Modélisation, classification et annotation d'images partiellement annotées avec un réseau Bayésien

Sabine Barrat, Salvatore Tabbone

Modélisation, classification et annotation d'images partiellement annotées avec un réseau Bayésien

Modeling, classifying and annotating weakly annotated images using Bayesian network

S. Barrat S. Tabbone

LORIA-UMR 7503
Université de Nancy 2
BP 239, 54506 Vandœuvre-les-Nancy

email : {barrat,tabbone@loria.fr}

Résumé

Dans cet article, nous proposons un modèle graphique probabiliste pour représenter des images partiellement annotées. Nous considérons une image comme partiellement annotée si elle ne possède pas le nombre maximal de mots-clés disponibles pour une image dans la vérité-terrain. Ce modèle est utilisé pour classifier des images et étendre automatiquement les annotations existantes à de nouvelles images, en prenant en compte les éventuelles relations sémantiques entre mots-clés. La méthode proposée a été évaluée en classification visuo-textuelle et en extension automatique d'annotations. La classification visuo-textuelle correspond à la classification effectuée en utilisant à la fois l'information visuelle et l'information textuelle, quand elle est disponible. Les résultats expérimentaux, obtenus à partir d'une base de plus de 30000 *images, montrent une amélioration de* 50.5% *en moyenne, en terme de taux de reconnaissance, par rapport à la classification basée sur l'information visuelle seule. La prise en compte des éventuelles relations sémantiques entre mots-clés améliore le taux de reconnaissance de* 10.5% *en moyenne et le taux de bonnes annotations de* 6.9% *en moyenne. Enfin, la méthode proposée s'est montrée compétitive, expérimentalement, avec des classificateurs de l'état de l'art.*

Mots Clefs

modèles graphiques probabilistes, réseaux Bayésiens, classification d'images, annotation d'images, similarité sémantique, Wordnet.

Abstract

In this paper, we propose a probabilistic graphical model to represent weakly annotated images. We consider an image as weakly annotated if the number of keywords defined for it is less than the maximum defined in the ground truth. This model is used to classify images and automatically extend existing annotations to new images by taking into account semantic relations between keywords. The proposed method has been evaluated in visual-textual classification and automatic annotation of images. The visual-textual classification is performed by using both visual and textual information. The experimental results, obtained from a database of more than 30000 *images, show an improvement by* 50.5% *in terms of recognition rate against only visual information classication. Taking into account semantic relations between keywords improves the recognition rate by* 10.5% *and the mean rate of good annotations by* 6.9%. *Finally, the proposed method is experimentally competitive with the state-of-art classifiers.*

Keywords

probabilistic graphical models, Bayesian networks, image classification, image annotation, semantic similarity, Wordnet.

1 Introduction

La croissance rapide d'Internet et de l'information multimédia a engendré un besoin en techniques de recherche d'information multimédia, et plus particulièrement en recherche d'images. On peut distinguer deux tendances. La première, appelée recherche d'images par le texte, consiste à appliquer des techniques de

