elle est présente, une nouvelle dimension qui remplace les dimensions dont cette régularité est formée. Contrairement à la plupart des modèles d'apprentissage, les exemples ne sont pas perçus comme des points dans un même

espace, puisque l'espace est différent pour chaque exemple. L'apprentissage est nécessairement hiérarchique : chaque régularité devient une unité de représentation distribuée, pouvant être utilisée pour la construction de traits plus complexe dans plusieurs contextes.

Cette proposition offre deux intérêts majeurs :

- Les traits utilisés ont divers niveaux de granularité, en fonction de l'exemple observé. Le niveau de détail varie donc en fonction des données observées.
- La construction des traits et l'algorithme d'apprentissage sont le même processus. De plus, les traits étant issus des régularités de l'ensemble d'apprentissage, on présuppose qu'ils sont pertinents par rapport à celui-ci.

Dans le cas d'images, les traits extraits sont à base d'histogramme de couleurs par exemple, mais il est également possible d'utiliser des indicateurs de couleurs nommées (Gong et al. 1996) par exemple pour facilité l'interaction avec l'utilisateur (cf. partie 5.3).

3.2. Les Régularités

Nous proposons de calculer de manière non supervisée les régularités apparaissant dans les caractéristiques extraites de régions des images d'un corpus. Cet apprentissage des régularités concerne les relations entre les attributs qui constituent les exemples, et non seulement les exemples eux-mêmes. Il doit être hiérarchique pour permettre le partage des représentations.

Les régularités extraites sont

- Soit conjonctives : elles sont basées sur des cooccurrences et représentent des corrélations. Elles définissent une *structure* dans les représentations. Un exemple de régularité conjonctive R_c peut indiquer qu'une texture en vague co-occurre souvent avec la couleur bleue.
- Soit disjonctives : elles caractérisent les informations qui sont similaires, ou interchangeables, dans le contexte particulier d'autres régularités. Un exemple de régularité disjonctive R_d peut composer une régularité conjonctive exprimant qu'une texture verticale co-occurre souvent avec une couleur bleue ou verte.

3.3. *Apprentissage non supervisé sur les régularités*

Les régularités représentent les configurations de caractéristiques qui apparaissent de manière récurrente dans le corpus d'image considéré. Tout le problème réside dans la construction de ces régularités. Nous décrivons ici comment nous les générons par l'intermédiaire de deux phases : le phase de propagation, et la phase de renforcement.

Les régularités sont représentées dans un graphe de composition de régularités. Chaque nœud correspond à une régularité, exprimable par une forme normale conjonctive des entrées du graphe. Les nœuds entrées du graphe (nœuds feuilles, représentés par des carrés dans les figures) sont des régularités atomiques. Un nœud entrée est associé à chaque caractéristique extraite d'une région d'une image. Les deux autres types de nœuds du graphe sont les nœuds *ET* correspondant à des régularités conjonctives et les noeud *OU* correspondant à des régularités disjonctives (ces régularités sont représentées par des cercles dans les graphiques). Par souci de simplification, chaque noeud ET possède deux fils, et chaque nœud OU possède deux fils ou plus. A chaque présentation d'un nouvel exemple d'apprentissage, tous les nœuds sont dans l'état non activé et non bloqué.

Une grande partie du processus décrit ici est basé sur des calculs de co-occurrences entre les nœuds représentant les régularités. Pour réaliser l'apprentissage des régularités, nous utilisons deux attributs des nœuds : l'attribut d'activation (booléen), qui dénote le fait qu'un nœud existant représente une configuration des nœuds entrée activés, et l'attribut de blocage (booléen), qui détermine si un nœud activé doit ou non être présent lors de la phase de renforcement. Nous allons décrire comment ces attributs sont valorisés dans la suite.

3.3.1. Phase de propagation.

L'objectif de la phase de propagation est de déterminer l'ensemble des régularités déjà connues présentes dans l'exemple présenté au graphe.

