Extension automatique d'annotation d'images en utilisant un modèle graphique probabiliste

A. Bouzaieni[1] S. Barrat[2] S. Tabbone[3]

[1] Université de Lorraine-LORIA, Xilopix, UMR 7503, Vandoeuvre-les-Nancy, France
abdessalem.bouzaieni@loria.fr
[2] Université François Rabelais- LI, Tours, France
sabine.barrat@univ-tours.fr
[3] Université de Lorraine-LORIA, UMR 7503, Vandoeuvre-les-Nancy, France
tabbone@loria.fr

Résumé

Avec le développement rapide des caméras numériques et des réseaux sociaux de partage d'images, l'annotation d'images est devenue un domaine de recherche d'un grand intérêt. Elle permet l'indexation et la recherche dans des grandes collections d'images d'une façon plus facile et plus rapide. Dans ce papier, nous proposons un modèle d'extension d'annotation d'images en utilisant un modèle graphique probabiliste. Ce modèle est un mélange de distributions multinomiales et de mélange de Gaussiennes. Les résultats du modèle proposé sont prometteurs sur trois ensembles de données standard : Corel-5k, ESP-Game et IAPRTC-12.

Mots Clef

Annotation d'image, extension d'annotation, modèle graphique probabiliste, mélange de Gaussiennes, caractéristiques visuelles, caractéristiques textuelles.

Abstract

With the fast development of digital cameras and social media image sharing, automatic image annotation has become a research area of great interest. It enables indexing, extracting and searching in large collections of images in an easier and faster way. In this paper, we propose a model for the annotation extension of images using a probabilistic graphical model. This model is based on a mixture of multinomial distributions and mixtures of Gaussians. The results of the proposed model are promising on three standard datasets : Corel-5k, ESP-Game and IAPRTC-12.

Keywords

Image annotation, annotation extension, probabilistic graphical models, Gaussian mixtures, visual characteristics, textual characteristics.

1 Introduction

La croissance rapide des archives de contenus visuels disponibles, par exemple les sites internet de partage de photos ou vidéos, a engendré un besoin en techniques d'indexation et de recherche d'information multimédia, et plus particulièrement en indexation et recherche d'images. L'annotation d'images est l'une des techniques d'indexation sémantique d'images. Nous distinguons trois types d'annotation d'images : l'annotation manuelle, l'annotation automatique et l'annotation semi-automatique. L'annotation manuelle consiste à assigner manuellement un ensemble de mots-clés à une image par un utilisateur. Elle est efficace mais très coûteuse pour un être humain. L'annotation automatique a pour but de générer de nouvelles métadonnées sémantiques pour les images par un système informatique. Ces métadonnées peuvent être utilisées par une requête de recherche d'images. Quand le processus d'annotation automatique d'images nécessite l'intervention d'un utilisateur, nous parlons d'une annotation semi-automatique.

L'annotation automatique d'images [1, 2, 3, 4] est devenue un sujet effectif de recherche depuis des années avec l'apparition de bases standards d'images annotées. Sur ces bases, des techniques d'apprentissage automatique sont utilisées de plus en plus dans les travaux d'annotation d'images. Ces techniques utilisent un processus d'apprentissage supervisé sur un ensemble de données d'images annotées manuellement. L'inconvénient majeur de ces méthodes est qu'elles nécessitent un grand nombre d'exemples annotés pour effectuer l'apprentissage. Un autre inconvénient est que l'annotation est effectuée en une seule fois et ne peut pas être affinée. Dans cet article, nous présentons notre modèle d'extension automatique d'annotation d'images en utilisant un modèle graphique probabiliste. Ce modèle permet de combiner les caractéristiques de bas et haut niveaux pour étendre l'annotation aux images partiellement annotées. Ce modèle ne

nécessite pas que toutes les images de la base d'apprentissage soient annotées. Au contraire, il permet de traiter le problème des données manquantes. Le modèle proposé peut être aussi utilisé pour la tâche de classification des images. L'article est organisé de la manière suivante. Dans la section 2, nous donnons un aperçu des travaux existants. Ensuite, en section 3, nous présentons notre modèle d'extension d'annotation. La section 4 est consacrée aux résultats expérimentaux et des mesures de performance sur trois ensembles de données. Enfin, nos conclusions et les futures directions de recherche sont données en section 5.