recherche de textes à partir d'ensembles d'images complètement annotés. L'efficacité de ces méthodes est étroitement liée à la qualité de l'indexation des images. Or, les méthodes d'indexation textuelle automatiques sont peu performantes et fournissent des ensembles d'images mal annotées, car elles utilisent l'URL, le titre de la page, ou d'autres attributs ou le texte proche de l'image dans le cas d'images provenant d'Internet, ou alors tout simplement le nom de l'image dans le cas d'images issues de collections personnelles. Quant à l'indexation textuelle manuelle, bien qu'elle soit plus performante que l'indexation textuelle automatique, elle est très coûteuse pour l'utilisateur et se révèle pratiquement inapplicable aux grandes bases d'images. De plus, les annotations obtenues manuellement peuvent être ambiguës car deux utilisateurs peuvent utiliser différents mots-clés pour décrire la même image. Par conséquent des approches utilisant Wordnet [6] ont été proposées [7] afin de réduire ces ambiguïtés potentielles. La seconde approche, appelée recherche d'images par le contenu, est un domaine plus récent et utilise une mesure de similarité (similarité de couleur, forme ou texture) entre une image requête et une image du corpus utilisé. Ces méthodes sont efficaces sur certaines bases d'images, mais leurs performances décroissent sur des bases d'images plus généralistes. Afin d'améliorer la reconnaissance, une solution consiste à combiner différentes sources d'informations. Dans un premier temps sont apparues des approches de combinaison de caractéristiques [15] et de combinaison de classificateurs[10]. Dans le cas de la combinaison de caractéristiques, un seul classificateur est utilisé pour combiner plusieurs caractéristiques. Au contraire, les approches de combinaison de classificateurs prennent une décision globale à partir des décisions individuelles prises par chaque classificateur. Par ailleurs, des approches de combinaison d'informations visuelles et sémantiques, appelées "approches visuo-textuelles", ont été proposées. L'annotation d'images par mots-clés constitue une manière possible d'associer de la sémantique à une image. En effet, elle consiste à assigner à chaque image, un mot-clé ou un ensemble de mots-clés, destiné(s) à décrire le contenu sémantique de l'image. Ainsi cette opération peut être vue comme une fonction permettant d'associer de l'information visuelle, représentée par les caractéristiques de bas niveau (forme, couleur, texture, ...) de l'image, à de l'information sémantique, représentée par ses mots-clés, dans le but de réduire le fossé sémantique ("semantic gap" en anglais) [13]. Donc, il est possible d'obtenir des bases d'images annotées et des approches visuo-textuelles peuvent être mises en place. Ces approches ont déjà fait l'objet de nombreux travaux [9]. Cependant, la plupart de ces méthodes présentent l'inconvénient majeur de nécessiter que toute la base d'images soit entièrement annotée. Or, de telles bases sont très difficiles à obtenir car elles requièrent un travail coûteux d'annotation manuelle de la base (les méthodes d'indexation textuelle automatique étant moins performantes que l'annotation manuelle). On préférera donc s'orienter vers des méthodes dédiées à des bases d'images partiellement annotées. De plus, des techniques d'annotation automatique d'images pourront être utilisées afin de compléter les annotations des images partiellement annotées. En effet, l'annotation automatique d'images peut être utilisée dans les systèmes de recherche d'images, pour organiser et localiser les images recherchées ou pour améliorer la classification visuo-textuelle. Cette méthode peut être vue comme un type de classification multi classes avec un grand nombre de classes, aussi large que la taille du vocabulaire. Plusieurs travaux ont été proposés dans ce sens. On peut citer, sans être exhaustif, les méthodes basées sur la classification [14], les méthodes probabilistes [4] et l'affinement d'annotations [12]. Dans cette direction, la contribution de ce papier est de proposer une méthode pour optimiser la classification d'images, en utilisant une approche de classification visuo-textuelle et en étendant automatiquement des annotations existantes. Plus précisément, le modèle présenté ici est dédié aux deux tâches de classification et d'annotation d'images partiellement annotées (images comportant moins de mots-clés que le nombre maximal de mots-clés disponibles dans la vérité-terrain pour une image). En effet, la plupart des méthodes de classification visuo-textuelles sont efficaces, mais requièrent que toutes les images, ou régions d'images, soient annotées. De plus, la plupart des modèles d'annotation automatique existants ne sont pas capables de classifier des images, car ils sont uniquement dédiés à l'annotation. Le modèle que nous proposons ne nécessite pas que toutes les images soient annotées : quand une image est partiellement annotée, les mots-clés manquants sont considérés comme des données manquantes. Notre modèle permet aussi d'étendre automatiquement des annotations existantes à des images partiellement annotées, sans l'intervention de l'utilisateur. Le modèle [1] est le plus proche de notre approche, car il permet de classifier des images sur la base de caractéristiques visuelles et textuelles, et d'annoter automatiquement de nouvelles images. Cependant, notre modèle est moins restrictif pour l'utilisateur. En effet, notre classificateur ne nécessite pas que toutes les images soient annotées. De plus, le modèle [1] suppose que les mots-clés sont indépendants étant donnés leurs parents. Au contraire, notre modèle a l'avantage de prendre en compte les éventuelles relations sémantiques entre mots-clés. En effet, des relations sémantiques, comme définies dans Wordnet, sont représentées par des arcs dans notre réseau Bayésien. Nous montrons que la prise en compte de ces relations améliore le taux de reconnaissance aussi bien que le

taux de bonnes annotations. Le reste de ce papier est organisé de la façon suivante : dans la section 2, les propriétés des classificateurs basés sur les réseaux Bayésiens sont introduites et nous conduisent à présenter notre propre réseau Bayésien pour la classification et l'extension d'annotations d'images (dans la section 3). Les résultats expérimentaux, obtenus sur une base de plus de 30000 images, sont présentés section 4. Une comparaison avec le modèle GM-Mixture [1] est aussi fournie. Enfin, les conclusions et perspectives de ce travail sont données section 5.