Pour chaque nœud d'entrée $n_{e,i}$, on détermine si ce nœud est activé ou non en seuillant par une valeur $t_{e,i}$ la valeur de la caractéristique d'entrée qui lui est associée : si la valeur de caractéristique est au dessus de $t_{e,i}$, alors le nœud $n_{e,i}$ est activé, sinon il n'est pas activé. A la fin de cette première étape, tout nœud d'entrée est soit activé, soit non activé. Le processus de propagation se fait des feuilles vers les racines du graphe de la manière suivante : Un nœud *ET* est activé si tous ses fils sont activés, et non activé sinon, un nœud *OU* est activé si au moins l'un de ses fils est activé, et non activé sinon.

3.3.2. Phase de renforcement.

Lors de la phase de renforcement, les règles de blocage que nous utilisons ont pour objectif de restreindre le nombre de nœuds considérés de manière à ce que ce nombre ne croisse pas exponentiellement en fonction du nombre d'exemples d'apprentissage. Ces règles ne s'applique qu'à des nœuds *ET*, car des nœuds *OU* activés sont toujours bloqués pour favoriser leurs composants activés qui décrivent avec plus de précision les régularités rencontrées. Les trois règles de blocage sont les suivantes :

- Règle 1 (Le tout remplace les parties) : Tout nœud appartenant à un sous arbre ne comprenant que des nœuds de type *ET* activés du graphe est bloqué, sauf la racine du sous-arbre. Cette règle favorise les régularités les plus contraintes à un moment donné, sans utiliser ses composantes conjonctives.

- Règle 2.1 (Règle des multiples niveaux de généralisation) : Les fils *ET* activés d'un noeud *OU* ne sont pas bloqués. Cette règle vise à maintenir divers niveaux d'abstraction dans l'apprentissage.
- Règle 2.2 (Règle des multiples niveaux de généralisation) : Un nœud *ET* n_1, fils d'un nœud *ET* n_0 et dont l'autre fils est un nœud *OU* n_2, n'est pas bloqué. Cette règle complémentaire de la règle 2.1, assure simplement l'exhaustivité de la représentation du vecteur d'entrée.

Une fois que les nœuds du graphes sont bloqués, nous pouvons passer à la création éventuelle des nœuds *ET* et *OU*.

L'apprentissage des nœuds *ET* porte sur les nœuds activés par le vecteur V, mais non bloqués par les règles précédentes. La figure 2 montre un exemple dans lequel une seule cooccurrence est renforcée, les autres nœuds étant désactivés ou bloqués. Dans cette figure, le seuil de déclenchement des entrées du vecteur V est noté s.

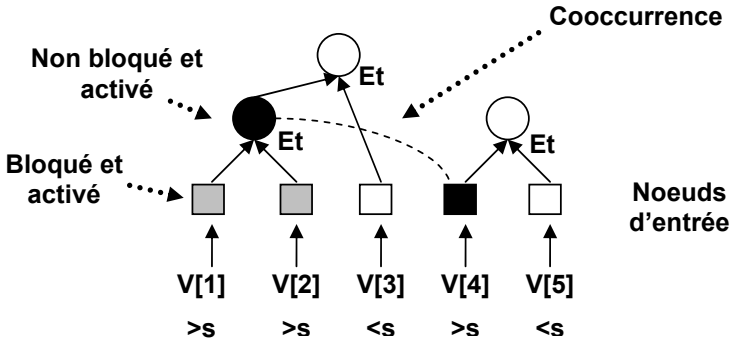

Figure 2 : Un exemple de cooccurrence renforcée.

Quand le nombre de cooccurrences entre deux nœuds dépasse un seuil fixé, un nouveau nœud de type 'Et' est créé, afin de dénoter dans le graphe qu'il s'agit d'une cooccurrence significative. Dorénavant, et en vertu de la règle 1, lorsque cette cooccurrence se produira, seul le nœud créé sera pris en compte, les nœuds fils seront bloqués. La figure 3 schématise la règle de création de nœud.