2 État de l'art

Le problème d'annotation d'images a été largement étudié durant ces dernières années, et de nombreuses approches ont été proposées pour résoudre ce problème. Ces approches peuvent être regroupées en des modèles génératifs et des modèles discriminatifs. Les modèles génératifs construisent une distribution conjointe des descripteurs et des mots-clés d'une image pour trouver une correspondance entre les descripteurs d'images et les mots-clés d'annotations. Dans le modèle de Duygulu et al. [18], les images sont segmentées en régions, ces dernières sont classifiées en utilisant différents descripteurs. Un apprentissage des correspondances entre les types de régions et les mots-clés fournis avec les images est effectué pour prédire des mots-clés à une nouvelle image. Liu et al. [19] ont proposé le modèle DCMRM (dual cross-media relevance model) qui estime la probabilité conjointe par l'espérance des mots dans un lexique prédéfini comme WordNet [15]. Ce modèle implique deux types de relations dans l'annotation d'images : la relation mot-à-image et la relation mot-à-mot. Ces deux relations peuvent être estimées en utilisant les techniques de recherche des données web ou à partir des données d'apprentissage disponibles. Dans [13], les auteurs ont utilisé deux types de graphes : graphe à base d'images dont les noeuds sont les images et les arcs sont les relations entre les images, et graphe à base de mots où les noeuds sont des mots et les arcs sont des relations entre mots. Le premier graphe est utilisé pour apprendre les relations entre les images et les mots, c'est à dire, pour obtenir les annotations candidates pour chaque image. L'autre graphe est utilisé pour affiner les relations entre les images et les mots pour obtenir les annotations finales pour chaque image. Le modèle SKL-CRM est présenté dans [21]. Ce modèle est une amélioration du modèle CRM [17]. Ce dernier est un modèle statique qui permet d'attribuer des mots-clés à une image en utilisant un ensemble d'apprentissage. Cela se fait par le calcul de $P(w|f)$, où W est un ensemble de mots-clés et f est un vecteur de caractéristiques de l'image annotée. Dans [7], Barrat et al. ont proposé le modèle GM-Mult pour l'extension d'annotation d'images. Dans ce modèle, l'échantillon des caractéristiques visuelles (variables continues) suit une loi dont la fonction de densité est une densité de mélange de Gaussiennes et les variables discrètes (mots-clés) suivent une distribution multinomiale.

Les modèles discriminatifs permettent de transformer le problème de l'annotation en un problème de classification. Plusieurs classificateurs ont été utilisés pour l'annotation comme SVM [20, 28], KNN [12, 24] et les arbres de décision [27, 29]. Lu et al. [23] ont proposé une heuristique appelée HSVM MIL de l'algorithme SVM pour apprendre les correspondances entre les régions d'images et les mots-clés. Fu et al. [27] ont présenté une méthode d'annotation d'images en utilisant les forêts aléatoires. Ils utilisent les annotations contenues dans les images d'apprentissage en tant qu'informations de contrôle afin de guider la génération des arbres aléatoires, permettant ainsi de récupérer les voisins les plus proches, non seulement au niveau visuel, mais aussi au niveau sémantique. Cette méthode considère la forêt aléatoire dans son ensemble et présente deux nouveaux concepts : les plus proches voisins sémantiques (SSN) et la mesure de similarité sémantique (SSM). Dans [24], l'objectif est d'annoter automatiquement des images en utilisant la méthode 2PKNN, variante de la méthode classique KNN, et de proposer un apprentissage métrique des poids et des distances. Pour une image non annotée, ses voisins sémantiques les plus proches correspondant à toutes les étiquettes sont identifiés. Ensuite, les étiquettes correspondant à cette image sont trouvées à partir des échantillons sélectionnés. Guillaumin et al. [26] ont proposé l'algorithme TagProp basé sur la méthode KNN et ont atteint une performance d'annotation très compétitive.

D'autres approches [20, 22] sont une combinaison d'un modèle discriminatif et d'un modèle génératif. Dans [20], les auteurs ont présenté un modèle hybride pour l'annotation des images. Plus précisément, ce modèle est basé sur un SVM utilisé comme modèle discriminatif pour résoudre le problème des images partiellement annotées ou non annotées. Un modèle DMBRM est utilisé comme modèle génératif pour résoudre le problème des données déséquilibrées (échantillons d'apprentissage insuffisants par étiquette).