2 Représentation et classification d'images

2.1 Contexte et objectifs

Dans ce travail, nous nous intéressons à la classification et à l'extension d'annotations d'images partiellement annotées. Étant donnée une base d'images, nous essayons de reconnaître l'objet représenté par l'image. Ce problème de reconnaissance peut être vu comme un problème de classification : notre but est d'affecter chaque image à une classe correspondant à un objet donné. Cependant nous ne disposons pas de modèle pour chaque classe. Par conséquent, nous ne pouvons pas nous contenter de minimiser une distance entre chaque image de la base et chaque modèle. Par contre, cette tâche de classification peut être résolue en utilisant une méthode d'apprentissage supervisée, à partir d'un sous-ensemble des images de la base pour lesquelles les étiquettes de classe sont connues. De plus, de façon à décrire plus précisément les images et d'améliorer le taux de reconnaissance, nous proposons de combiner deux descripteurs (un de forme et un de couleur), afin de représenter l'information visuelle contenue dans l'image, et d'utiliser les mots-clés annotant certaines images afin de prendre en compte l'information sémantique qu'elles véhiculent. Les descripteurs fournissent en général des vecteurs de caractéristiques continues, et les mots-clés sont considérés comme des variables discrètes. Il semble donc approprié de proposer un classificateur qui permet de combiner caractéristiques discrètes et continues. De plus, le classificateur proposé se doit d'être robuste aux données manquantes, car toutes les images de la base ne sont pas annotées ou ne le sont que partiellement. La plupart des méthodes de classification ne permettent de traiter que les données discrètes et requièrent ainsi un pré-traitement de discrétisation des données de façon à transformer chaque variable à valeurs continues en variable à valeurs discrètes. Cependant, il existe quelques méthodes de classification permettant de combiner les deux types de variables. C'est le cas, par exemple, des Machines à Vecteurs Supports SVM [3], des forêts aléatoires [2], de l'algorithme des k plus proches voisins (notés KPPV, ou KNN en anglais), et des classificateurs Bayésiens. Les SVM et les forêts aléatoires sont réputés pour être performants en présence d'un grand nombre de variables. Par contre, l'utilisation des SVM devient difficile lorsque le nombre d'observations de la base d'apprentissage est important. Concernant les KNN, la procédure de classification est lourde car chaque image requête est comparée (sur la base de ses caractéristiques) à toutes les images stockées. Par contre cette méthode a l'avantage de ne pas nécessiter d'apprentissage : c'est l'échantillon qui constitue le modèle. Enfin, les classificateurs Bayésiens, quant à eux, sont sensibles à la dimensionnalité des données. Par contre, ils sont efficaces avec beaucoup de données d'apprentissage. Enfin, les classificateurs Bayésiens sont adaptés à la résolution de problèmes en présence de données manquantes, contrairement aux SVM. Par conséquent, nous avons choisi de construire un classificateur Bayésien pour sa capacité à combiner variables discrètes et continues, en présence de nombreuses données d'apprentissages et de données manquantes. De plus, nous montrerons (voir section 4) que ce classificateur Bayésien est compétitif avec le réseau Bayésien décrit dans [1]. Enfin, le modèle proposé sera utilisé pour étendre des annotations existantes à des images sans mots-clés ou partiellement annotées, afin d'augmenter le nombre d'annotations de la base existant, en vue d'effectuer des classifications visuotextuelles plus efficaces.

2.2 Les classificateurs Bayésiens

Soit I une image caractérisée par une observation particulière $f = \{f_1, ..., f_n\}$ d'un vecteur caractéristique $F = \{F_1, ..., F_n\}$. Notre but est d'affecter l'image I à la classe c_i parmi k classes. Chaque c_i est une observation particulière de la variable C. Le Naïve Bayes (NB) est un simple algorithme de classification probabiliste qui a montré de bonnes performances dans de nombreux domaines. Ce classificateur encode la distribution $P_{NB}(F_1, ..., F_n, C)$, d'un échantillon d'apprentissage donné (composé de données étiquetées). Le modèle probabiliste résultant peut être utilisé pour classifier une nouvelle observation I. En effet, la règle de Bayes est appliquée pour calculer la probabilité de c_i étant donnée l'observation f. Le classificateur basé sur le modèle NB retourne la classe c_i, $i \in \{1, ..., k\}$, qui maximise la probabilité *a posteriori* $P_i = P_{NB}(c_i|f_1, ..., f_n)$, où $P_i = \frac{P_{NB}(f_1,...,f_n|c_i) \times P_{NB}(c_i)}{P_{NB}(f_1,...,f_n)}$ et $P_{NB}(f_1, ..., f_n) = \sum_{j=1}^{k} P_{NB}(f_1, ..., f_n|c_j) \times P_{NB}(c_j)$. Cependant, nous nous intéressons aux distributions de probabilités de caractéristiques discrètes et continues, et de leurs relations de dépendance conditionnelle. Considérons chaque composante des vecteurs de caractéristiques continues comme une variable discrète continue et les valeurs discrètes provenant des mots-clés comme des variables discrètes. Ce modèle est trop grand (il possède trop de variables) pour être

représenté par une unique distribution de probabilité jointe. Par conséquent, il est nécessaire d'introduire de la connaissance structurelle *a priori* : le Naïve Bayes doit être étendu pour prendre en compte les variables discrètes et continues. Les modèles graphiques probabilistes, et en particulier les réseaux Bayésiens, sont un bon moyen de résoudre ce genre de problème. En effet, dans un réseau Bayésien, la distribution de probabilité jointe est remplacée par une représentation graphique des relations entre variables, uniquement pour les variables s'influençant les unes les autres. Les interactions indirectes entre variables sont ensuite calculées en propageant la connaissance à travers le graphe de ces connections directes. Par conséquent, les réseaux Bayésiens sont un moyen simple de représenter une distribution de probabilité jointe d'un ensemble de variables, de visualiser les propriétés de dépendance conditionnelle et d'effectuer des calculs complexes comme l'apprentissage ou l'inférence, grâce à des manipulations graphiques.