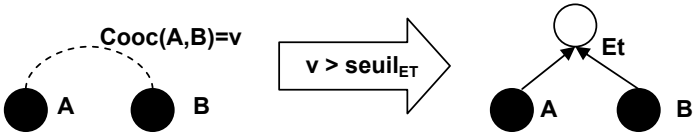

Figure 3 : Création d'un nœud de type 'Et'.

Lorsqu'un nœud n est créé, des liens de cooccurrences le sont également, afin de permettre l'apprentissage des cooccurrences entre n et les nœuds existants. n est

donc relié à tous les nœuds existants, à l'exception de ses descendants. Dans le cas où l'arbre des nœuds *ET* est un arbre binaire complet, la complexité de la mise à jour des valeurs de cooccurrences est en $O(i^2)$ avec i le nombre d'entrées, pour un exemple présenté.

La mise à jour ou création des nœuds *OU* intervient après le traitement des conjonctions. Un nœud disjonctif est créé lorsque deux nœuds conjonctifs ont un constituant commun. Si plus de deux nœuds conjonctifs ont un constituant commun, le nœud disjonctif existe déjà et il est simplement mis à jour.

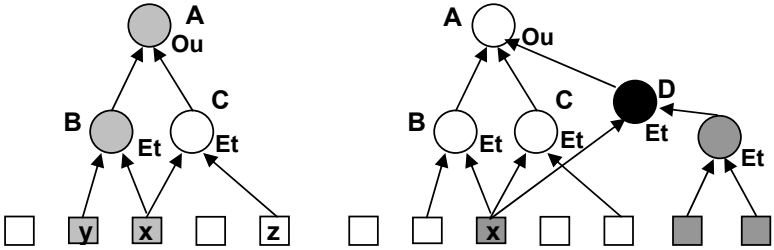

Figure 4 : a. Création du nœud OU. b. Mise à jour d'un nœud OU.

La figure 4a illustre la création d'un nœud 'Ou' : lors de la phase de renforcement des cooccurrences, une cooccurrence a dépassé le seuil de création de nœud et le nœud de type 'Et' B a été créé. Or, il existait déjà un nœud de type 'Et' (le nœud C) ayant pour fils un constituant commun 'x'. Le nœud A, de type 'Ou', est créé, et représente une abstraction par rapport à ses fils B et C, c'est-à-dire que dans le contexte 'x', 'y' et 'z' sont interchangeables. Etant une généralisation de B et C, A est une régularité plus probable. Par conséquent, il est plus probable que A serve de constituant à de nouvelles régularités que B et C. La figure 4b présente la mise à jour d'un nœud 'Ou' à partir de la configuration 4.a : un nœud conjonctif D vient d'être construit, et il a pour constituant un nœud x qui est lui-même constituant de deux autre nœuds conjonctifs (B et C). Le nœud disjonctif A reflète déjà ce constituant commun. La mise à jour consiste donc à ajouter une entrée à A, en provenance de D.

3.4. *Apprentissage supervisé avec les régularités*

L'apprentissage supervisé consiste à comptabiliser les cooccurrences entre concepts et régularités. La figure 5 illustre ce principe avec 6 régularités (de R_1 à R_6) et 3 termes (de C_1 à C_3) : à chaque fois qu'un vecteur, instance d'un terme C_i est présenté au réseau, contient une régularité R_j, le compteur R_jC_i est incrémenté. Pour déterminer si un vecteir contient une régularité, on effectur un parcours à partir de entrées du graphe. Le résultat de l'apprentissage supervisé est simplement

l'ensemble de valeurs R_iC_i. Dans la figure 5, deux flèches montrent que la régularité R_2 participe à la détermination de C_2 mais également de C_3, et on remarque que la régularité R_3 n'est utilisée pour la détermination d'aucun des 3 termes.