À partir d'un ensemble d'apprentissage d'images annotées, de nombreuses méthodes d'annotation automatique [7, 13, 18, 19, 5] visent à apprendre des relations entre des caractéristiques visuelles et des concepts sémantiques (mots-clés). Ces relations sont utilisées afin de prédire des mots-clés à des nouvelles images. Le problème de ces méthodes est qu'elles nécessitent un grand nombre d'exemples pour effectuer l'apprentissage. Cependant, dans le monde réel, nous ne disposons pas systématiquement de suffisamment d'images annotées. Plusieurs autres méthodes d'annotation automatique sont basées sur le KNN [12, 24, 26, 27]. Ces méthodes souffrent du problème du fossé sémantique [4, 27]. En effet, les voisins les plus proches récupérés en se basant sur les similarités des caractéristiques visuelles ne partagent pas nécessairement le même concept sémantique. Une solution à ce problème pourrait être de prendre un ensemble d'images partiellement annotées et d'étendre l'annotation à d'autres images. Dans cette perspective, nous proposons dans cet article un modèle d'extension automa-

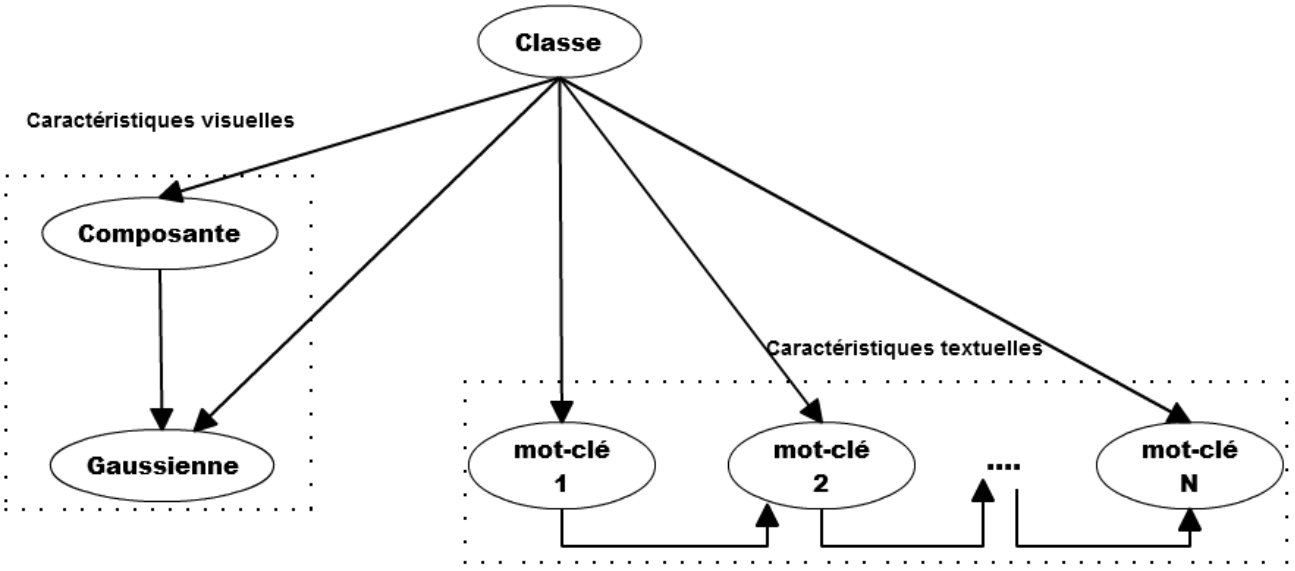

FIGURE 1 – Modèle de mélange de lois multinomiales et de mélange de Gaussiennes

tique d'annotation d'images.

Le travail présenté dans cet article est dans le même esprit que [7]. Cependant, dans notre modèle d'extension d'annotation d'images, les caractéristiques visuelles extraites et la structure des deux modèles sont différentes. Notre modèle a l'avantage d'utiliser des caractéristiques visuelles plus pertinentes et d'ajouter des dépendances conditionnelles dans la structure pour représenter les relations sémantiques entre les mots-clés. En outre, un avantage des modèles graphiques probabilistes est de traiter le problème des données manquantes (mots-clés manquants dans notre cas) et d'offrir la possibilité de combiner plusieurs sources d'informations.

3 Extension d'annotation d'images

Dans cette section, nous détaillons notre méthode d'extension d'annotation d'images en utilisant un modèle graphique probabiliste.

Le modèle proposé est un mélange de distributions multinomiales et de mélange de Gaussiennes. Le modèle proposé est présenté dans la Figure 1. Nous supposons que les caractéristiques visuelles sont considérées comme des variables continues. Elles suivent une loi dont la fonction de densité est une densité de mélange de Gaussiennes. Les caractéristiques textuelles (mots-clés) sont considérées comme des variables discrètes. Elles suivent une distribution multinomiale. Les caractéristiques visuelles d'une image sont représentées par deux noeuds :

— Le noeud *Gaussienne* est modélisé par une variable aléatoire continue qui est utilisée pour représenter les descripteurs calculés sur l'image.

— Le noeud *Composante* est modélisé par une variable aléatoire cachée qui est utilisée pour représenter le poids des Gaussiennes utilisées. Il peut prendre g valeurs correspondant au nombre de Gaussiennes utilisées dans le mélange (ce nombre est déterminé expérimentalement en réalisant le meilleur compromis entre temps d'exécution et précision d'annotation).