3 Modèle proposé

Nous présentons un modèle hiérarchique probabiliste multimodal (images et mots-clés associés) pour classifier de grandes bases de données d'images annotées. Nous rappelons que les caractéristiques visuelles sont considérées comme des variables continues, et les mots-clés du vocabulaire comme des variables discrètes. De plus, on considère que notre échantillon de caractéristiques visuelles suit une loi dont la fonction de densité est une densité de mélange de Gaussiennes. Les variables discrètes sont supposées suivre une loi de Bernoulli. En effet, pour une image donnée, chaque variable mot-clé peut prendre deux états : "présent" quand le mot annote l'image donnée, "absent" sinon. La distribution de Bernoulli associe une probabilité à p à la présence du mot-clé dans l'annotation et une probabilité $1-p$ à son absence. Cette distribution a été préférée à une distribution multinomiale (utilisée dans [1]) car elle nous permet de représenter des dépendances entre mots-clés. De plus, aucun nombre maximal de mots-clés par annotation n'est fixé à la création du modèle. Au contraire, la distribution multinomiale considère chaque mot-clé d'une annotation comme une variable discrète et associe une probabilité à chaque mot-clé du vocabulaire. Il est donc nécessaire de fixer un nombre maximal de mots-clés par image à la création du modèle (car ce nombre est égal au nombres de variables mots-clés). De plus, elle ne permet pas de représenter des dépendances entre mots-clés du vocabulaire.

Nous proposons d'étendre le Naïve Bayes afin de prendre en compte ces distributions de probabilités : le modèle proposé est un modèle de mélange de lois de Bernoulli et de mélanges de Gaussiennes (noté "modèle de mélange GM-B"). La structure du Naïve Bayes est conservée c'est-à-dire que l'on dispose d'une variable "Classe", connectée à chaque variable caractéristique (cf. Figure 2). Soit F un échantillon d'apprentissage composé de m individus $f_{1_i}, ..., f_{m_i}, \forall i \in \{1, ..., n\}$, où n est la dimension des signatures obtenues par concaténation des vecteurs caractéristiques issus du calcul des descripteurs sur chaque image de l'échantillon. Chaque individu $f_j, \forall j \in \{1, ..., m\}$ est caractérisé par n variables continues. Comme nous l'avons vu dans la section 2.1, nous sommes dans le cadre d'une classification supervisée. Les m individus sont donc divisés en k classes $c_1, ..., c_k$. Soient $G_1, ..., G_g$ les g groupes dont chacun a une densité Gaussienne avec une moyenne $\mu_l, \forall l \in \{1, ..., g\}$ et une matrice de covariance $\sum_l$. De plus, soient $\pi_1, ..., \pi_g$ les proportions des différents groupes, $\theta_l = (\mu_l, \sum_l)$ le paramètre de chaque Gaussienne et $\Phi = (\pi_1, \pi_1, ..., \pi_g, \theta_1, ..., \theta_g)$ le paramètre global du mélange. Alors la densité de probabilité de F conditionnellement à la classe $c_i, \forall i \in \{1, ..., k\}$ est définie par $P(f, \Phi) = \sum_{l=1}^{g} \pi_l p(f, \theta_l)$ où $p(f, \theta_l)$ est la Gaussienne multivariée définie par le paramètre θ_l. Ainsi, nous avons un modèle de mélange de Gaussiennes (GMM) par classe. Ce problème peut être représenté par le modèle probabiliste de la Figure 1, où :

- Le nœud "Classe" est un nœud discret, pouvant prendre k valeurs correspondant aux classes prédéfinies $c_1, ..., c_k$.
- Le nœud "Composante" est un nœud discret correspondant aux composantes (i.e les groupes $G_1, ..., G_g$) des mélanges. Cette variable peut prendre g valeurs, i.e le nombre de Gaussiennes utilisé pour calculer les mélanges. Il s'agit d'une variable latente qui représente le poids de chaque groupe (i.e les $\pi_l, \forall l \in \{1, ..., g\}$).
- Le nœud "Gaussienne" est une variable continue représentant chaque Gaussienne $G_l, \forall l \in \{1, ..., g\}$ avec son propre paramètre ($\theta_l = (\mu_l, \sum_l)$). Il correspond à l'ensemble des vecteurs caractéristiques dans chaque classe.
- Enfin, les arêtes représentent l'effet de la classe sur le paramètre de chaque Gaussienne et son poids associé. Le cercle vert sert à montrer la relation entre le modèle graphique proposé et les GMMs : nous avons un GMM (entouré en vert), composé de Gaussiennes et de leur poids associé, par classe. Chaque GMM a son propre paramètre global.