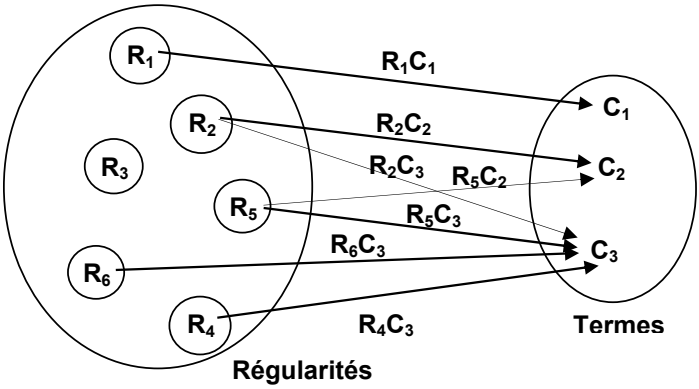

Figure 5 : Un apprentissage avec 6 régularités et 3 termes.

4. Classification à base de régularités non supervisées

La classification que nous proposons est également simple : un vecteur inconnu V est tout d'abord traduit en termes de régularités. Pour cela, nous appliquons la phase de propagation ainsi que les règles de blocage vues précédemment.

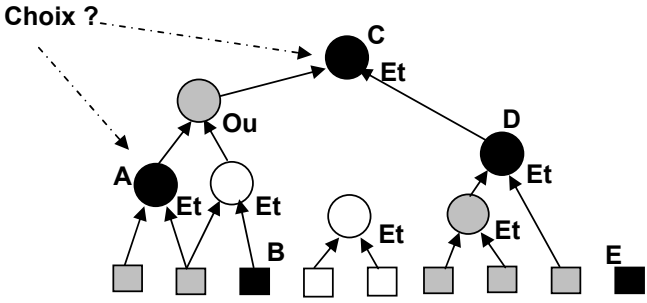

Figure 6 : Choix entre divers niveaux d'abstraction pour la classification.

Cependant, contrairement à l'apprentissage non supervisé, nous ne pouvons pas ici conserver plusieurs niveaux d'abstraction dans la représentation du vecteur. Par exemple, la figure 6 montre la traduction d'un vecteur après la phase de propagation et l'application des règles de blocage. Cette traduction présente une redondance car les deux nœuds indiqués par les flèches possèdent des constituants communs. Dans la mesure où la classification d'un vecteur nécessite que celui-ci soit représenté de manière unique et non ambiguë, nous forçons le choix d'un seul de ces nœuds. L'idée est de choisir la représentation qui donne lieu à la classification la moins ambiguë.

La première étape calcule l'activation de chaque terme, en représentant le vecteur de la manière la plus abstraite possible. Dans la figure, la représentation la plus abstraite est (C , E), car A, B et D sont à la fois des constituants et des spécifiques de C. L'activation d'un terme j se calcule de la manière suivante :

$$Activation(Cj) = \sum_i \left(\frac{R_i C_j}{\sum_k R_k C_j} \right)$$

Afin de mesurer l'ambiguïté de la classification, nous nous intéressons à la valeur de la différence d'activation entre les deux termes les plus activés. Plus cette différence est élevée, moins la classification sera considérée comme ambiguë.

L'activation des termes est une nouvelle fois calculée, en représentant le vecteur de manière plus spécifique. Dans la figure 7, la représentation immédiatement plus spécifique est (A, B, D, E). Si cette représentation est moins ambiguë, l'itération continue et l'activation des termes est calculée pour une représentation encore plus spécifique. Si par contre, cette représentation est plus ambiguë, la représentation précédente est conservée. Cette procédure est appliquée de la représentation la plus générique à la plus spécifique, et s'arrête lorsque la représentation spécifique est plus ambiguë que la représentation générique. L'algorithme de classification tente donc d'abord de reconnaître une observation comme « un tout » et, cas d'ambiguïté, décompose ce tout en « parties » puis évalue à nouveau l'observation.