Les caractéristiques textuelles (mots-clés) sont modélisées par N noeuds discrets, où N est le nombre maximum de mots-clés utilisés pour annoter une image. Des arcs sont ajoutés entre les N noeuds pour représenter les dépendances conditionnelles entre les mots-clés. Un noeud racine *Classe* est utilisé pour représenter le type d'image, il peut prendre k valeurs correspondant aux classes prédéfinies $C_1, ..., C_k$.

Pour un tel réseau, on peut écrire la probabilité jointe :

$$P(C, TC, LLC) = P(C) \prod_{i=1}^{M} P(LLC_i|C) \prod_{i=1}^{N} P(TC_i|C) \quad (1)$$

où TC représente les caractéristiques textuelles (les mots-clés $Kw_1, ..., Kw_N$) et $LLC_1, ..., LLC_M$ représentent les caractéristiques de bas niveau (caractéristiques visuelles).

Soit $Kw_1, ..., Kw_N$ l'ensemble des mots-clés dans une image. Chaque variable $Kw_j, \forall j \in \{1, ..., N\}$ peut être représentée par un espace vectoriel booléen des mots du vocabulaire :

$Kw_j = \{m_1, ..., m_n\}$, où $m_i = 0$ ou $1, \forall i \in \{1, ..., n\}$ et $\sum_{i=1}^{n} m_i = k$.

Chaque variable $Kw_j, \forall j \in \{1, ..., N\}$ suit une distribution multinomiale avec les paramètres $\Phi_{TC} = (k, p_1, ..., p_n)$, où p_i est la probabilité associée à chaque valeur m :

$$p(m_1 = p_1, ..., m_n = p_n) = \frac{k!}{m_1! m_2! ... m_n!} p_1^{m_1} p_2^{m_2} ... p_n^{m_n} \quad (2)$$

TABLE 1 – Détail des trois bases de données (Corel-5k, ESP-Game et IAPRTC-12)

Base	nombre d'images	taille du vocabulaire	taille d'apprentissage	taille de test	mots par image	images par mot
Corel-5K	5 000	260	4 500	500	3.4	58.6
ESP-Game	20 770	268	18 689	2 081	4.7	362.7
IAPRTC-12	19 627	291	17 665	1 962	5.7	347.7

Soit I un ensemble de m images $(im_1, ..., im_m)$ et g groupes $(G_1, ..., G_g)$ dont chacun a une densité Gaussienne avec une moyenne $\mu_l, \forall l \in \{1, ..., g\}$ et une matrice de covariance $\sum_l$.
Soit $\pi_1, ..., \pi_g$ les proportions des différents groupes, on note par $\theta_k = (\mu_k, \sum_k)$ le paramètre de chaque Gaussienne, et $\Phi_{LLC} = (\pi_1, ..., \pi_g, \theta_1, ..., \theta_g)$ le paramètre global du mélange. Alors, la densité de probabilité de I conditionnellement à la classe $c_i, \forall i \in \{1, ..., k\}$ est définie par :

$$P(im, \Phi_{LLC}) = \sum_{l=1}^{g} \pi_l p(im, \theta_l) \tag{3}$$

où $p(im, \theta_l)$ est la Gaussienne multivariée définie par le paramètre θ_l.
On note par Φ le paramètre global de ce modèle :

$$\Phi = (\Phi_{LLC}, \Phi_{TC}) = (\pi_1, ..., \pi_g, \theta_1, ..., \theta_g, k, p_1, ..., p_n) \tag{4}$$

L'équation (1) peut être réécrite :

$$P(C, TC, LLC) = P(C) f(\pi_1, .., \pi_g, \theta_1, .., \theta_g, k, p_1, .., p_n) \tag{5}$$

Les paramètres du modèle peuvent être appris d'un ensemble d'images d'apprentissage pour estimer la probabilité jointe de chaque image et chaque classe. Nous devons maximiser la log-vraisemblance L_I de I :

$$\begin{aligned} L_I &= log(P(C, TC, LLC) \\ &= \sum logP(C) + \sum logf(\pi_1, ..., \pi_g, \theta_1, ..., \theta_g) \\ &+ \sum logf(k, p_1, ..., p_n) \end{aligned} \tag{6}$$