Maintenant le modèle peut être complété par les variables discrètes, notées $KW\ 1, ..., KW\ n$, correspondant aux éventuels mots-clés associés aux images. Des *a priori* de Dirichlet [11], ont été utilisés pour l'estimation de ces variables. Plus précisément, on introduit des pseudo comptes supplémentaires à chaque instance de façon à ce qu'elles soient toutes virtuellement représentées dans l'échantillon d'apprentissage. Ainsi, chaque observation, même si elle n'est pas représentée dans l'échantillon d'apprentissage, aura une

probabilité non nulle. Comme les variables continues correspondant aux caractéristiques visuelles, les variables discrètes correspondant aux mots-clés sont incluses dans le réseau en les connectant à la variable classe. Notre classificateur peut alors être décrit par la Figure 2. La variable latente "α" montre qu'un *a priori* de Dirichlet a été utilisé. La boîte englobante autour de la variable KW indique n répétitions de KW, pour chaque mot-clé. Les arcs représentant les relations sémantiques entre mots-clés ne sont pas dessinés dans la boîte englobante, dans un soucis de clarté. Par contre, la Figure 3 représente plus précisément les variables mots-clés et leurs éventuelles dépendances. Les n nœuds correspondent aux n mots-clés du vocabulaire : $KW\ 1, ..., KW\ n$. Seul un sous-ensemble des dépendances entre mots-clés est représenté. Par exemple, "bird" et "animal" ont une relation sémantique, qui est représentée par l'arc orienté du nœud "bird" vers le nœud "animal". De la même façon un arc est observé entre les nœuds "duck" et "animal" et les nœuds "duck" et "bird". En général, une arc est ajouté entre deux mots-clés du même groupe sémantique (ou synset, comme défini dans Wordnet [6]). Chaque arc entre deux mots-clés est orienté du mot-clé le plus spécifique (hypernyme) vers le mots-clé le plus général (hyponyme). Concernant les arcs entre deux mots-clés où un mot-clé est une partie de l'autre, l'arc est orienté du mot-clé qui est la part de l'autre (meronyme) vers le mot-clé représentant le "tout" (holonyme). De cette façon nous représentons les ontologies de Wordnet. Quelques dépendances de notre base sont données Tableau 2. La structure de ce modèle, tout comme les relations sémantiques, ont été établies "à la main". Aucun algorithme d'apprentissage de structure n'a été utilisé.

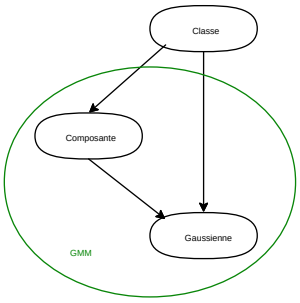


FIG. 1 – GMMs représentés par un modèle graphique probabiliste

3.1 Classification

Pour classifier une nouvelle image f_j, le nœud classe C est inféré grâce à l'algorithme de passage de messages [8]. Ainsi, une image requête f_j, représentée par ses caractéristique visuelles $v_{j_1}, ..., v_{j_m}$ et ses éventuels mots-clés $KW\ 1_j, ..., KW\ n_j$, est considérée comme une observation (aussi appelée "évidence") représentée par :

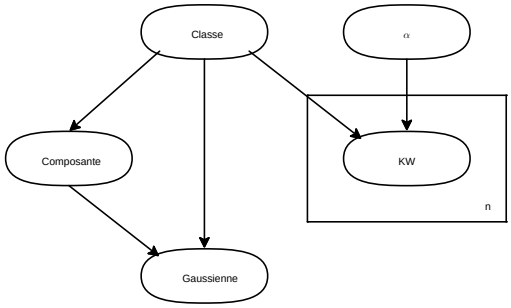


FIG. 2 – Le modèle de mélange de lois de Bernoulli et de mélanges de Gaussiennes

$P(f_j) = P(v_{j_1}, ..., v_{j_m}, KW\ 1_j, ..., KW\ n_j) = 1$ quand le réseau est évalué. En effet, une évidence correspond à une information nous donnant avec une certitude absolue la valeur d'une variable, d'où la probabilité égale à 1. Grâce à l'algorithme d'inférence (i.e. l'algorithme de passage de messages [8]), les probabilités de chaque nœud sont mises à jour en fonction de cette évidence. On parle de "propagation de croyance ou de propagation de l'évidence". Il s'agit de la phase de calcul probabiliste à proprement parler où les nouvelles informations concernant les variables observées sont propagées à l'ensemble du réseau, de manière à mettre à jour l'ensemble des distributions de probabilités du réseau. Ceci ce fait en passant des messages contenant une information de mise à jour entre les nœuds du réseau. A la fin de cette phase, le réseau contiendra la distribution de probabilité sachant les nouvelles informations. Après la propagation de croyances, on connaît donc, $\forall i \in \{1, ..., k\}$, la probabilité *a posteriori* : $P(c_i|f_j) = P(c_i|v_{j_1}, ..., v_{j_m}, DF\ 1_j, ..., DF\ n_j)$. L'image requête f_j est affectée à la classe c_i maximisant cette probabilité. L'algorithme EM, dont le principe général est expliqué en détail dans [5], a été utilisé pour apprendre les paramètres des mélanges de Gaussiennes et les données manquantes correspondant aux mots-clés manquants dans les annotations partielles.