5. Expérimentations

5.1. *Evaluation de l'apprentissage non-supervisé*

Cette première expérimentation a pour but d'évaluer l'intérêt de découvrir et utiliser des régularités. Pour cela, nous avons choisi de travailler sur 4 termes visuellement similaires, donc a priori difficilement séparables. Dans cette expérimentation, nous n'utilisons pas d'images entière mais directement des

régions. Ces régions sont des blocs de 64×64 pixels extraits d'images de la base utilisée dans (Paterno et al., 2004). Les classes choisies sont : Champ, Feuillages, Fleur, Herbe.

Nous extrayons de chaque bloc un vecteur de 120 valeurs : a) 100 valeurs couleurs extraites dans l'espace HSV, b) 10 réels compris entre 0 et 1 représentant la rugosité de la texture (Tamura 78). Ces valeurs sont normalisées de manière à ce que la somme des 10 valeurs soit égale à 1, et c) 10 réels compris entre 0 et 1 caractérisant la directionnalité (Tamura 78) sur 5 directions. Ces valeurs sont normalisées de manière à ce que la somme des 10 valeurs soit égale à 1.

Pour chaque terme, nous utilisons 120 exemples d'apprentissage, pour le test nous avons entre 12 et 94 échantillons. Chaque test est répété 10 fois, les exemples étant répartis aléatoirement entre l'ensemble d'apprentissage et l'ensemble de test.

Nus nous comparons l'algorithme 1-NN, suivant les recommandations de (Jai et al. 2000). Pour 1-NN, la précision moyenne est de 61,2% et le rappel moyen de 66,2%, ce qui signifie que 1-NN identifie les régions correctement 6 fois sur 10, et qu'en tout on reconnaît 66% de toutes les bonnes étiquettes.

Nous avons tout d'abord testé notre approche en se limitant à des régularités atomiques avec un seuil de 0.1, sans utiliser d'apprentissage non supervisé, et nous avons obtenus une valeur de précision moyenne de 64,4% et un rappel moyen de 65,4%. Nous obtenons donc de meilleurs résultats en terme de précision moyenne.

Les résultats de l'utilisation de notre approche complète sont présenté en table 1 (les deux premières lignes présentant les résultats préliminaires 1-NN et sans apprentissage non supervisé). Dans ce tableau, nous indiquons les résultats obtenus en faisant varier le nombre de nœuds no atomiques en plus des 120 en entrée. Les résultats montrent clairement que les régularité apprises améliorent le résultats : pour 280 nœuds on passe à plus de 70% en rappel et en précision. Un test de Student d'égalité des espérances pour la précision montre que les différences obtenues pour les valeurs de rappel entre notre meilleur résultat et 1-NN sont statistiquement significatives ($t=5,9^{E}$-7 pour la précision, t=0,001 pour le rappel).

	# noeuds	Précision	Rappel
1-NN	~	61.2	66.2
Notre approche	0	64.4	65.4
	160	68.8	68.2
	280	**72.2**	**71.2**
	360	70.1	69.4

Table 1. Résultats 1-NN, avec régularités atomiques et non atomiques.

5.2. *Impact de la taille de l'ensemble d'apprentissage*

Le but ici est d'évaluer la réactivité de notre approche, c'est-à-dire sa capacité à généraliser avec peu d'exemples. Afin de rendre comparables les résultats obtenus, nous utilisons parallèlement l'algorithme 1-NN pour l'apprentissage et la classification. Nous avons utilisé une base réelle de 744 photographies personnelles. Chaque image est découpée en 100 blocs rectangulaires sans chevauchement, et ces blocs on été étiquetés à partir d'une annotation manuelle des régions des images.