Étant donné que l'ensemble des images d'apprentissage est incomplet, nous utilisons l'algorithme EM [35, 34]. Cet algorithme est le plus utilisé dans le cas de données manquantes. Des *a priori* de Dirichlet ont été utilisés avec les variables mots-clés du modèle. Pour une image partiellement annotée, représentée par ses caractéristiques visuelles $LLC_1, LLC_2, ..., LLC_M$ et ses mots-clés existants $Kw_1, Kw_2, ..., Kw_n$, nous pouvons utiliser l'algorithme d'inférence [36] pour étendre l'annotation de cette image avec d'autres mots-clés. Nous pouvons calculer la probabilité *a posteriori*

$$P(Kw_i | LLC_1, ..., LLC_M, Kw_1, ..., Kw_n) \, \forall i \in \{1, ..., N\} \tag{7}$$

où N est la taille du vocabulaire utilisé. Le mot-clé ayant la probabilité maximale sera retenu comme une nouvelle annotation de l'image. Nous pouvons également calculer la probabilité *a posteriori*

$$P(C_i | LLC_1, ..., LLC_M, Kw_1, ..., Kw_n) \tag{8}$$

dans le but d'identifier la classe de l'image. L'image requête est affectée à la classe C_i maximisant cette probabilité.

4 Expérimentation

Dans cette section, nous présentons d'abord les bases d'images utilisées dans nos expériences et les différents critères d'évaluation des performances. Ensuite, nous présentons les résultats expérimentaux obtenus avec notre modèle.

4.1 Bases de données et critères d'évaluation

Nous avons effectué nos expériences sur les trois bases d'images les plus populaires en annotation d'images : Corel-5K, ESP-Game et IAPRTC-12.

- **Corel-5K :** Cette base est la plus utilisée pour l'annotation et la recherche d'images. Elle est divisée en 4500 images pour l'apprentissage et 500 images pour les tests avec un vocabulaire de 260 mots-clés. Les images sont regroupées en 50 catégories, chacune d'elles contient 100 images. Chaque image est annotée manuellement avec 1 à 5 mots-clés avec une moyenne de 3.4 mots-clés par image.
- **ESP-Game :** Cette base est obtenue à partir d'un jeu en ligne. Nous utilisons un sous-ensemble de 20770 images utilisées dans [20, 24, 26, 8]. Ce sous-ensemble est divisé en 18689 images pour l'apprentissage et 2081 images pour les tests avec un vocabulaire de 268 mots-clés. Chaque image est annotée avec une moyenne de 4.7 mots-clés par image.
- **IAPRTC-12 :** Cette base est une collection d'environ 20000 images naturelles. Elle est divisée en 17665 images pour l'apprentissage et 1962 images pour les tests avec un vocabulaire de 291 mots-clés. Chaque image est annotée avec une moyenne de 5.7 mots-clés par image.

Pour évaluer notre modèle d'annotation d'images, nous utilisons les quatre mesures d'évaluation standards utilisées dans l'annotation d'images. Nous annotons automatiquement par notre modèle chaque image dans la base de

		
sky, sun, clouds, tree	sky, jet, plane	bear, polar, snow, tundra
sky, **sun**, **clouds**, **tree**, **palm**	**sky**, **jet**, **plane**, **f-16**, **kit**	**bear**, **polar**, **snow**, **ice**, **slope**
		
water, boats, bridge	tree, horses, mare, foals	sky, buildings, flag
water, **boats**, **bridge**, **arch**, **pyramid**	**tree**, **horses**, **mare**, **foals**, **field**	**sky**, **buildings**, **skyline**, **street**, **outside**
		
flowers, house, garden	plants	mountain, sky, water, tree
flowers, **house**, **garden**, **lawn**, **reptile**	**plants**, **leaf**, **stems**, **interior**, **zebra**	**mountain**, **sky**, **clouds**, **tree**, **whales**

FIGURE 2 – Exemples d'annotation d'images de la base Corel-5k

TABLE 2 – Performance de notre modèle par rapport aux autres modèles d'annotation de l'état de l'art sur trois ensembles de données (Corel-5k, ESP-Game et IAPRTC-12)

Méthode	Corel-5K				ESP-Game				IAPRTC-12			
	P	R	$F1$	$N+$	P	R	$F1$	$N+$	P	R	$F1$	$N+$
CRM [17]	16	19	17	107	–	–	–	–	–	–	–	–
MBRM [16]	24	25	25	122	18	19	18	209	24	23	23	223
SML [11]	23	29	26	137	–	–	–	–	–	–	–	–
JEC [25]	27	32	29	139	22	25	23	224	28	29	28	250
GS [10]	30	33	31	146	–	–	–	–	32	29	30	252
RF [27]	29	40	34	157	41	26	32	235	44	31	36	253
TagProp [26]	33	42	37	160	39	27	32	239	46	35	40	266
GMM-Mult [7]	19	31	24	104	29	26	27	224	28	22	25	227
NMF-KNN [12]	38	**56**	**45**	150	33	26	29	238	–	–	–	–
SVM-DMBRM [20]	36	48	41	**197**	**55**	25	34	259	**56**	29	38	**283**
SKL-CRM [21]	39	46	42	184	41	26	32	248	47	32	38	274
2PKNN [24]	**44**	46	**45**	191	53	27	**36**	252	54	**37**	**44**	278
Notre méthode	27	42	33	165	30	**35**	32	248	32	35	33	256