3.2 Extension automatique d'annotations

Étant donnée une image sans mot-clé, ou partiellement annotée, le modèle proposé peut être utilisé pour calculer une distribution des mots-clés conditionnellement à une image et ses éventuels mots-clés existants. En effet, pour une image f_j annotée par $k, \forall k \in \{0, ..., n\}$ mots-clés noté EKW(pour mots-clés Existants), où n est le nombre maximum de mots-clés par image, l'algorithme d'inférence permet de calculer la probabilité *a posteriori* $P(KW_{i_j}|f_j, EKW)\ \forall KW_{i_j} \notin EKW$. Cette distribution représente une prédiction des mots-clés manquants d'une image. Par exemple, considérons le tableau 1 présentant 2 images avec leurs éventuels mots-clés existants et les mots-clés obtenus

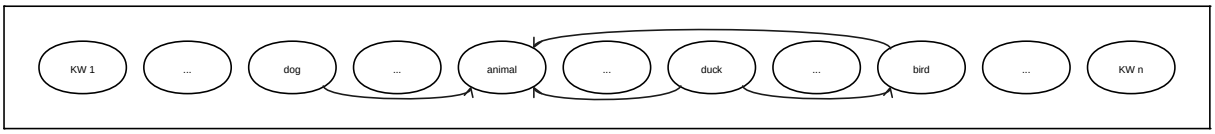
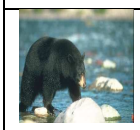


FIG. 3 – Dépendances entre mots-clés

après extension automatique d'annotations avec (colonne 3) ou sans (colonne 2) prise en compte des éventuelles relations sémantiques entre mots-clés. L'annotation de la première image, composée de 3 mots-clés au début, a été étendue par un mot-clé incorrect. En effet, le mot-clé manquant correct est "shrubs". Cette erreur est due au grand nombre d'images de la base annotées par les 4 mots-clés "bear", "black", "water" and "grass", ce qui génère une grande probabilité jointe de cet ensemble de mot-clés. Considérons la seconde image. Son annotation n'a pas été étendue sans prise en compte des relations sémantiques entre mots-clés. Ceci est dû au seuil utilisé pour sélectionner les mots-clés. En effet, un mot-clé est sélectionné comme annotation si la probabilité de ce mot-clé comme annotation est strictement supérieure à un seuil. Dans le cas de la seconde image, sans prise en compte des relations sémantiques entre mots-clés, aucun mot-clé du vocabulaire n'avait de probabilité supérieure au seuil fixé. Au contraire, en prenant en compte les relations sémantiques, la deuxième image a été correctement annotée par le mot-clé "water", grâce à la relation sémantique entre les mots-clés "river" et "water" : l'image, déjà annotée par le mot-clé "river", a vu la probabilité du mot-clé "water" augmenter et dépasser le seuil grâce à cette relation sémantique.

image	mots-clés initiaux	extension sans RS	extension avec RS
	bear black water	bear black water grass	bear black water grass
	bear black river	bear black river	bear black river water

TAB. 1 – Exemples d'images et leurs éventuels mots-clés avant et après extension d'annotation, avec et sans prise en compte des relations sémantiques

4 Résultats expérimentaux

Dans cette section, nous présentons une évaluation de notre modèle sur plus de 30000 images provenant de la librairie d'images Corel et fournies par Vasconcelos et al. [4]. Nous avons préféré cette base à celles

Source de la dépendance	destination	type de la relation
autumn	season	direct hypernym
bird	animal	inherited hypernym
beach	sand	substance meronym
chicken	animal	inherited hypernym
chicken	bird	inherited hypernym

TAB. 2 – Exemples de relations sémantiques entre mots-clés et types des relations

utilisées dans les campagnes d'évaluation pour la recherche d'images et l'annotation, comme ImageCLEF et VOC, car elles comportent au plus 20000 images et nous souhaitions travailler sur une plus grande base afin de tester la robustesse à l'échelle et les performances de notre méthode. Ces images sont réparties en 306 classes. Par exemple, la Figure 4 présente quatre images de la classe "arabian horses". La connaissance

FIG. 4 – Exemples d'images de la classe "arabian horses"

d'un mot-clé pour une image ne détermine pas la classe de cette image. En effet, un même mot-clé peut apparaître dans l'annotation d'images de différentes classes. Par exemple, le mots-clés "duck" apparaît dans l'annotation d'images des classes "beautiful Bali", "waterfowl", "african birds" ou "cuisine". 72% des image de la base sont annotées par 4 mots-clés, 23% par 3 mots-clés, 4% par 2 mots-clés et 0.5% par 1 mot-clé (i.e. 99.5% des images de la base sont annotées par au moins 1 mot-clé), en utilisant un vocabulaire de 1036 mots-clés. Par conséquent, dans cette base, les images annotées par moins de 4 mots-clés sont considérées comme partiellement annotées. Les caractéristiques visuelles choisies sont issues d'un descripteur de couleur et d'un descripteur de forme basé sur la transformée de Radon. Le choix des caractéristiques visuelles n'est pas très important dans le sens où l'objectif de ce papier est de montrer que la combinaison d'informations visuelles et sémantiques améliore la reconnaissance, quelles que soient les caractéristiques visuelles. Tout d'abord, des dépendances entre mots-clés ont été éta-