Nous utilisons toutes les images de la collection. Dans une première phase, un apprentissage non supervisé est effectué sur 5000 blocs choisis aléatoirement dans la collection. Cette phase a pour but d'apprendre les régularités présentes dans la collection. Dans une seconde phase, nous simulons l'interaction utilisateur/système : des images sont sélectionnées au hasard dans la collection (par groupe de 5) et sont considérées comme indexées par l'utilisateur. Les blocs provenant de ces images sont utilisés pour l'apprentissage, d'une part selon notre approche et d'autre par selon le 1-NN. Ces images sont ensuite éliminées et la classification, selon les deux approches, est effectuée sur les images restantes. Dans cette expérimentation, nous limitons le nombre de termes aux 15 termes les plus représentés dans la collection : arbre avec feuilles, cailloux, ciel brumeux, ciel clair, ciel couvert, ciel crépusculaire, ciel de nuit, cours d'eau, façade d'immeuble, feuillage, nuage, herbe, mer, rocher, terre. Nous limitons le nombre d'images indexées à 50 pour deux raisons : la première est que l'indexation manuelle de 50 images par un utilisateur, même patient, est déjà à la limite du réalisme, l'autre raison étant que nous désirons surtout caractériser la réactivité initiale de l'apprentissage.

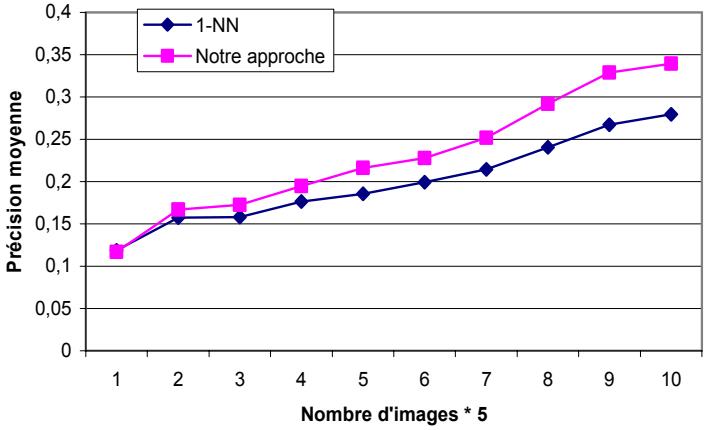

Figure 7 : Précision moyenne en fonction du nombre d'images indexées selon 15 termes.

La figure 7 montre la précision moyenne obtenue pour les deux méthodes en fonction du nombre d'images indexées. Ces courbes montrent que si les précisions initiales (après apprentissage de 5 images), sont quasi identiques (12%), l'écart se creuse au fur et à mesure de l'augmentation de l'ensemble d'apprentissage, jusqu'à atteindre 6% d'écart (entre 28% et 34%).

En ce qui concerne l'évolution du rappel en figure 8, les deux courbes sont décalées mais évoluent parallèlement, avec toujours un avantage clair pour notre proposition. Nous avons effectué 5 cycles apprentissage/classification, et un test de Sudent montre que les différences pour la précision et le rappel sont significatives (resp. $t = 0.001$, et $t = 7^{E}\text{-}08$).

Nous avons donc montré que notre proposition obtenait toujours de meilleurs résultat que le 1-NN, même avec très peu d'images exemples, ce qui était l'un de nos objectifs.

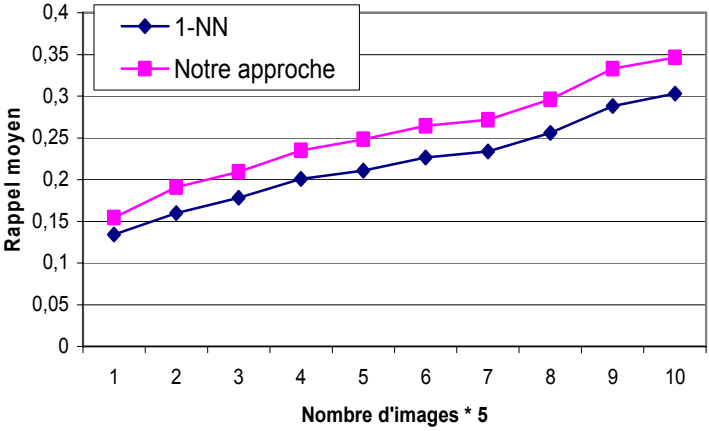

Figure 8 : Rappel moyen en fonction du nombre d'images indexées selon 15 termes.