test par 5 mots-clés et nous calculons le rappel, la précision, $F1$ et $N+$. Supposons qu'une étiquette soit présente m_1 fois dans les images de la vérité terrain, et apparaisse dans m_2 images lors des tests à partir desquels m_3 prédictions sont correctes. La précision (P) est le rapport entre les images correctement annotées par un mot-clé et toutes les images annotées par ce mot-clé par le modèle : $P = m_3/m_2$. Le rappel (R) est le rapport entre les images correctement annotées par un mot-clé et toutes les images annotées par ce mot-clé dans les images de vérité terrain : $R = m_3/m_1$. $N+$ est le nombre de mots qui sont correctement affectés à au moins une image de test (nombre de mots avec rappel strictement positif). La mesure F_1 est une moyenne harmonique entre le rappel et la précision : $F_1 = 2(PR)/(P + R)$.

4.2 Caractéristiques visuelles

Nous avons utilisé des descripteurs globaux et locaux. Le descripteur GIST a été proposé par Oliva et Torralba [30]. Il capture la forme globale d'une image en caractérisant l'orientation des différents contours qui y apparaissent. La méthode SIFT [31] permet de détecter et identifier des éléments similaires entre des images (paysages, objets, personnes,..). Les descripteurs SIFT sont des informations numériques qui dérivent de l'analyse locale d'une image, et qui caractérisent le contenu visuel de cette image indépendamment de l'échelle, de l'angle d'observation et de la luminosité. Ils sont généralement calculés autour des points d'intérêts. Le descripteur SURF [32] est inspiré du descripteur SIFT, il est plus rapide et plus robuste pour différentes transformations d'images. Il est basé sur la détermination de l'angle de rotation de la zone d'analyse avec la construction d'un histogramme de gradients orientés. Le descripteur de texture LBP [33] compare le niveau de luminance d'un pixel avec les niveaux de ses voisins. Grâce à son pouvoir discriminant et la simplicité de calcul, LBP est devenu populaire dans diverses applications.

4.3 Résultats

La figure 2 illustre l'annotation de quelques images de la base Corel-5k où les étiquettes de la vérité terrain sont données. Les mots-clés en gras ont été trouvés automatiquement par notre modèle d'annotation. Les mots-clés en gras et soulignés ont été ajoutés automatiquement par notre extension d'annotation. Par exemple, la quatrième image est annotée manuellement par trois mots-clés, deux nouveaux mots-clés "arch" et "pyramid" ont été ajoutés automatiquement après l'extension automatique.

Les résultats de notre modèle par rapport à d'autres approches sur trois ensembles de données Corel-5k, ESP-Game et IAPRTC-12 sont présentés dans le tableau 2. Dans ce tableau, P représente la précision moyenne, R le rappel moyen, F1 la moyenne harmonique entre le rappel et la précision, et N+ le nombre de mots-clés qui ont été correctement assignés aux images de test.

La méthode proposée fournit des résultats concurrentiels pour les deux critères R et N+ comparés aux autres méthodes. Cela signifie que notre approche ne pénalise pas les étiquettes qui ne sont présentes que dans quelques images d'apprentissage. Il fournit le meilleur rappel sur la base ESP-Game.

Nous pouvons remarquer dans le tableau 2 que la méthode [24] présente des meilleurs résultats que la nôtre sur les trois bases. Cependant, cette méthode présente l'inconvénient du grand temps d'annotation. En effet, chaque image à annoter doit être comparée à toutes les images de la base. Au contraire, pour notre méthode, l'apprentissage est effectué une fois pour toutes et, pour annoter une image, nous calculons seulement la probabilité présentée dans l'équation (7). En outre, comme toutes les méthodes basées sur KNN, cette méthode souffre du problème du choix du nombre de voisins et la distance à utiliser entre les caractéristiques visuelles. La méthode [20] présente des bons résultats sur les trois bases. Cependant, cette méthode est une combinaison d'un modèle discriminatif et d'un modèle génératif. C'est un modèle de complexité élevée. De plus, comme toutes les méthodes à base de SVM, cette méthode est mal adaptée aux problèmes avec données manquantes et est perturbée avec un grand nombre de classes.