blies, à la main, à partir du vocabulaire. Nous définissons une relation de dépendance entre deux mots-clés du même groupe sémantique (synset) comme défini dans Wordnet [6]. Wordnet est une grande base de données lexicales provenant de la langue anglaise, où les mots (noms, verbes, adjectifs et adverbes) sont regroupés en ensembles de synonymes cognitifs (appelés "synset"). Chaque synset exprime un concept distinct. Ainsi, deux mots-clés ayant une relation sémantique sont regroupés dans le même synset. Ces relations sémantiques sont représentées par des dépendances dans notre modèle, i.e. par des arcs dans notre réseau Bayésien. Certaines de ces dépendances sont données dans le Tableau 2. La première colonne contient les mots-clés qui sont source d'une dépendance avec le mot-clé de la même ligne dans la seconde colonne. La dernière colonne donne le type de la relation sémantique liant les deux mots-clés. Notre méthode a été évaluée en effectuant 6 validations croisées, dont chaque proportion de l'échantillon d'apprentissage est fixée à 25%, 35%, 50%, 65%, 75% et 90% de la base. Les 75%, 65%, 50%, 35%, 25% et 10% respectivement restants sont retenus pour l'échantillon de test. Dans chaque cas, les tests ont été répétés 10 fois, de façon à ce que chaque observation ait été utilisée au moins une fois pour l'apprentissage et les tests. Pour chacune des 6 tailles de l'échantillon d'apprentissage, nous calculons le taux de reconnaissance moyen en effectuant la moyenne des taux de reconnaissance obtenus pour les 10 tests. Pour chaque test, le taux de reconnaissance correspond au ratio entre le nombre d'images bien classées et le nombre d'images test. Dans tous les tests, notre modèle de mélange GM-B a été exécuté avec des mélanges de 2 Gaussiennes et des matrices de covariances diagonales. Considérons le Tableau 3. Notre modèle de mélange de lois de Bernoulli et de mélanges de Gaussiennes a été utilisé pour combiner différents types d'information. La notation "C + F" signifie que les descripteurs de forme et de couleur ("C" pour couleur et "F" pour forme) ont été combinés. La notation "C + F + KW" indique la combinaison des informations visuelles et textuelles. Les taux de reconnaissance confirment que la combinaison des caractéristiques visuelles et sémantiques est toujours plus performante que l'utilisation d'un seul type d'information. Le Tableau 4 montre les taux de reconnaissance de notre modèle GM-B, en prenant en compte les éventuelles relations sémantiques entre mots-clés (colonne "avec RS", RS pour relations sémantiques), ou sans (colonne "sans"). Les résultats montrent que la prise en compte des relations sémantiques améliore le taux de reconnaissance de 10.5% en moyenne. De plus, le Tableau 4 montre l'efficacité de notre approche (modèle GM-B) comparé au modèle de mélange de lois multinomiales et Gaussiennes (noté GM-Mixture) [1]. Le modèle GM-Mixture a été utilisé sans segmentation d'images, comme dans notre approche : les descripteurs de forme et de couleur ont été calculés sur les images entières et les mots-clés sont également associés à une image entière. De plus, comme nous considérons, dans ce papier, un problème de classification supervisée, la variable discrète z, utilisée dans [1] pour représenter la classification jointe d'une image et de sa légende, n'est pas cachée pour les images des échantillons d'apprentissage. De même le nombre de clusters est connu. En fait, cette variable discrète z correspond à notre variable classe "Classe". Les résultats ont été obtenus en utilisant les caractéristiques visuelles de chaque image et leurs éventuels mots-clés. Il apparaît qu'avec les relations sémantiques entre mots-clés, notre modèle GM-B a un meilleur taux de reconnaissance que le modèle GM-Mixture. Considérons maintenant le problème d'extension d'annotations. Au moins un mot-clé par image est nécessaire pour comparer les annotations après extension automatique aux annotations de la vérité terrain. Par conséquent, 99.5% des images de la base, annotées par au moins 1 mot-clé, sont sélectionnées comme vérité terrain. Comme pour l'évaluation de la classification, 60 validations croisées ont été effectuées. Les tests sont répétés 10 fois de façon à ce que chaque image soit utilisée pour l'apprentissage et les tests. Le taux moyen de bonnes annotations est obtenu en effectuant la moyenne des taux de bonnes annotations obtenus pour les 10 tests. Pour chaque test, le taux de bonnes annotations correspond à la proportion de mots-clés corrects parmi les mots-clés obtenus par extension. Pour chaque test, on restreint l'extension d'annotations à 4 mots-clés maximums, car c'est le nombre maximum de mots-clés observés par image dans la vérité-terrain. Le seuil utilisé pour l'annotation a été fixé à 0.5. C'est-à-dire que, pour une image donnée, un mot-clé est sélectionné comme annotation si sa probabilité d'annoter cette image, étant données ses caractéristiques visuelles et ses éventuels mots-clés existants, est strictement supérieure à 0.5. Le Tableau 5 compare les taux de bonnes annotations obtenus avec ou sans prise en compte des relations sémantiques entre mots-clés. On observe que la prise en compte de ces relations améliore le taux de bonnes annotations de 6.9% en moyenne. Le Tableau 5 compare également les taux de bonnes annotations obtenus par le modèle GM-Mixture à ceux obtenus par notre modèle GM-B. On peut voir que notre modèle est meilleur que le modèle GM-Mixture, même sans prendre en compte les relations sémantiques entre mots-clés. Ceci peut s'expliquer par l'utilisation de lois de Bernoulli pour estimer la distribution des mots-clés. En effet, une décision binaire pour chaque mot-clé du vocabulaire (solution proposée dans notre modèle) est plus précis, en annotation, que le choix d'un mot-clé parmi tous les mots-clés du vocabulaire (solution utilisée dans le modèle GM-Mixture).