5.3. Visualisation de régularités pour personnalisation

Nous présentons en figure 9 un exemple réel d'apprentissage sur le terme "ciel de nuit" effectué sur la collection décrite en 5.2. Cette figure montre qu'il est relativement simple de comprendre les régularités apprises, pour éventuellement ajouter de nouveaux exemples pour améliorer les résultats d'annotation d'images. On

y voit que qu'il y a peu de couleurs utilisées (le teintes sont une entrée de rouge, une de rouge-vert et de nombreuses entrées de teintes violet-bleue), que toutes les saturations sont utilisées (ce qui est en fait courant quand on rencontre des couleurs très peu lumineuses), que les couleurs sont peu lumineuses (une seule entrée de la luminosité la plus sombre). Au niveau des textures, on voit qu'il y a des motifs peu texturés (une entrée de rugosité est utilisée : celle qui est la plus faible), et que les directions de textures sont plutôt horizontales (deux premières et dernières entrées de rugosité utilisées). D'autre part, on se rend compte sur cette figure que les premières régularités non atomiques crées (celles les plus courantes par construction), sont liée à la teinte violet-bleue, à la saturation très sombre et au fait qu'il y a peu de texture, ce qui reflète bien l'idée de la description d'un ciel de nuit. Par ailleurs, la caractéristique apportant le moins d'information (la saturation) est prise en compte en dernier.

A partir d'un telle visualisation, enrichie d'autres informations, nous estimons qu'un utilisateur peut être à même de se rendre compte que le système n'a pas pris compte des caractéristiques car l'utilisateur ne lui a pas donné d'exemples suffisants. L'utilisateur peut donc alors interagir avec le système de manière simple pour ajouter des exemples pertinents. Ce point est fondamental si nous voulons proposer des moyens de personnaliser les indexations des images.

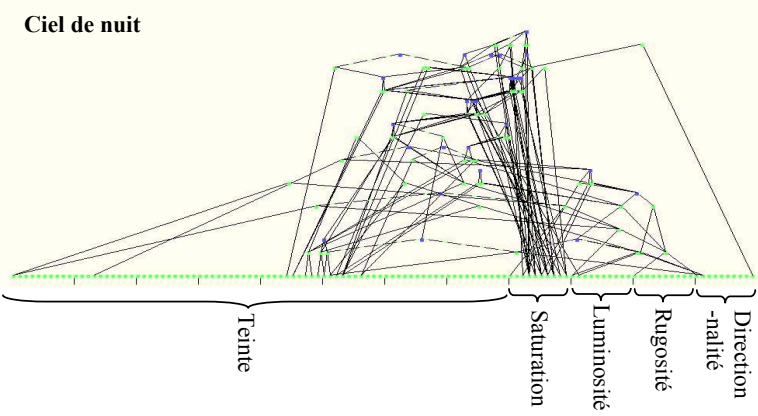

Figure 9. Exemple de régularités générées pour "ciel de nuit".