Par rapport aux méthodes citées dans le tableau 2, notre méthode présente l'avantage d'être utilisée pour les deux tâches d'annotation et de classification des images. Notre méthode peut être utilisée aussi pour l'extension d'annotation en combinant des caractéristiques visuelles et textuelles. Par rapport au modèle GMM-mult [7], les résultats montrent une amélioration significative pour tous les critères d'évaluation et pour les trois ensembles de données. Cela signifie que l'addition des relations sémantiques entre les mots-clés améliore les résultats de l'annotation.

Pour rendre les résultats plus stables, nous avons évalué notre méthode en effectuant quatre validations croisées (tableau 3). Chaque proportion de l'échantillon d'apprentissage est fixée à 80%, 70%, 60% et 50% de l'ensemble de données. Les 20%, 30%, 40% et 50% restant respectivement sont sélectionnés pour l'échantillon de test. Dans chaque cas, les essais ont été répétés cinq fois. Le résultat final est une moyenne de ces répétitions. Nous pouvons remarquer dans le tableau 3 que les résultats restent acceptables si les échantillons d'apprentissage diminuent. Cela montre que notre approche est efficace même en présence de quelques échantillons d'apprentissage.

Pour améliorer la précision de la méthode proposée, nous pouvons définir un seuil λ sur la probabilité d'un mot-clé. Une image sera annotée par un mot-clé Kw seulement si :

$$P(KW|im_1, im_2, ..., im_m, Kw_1, Kw_2, ..., Kw_n) > \lambda \quad (9)$$

Le tableau 4 montre la précision de notre modèle avec différents seuils sur les trois ensembles de données. Nous notons clairement que l'utilisation d'un seuil améliore considérablement la précision du modèle. L'utilisation du seuil pénalise les annotations qui ont une probabilité inférieure à λ. Par exemple, pour ESP-Game, un seuil de 0,1 supprime en moyenne un mot-clé par image.

TABLE 3 – Performance de notre modèle avec des validations croisées sur trois ensembles de données (Corel-5k, ESP-Game et IAPRTC-12)

Test	Apprentissage	Corel-5K				ESP-Game				IAPRTC-12			
(%)	(%)	P	R	$F1$	$N+$	P	R	$F1$	$N+$	P	R	$F1$	$N+$
20	80	22	39	28	175	29	35	32	257	29	32	30	269
30	70	21	37	27	185	28	34	31	263	28	31	29	278
40	60	19	35	25	187	27	33	30	261	27	31	29	281
50	50	19	34	24	194	26	33	29	264	25	30	27	283

TABLE 4 – Précision de l'annotation avec différents seuils sur trois ensembles de données (Corel-5k, ESP-Game et IAPRTC-12)

Base	Corel-5K			ESP-Game			IAPRTC-12		
Seuil (λ)	0.1	0.2	0.3	0.1	0.2	0.3	0.1	0.2	0.3
Précision	37	37	37	38	41	50	36	40	44
Mots-clés par image	3.43	3.28	3.0	4.07	3.67	3.21	4.43	3.87	2.75

5 Conclusion et perspectives

Nous avons proposé un modèle graphique probabiliste pour l'extension d'annotation des images. Ce modèle est un mélange de distributions multinomiales et de mélange de Gaussiennes où nous avons combiné des caractéristiques visuelles et textuelles. Les résultats expérimentaux sur Corel-5k, ESP-Game et IAPRTC-12 ont démontré que la considération des relations sémantiques entre les mots-clés améliore significativement les performances d'annotation. Nos futurs travaux seront consacrés à utiliser les hiérarchies sémantiques pour enrichir l'annotation des images.

Remerciments

Ce travail est réalisé dans le cadre d'un contrat CIFRE avec la société Xilopix d'Épinal.

Références

[1] Wang, F., A survey on automatic image annotation and trends of the new age. Procedia Engineering 23, pp. 434-438, 2011.

[2] Hanbury, A., A survey of methods for image annotation. Journal of Visual Languages & Computing 19(5), pp. 617-627, 2008.

[3] Zhang, D., Islam, M. M., Lu, G., A review on automatic image annotation techniques. Pattern Recognition 45(1), pp. 346-362, 2012.

[4] Tousch, A. M., Herbin, S., Audibert, J. Y., Semantic hierarchies for image annotation : A survey. Pattern Recognition 45(1), pp. 333-345, 2012.

[5] Bouzaieni, A., Barrat, S., Tabbone, S., Automatic annotation extension and classification of documents using a probabilistic graphical model, International Conference on Document Analysis and Recognition, pp. 316-320, 2015.

[6] Wang, C., Blei, D., Li, F. F., Simultaneous image classification and annotation. Computer Vision and Pattern Recognition, pp. 1903-1910, 2009.