5 Conclusion et perspectives

Nous avons proposé une méthode de modélisation, classification et annotation d'images partiellement annotées. Cette approche a l'avantage de prendre en compte les relations sémantiques entre les mots-clés constituant les annotations. Les résultats expérimentaux ont démontré que la prise en compte de ces relations sémantiques améliore les taux de reconnaissance et de bonnes annotations. De plus, l'évaluation a montré une amélioration prometteuse des performances par rapport à un classificateur de l'état de l'art. Nos travaux futurs seront dédiés au calcul ou à l'extraction automatique des relations sémantiques à partir de Wordnet. On souhaiterait aussi capturer les préférences de l'utilisateur en intégrant un processus de retour de pertinence. Plus précisément, les préférences de l'utilisateur peuvent être prises en comptes grâce à la modification des paramètres du réseau lors de l'inférence.

part app	C	F	KW	C + F	C + F + KW
25%	20.6	16.5	48.3	23.6	58.5
35%	22.8	16.8	54.5	24	59
50%	23.4	18.4	61.4	24.3	64.2
65%	24.1	19.1	62.4	26	65.6
75%	26	19.9	67.8	26.4	69.8
90%	26	24	69.2	28.8	76

Tab. 3 – Taux de reconnaissance moyens (en %) de notre modèle GM-B sans prise en compte des relations sémantiques

Part apprentissage	GM-Mixture	GM-B	
		sans	avec RS
25%	61	58.5	68.7
35%	62.4	59	69.5
50%	67.2	64.2	76.2
65%	67.7	65.6	75.4
75%	72.2	69.8	80.4
90%	78.6	76	86

Tab. 4 – Taux de reconnaissance moyens (en %) du modèle GM-Mixture vs. notre modèle GM-B

part apprentissage	GM-Mixture	GM-B	
		sans	avec RS
25%	40	52	71
35%	56.2	72.6	78.9
50%	60	72.8	79.6
65%	61.7	77.1	79.7
75%	66	78.9	82.3
90%	68.7	79	82.4

Tab. 5 – Taux moyens (en %) de bonnes annotations du modèle GM-Mixture vs. notre modèle GM-B

Références

[1] D.M. Blei and M.I. Jordan. Modeling annotated data. In *SIGIR '03*, pages 127–134, 2003.

[2] L. Breiman. Random forests. In *Machine Learning*, pages 5–32, 2001.

[3] C.J.C Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2) :121–167, 1998.

[4] G. Carneiro, A.B. Chan, P.J. Moreno, and N. Vasconcelos. Supervised learning of semantic classes for image annotation and retrieval. *IEEE PAMI*, 29(3) :394–410, 2007.

[5] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B*, 39(1) :1–38, 1977.

[6] C. Fellbaum, editor. *WordNet - An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press, 1998.

[7] Y. Jin, L. Khan, L. Wang, and M. Awad. Image annotations by combining multiple evidence & wordnet. In *MULTIMEDIA '05*, 2005.

[8] J. H. Kim and J. Pearl. A computational model for combined causal and diagnostic reasoning in inference systems. In *IJCAI '83*, 1983.

[9] J. Magalhaes and S. Rüger. Information-theoretic semantic multimedia indexing. In *CIVR '07*, pages 619–626, 2007.

[10] O. Ramos-Terrades, E. Valveny, and S. Tabbone. Optimal classifier fusion in a non-bayesian probabilistic framework. *IEEE PAMI*, 31(9) :1630–1644, 2009.

[11] C. Robert. *A decision-Theoretic Motivation*. Springer-Verlag, 1997.

[12] X. Rui, M. Li, Z. Li, W.Y. Ma, and N. Yu. Bipartite graph reinforcement model for web image annotation. In *MULTIMEDIA '07*, pages 585–594, 2007.

[13] A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta, and R.Jain. Content-based image retrieval at the end of the early years. *IEEE PAMI*, 22(12) :1349–1380, December 2000.

[14] A. Torralba and A. Oliva. Statistics of natural image categories. In *Network : Computation in Neural Systems*, pages 391–412, 2003.

[15] L. Wendling, J. Rendek, and P. Matsakis. Selection of suitable set of decision rules using choquet integral. In *SSPR/SPR '08*, pages 947–955, 2008.