6. Conclusion

La proposition que nous avons faite pour l'indexation des images intègre un aspect d'apprentissage non supervisé suivi d'une approche supervisée. La partie non supervisée permet d'extraire et de structurer les caractéristiques qui occurrent dans

un corpus d'images en utilisant des régularités, et la seconde étape permet d'associer des termes à ces régularités. Les objectifs fixés étaient de permettre une indexation de bonne qualité sans nécessiter de trop nombreux exemples, et de fournir des bases pour qu'un utilisateur non expert puisse interagir avec le système d'apprentissage. Les résultats obtenus sont positifs et montrent l'intérêt de notre approche. Même si notre proposition ne garanti pas d'obtenir le meilleur apprentissage, elle permet de mettre en place une interaction avec l'utilisateur pour une personnalisation de l'indexation, ce qui est un atout non négligeable pour une telle approche.

A partir du travail décrit ici, nous devons étudier l'impact de certains paramètres sur nos propositions : l'étude de différentes caractéristiques signal fournies au processus d'apprentissage de régularités devra permettre de définir si un certaine indépendance vis-à-vis de ces caractéristiques existe pour l'étiquetage, l'étude du comportement de nos propositions avec de plus grands ensembles de termes et dans des cas plus difficiles pour l'apprentissage que jusqu'à présent, ainsi que l'étude du comportement général en fonction de l'ordre des exemples fournis.

Nous allons étendre cette approche à des vocabulaires plus riches, et intégrer ces résultats dans un système de recherche d'images. Nous allons aussi étudier plus en détails comment une interaction réelle avec un utilisateur peut amener à corriger des annotations incorrectes dues à un mauvais ensemble d'apprentissage fourni, de manière à proposer un véritable environnement pour l'indexation et la recherche de photographies personnelles.

7. Bibliographie

Barnard K., Duygulu P., Forsyth D. A., de Freitas N., Blei D. M., Jordan M. I., Matching Words and Pictures. Journal of Machine Learning Research 3: 1107-1135 (2003).

Gong, Y. & Chuan, H. & Xiaoyi, G., Image Indexing and Retrieval Based on Color Histograms. Multimedia Tools and Applications II, p. 133-156, 1996.

Jain A., Duin P., Mao J., Satistical Pattren Recognition : A Review, IEEE Trans. On PAMI 22(1), 2000, pp. 4-37.

Jeon, J., Lavrenko V., Manmatha M., Automatic Image Annotation and Retrieval using CrossMedia Relevance Models, SIGIR'03, July 28–August 1, 2003, Toronto, Canada, pp. 119-126.

Lim, J.H., Building visual vocabulary for image indexation and query formulation. Pattern Analysis and Applications (Special Issue on Image Indexation), 4(2/3): 125-139, 2001.

Mooney, R. J., Encouraging experimental results on learning CNF. Machine Learning, Vol. 19, 1, 1995, pp. 79-92.

Mulhem P., Lim J.-H., Leow W. –K., Kankanhalli M.S., Advances in Digital Home Photo Albums, Chapter IX of Multimedia Systems and Content-Based Image Retrieval, S. Deb Editor, IDEA Publishing, pp. 201-226, 2004.

Paterno M. C. S., Lim F. S., Leow W.-K., Fuzzy semantic Labeling for Image Retrieval, IEEE ICME, 2004, pp. 767-770.

Smeulders A. W. M., Worring M., Santini S., Gupta A., Jain R.: Content-Based Image Retrieval at the End of the Early Years. IEEE Trans. Pattern Anal. Mach. Intell. 22(12): 1349-1380 (2000).

Snoek C. G.M., Worring M., van Gemert J. C., Geusebroek J.-M., Smeulders A. W. M., The Challenge Problem for Automated Detection of 101 Semantic Concepts in Multimedia, ACM Multimedia, Santa Barbara, USA, October 2006, pp. 421430.

Tollari S., Une méthode de recherche des caractéristiques vbisuelles d'un mot, Chapitre 6 de Indexation et recherche d'images par fusion d'informations textuelles et visuelles, thèse de doctorat, Université du Sud Toulon-Var, octobre 2006.

Town C.P., Sinclair D., Content based image retrieval using semantic visual categories. Technical Report TR2000-14, AT&T Laboratories Cambridge, 2000.