[7] Barrat, S., Tabbone, S., Classification and Automatic Annotation Extension of Images Using Bayesian Network. SSPR/SPR, pp. 937-946, 2008.

[8] Bouzaieni, A., Tabbone, S., Barrat, S., Automatic Images Annotation Extension Using a Probabilistic Graphical Model. In Computer Analysis of Images and Patterns, pp. 579-590, 2015.

[9] Wang, C., Yan, S., Zhang, L., Zhang, H. J., Multi-label sparse coding for automatic image annotation. CVPR, 2009.

[10] Zhang, S., Huang, J., Huang, Y., Yu, Y., Li, H., Metaxas, D., Automatic Image Annotation Using Group Sparsity. CVPR, 2010.

[11] Carneiro, G., Chan, A., Moreno, P., Vasconcelos, N., Supervised learning of semantic classes for image annotation and retrieval. PAMI 29(3), pp. 394-410, 2007.

[12] Kalayeh, M., Idrees, H., Shah, M., NMF-KNN : Image Annotation using Weighted Multi-view Non-Negative Matrix Factorization. CVPR, 2014.

[13] Liu, J., Li, M., Liu, Q., Lu, H., Ma, S., Image annotation via graph learning. Pattern Recognition 42(2), pp. 218-228, 2009.

[14] Zhang, S., Tian, Q., Hua, G., Huang, Q., Gao, W., ObjectPatchNet : Towards scalable and semantic image annotation and retrieval. Computer Vision and Image Understanding 118, pp. 16-29, 2014.

[15] Miller, G., WordNet : A Lexical Database for English. Communications of the ACM, 1995.

[16] Feng, S., Manmatha, R., Lavrenko, V., Multiple bernoulli relevance models for image and video annotation. Conference on Computer Vision and Pattern Recognition, pp. 1002-1009, 2004.

[17] Lavrenko, V., Manmatha, R., Jeon, J., A model for learning the semantics of pictures. NIPS, pp. 553-560, 2004.

[18] Duygulu, P., Barnard, K., de Freitas, J., Forsyth, D. A., Object recognition as machine translation : Learning a lexicon for a fixed image vocabulary. Proceedings of 7th Europe Conference on Computer Vision, pp. 97-112, 2002.

[19] Liu, J., Wang, B., Li, M., Li, Z., Ma, W., Lu, H., Ma, S., Dual cross-media relevance model for image annotation. Proceedings of the 15th international conference on Multimedia, pp. 605-614, 2007.

[20] Murthy, V. N., Can, E. F., Manmatha, R., A Hybrid Model for Automatic Image Annotation. International Conference on Multimedia Retrieval, pp. 369-376, 2014.

[21] Moran, S., Lavrenko, V., Sparse kernel learning for image annotation. International Conference on Multimedia Retrieval, pp. 113-120, 2014.

[22] Wang, M., Xia, X., Le, J., Zhou, X., Effective automatic image annotation via integrated discriminative and generative models. Inf. Sci. 262, pp. 159-171, 2014.

[23] Jing, L., Shaoping, M., Region-Based Image Annotation Using Heuristic Support Vector Machine in Multiple-Instance Learning. Journal of Computer Research and Development 46(5), pp. 864-871, 2009.

[24] Verma, Y., Jawahar, C. V., Image Annotation Using Metric Learning in Semantic Neighbourhoods. Computer Vision ECCV, LNCS 7574, pp. 836-849, 2012.

[25] Makadia, A., Pavlovic, V., Kumar, S., Baselines for Image Annotation. ECCV 90, 2008.

[26] Guillaumin, M., Mensink, T., Verbeek, J., Schmid, C., TagProp : Discriminative metric learning in nearest neighbor models for image auto-annotation. ICCV, 2009.

[27] Fu, H., Zhang, Q., Qiu, G., Random Forest for Image Annotation. Computer Vision ECCV, LNCS Volume 7577, pp. 86-99, 2012.

[28] Alham, N. K., Li, M., Liu, Y., Qi, M., A MapReduce based distributed SVM ensemble for scalable image classification and annotation. Computers & Mathematics with Applications, 66(10), pp. 1920-1934, 2013.

[29] Fakhari, A., Moghadam, A., Combination of classification and regression in decision tree for multi-labeling image annotation and retrieval. Appl. Soft Comput 13(2), pp. 1292-1302, 2013.

[30] Oliva, A., Torralba, A., Modeling the shape of the scene : A holistic representation of the spatial envelope. IJCV 42(3), 2001.

[31] Lowe, G. D., Object recognition from local scale-invariant features. Proceedings of the International Conference of Computer Vision 2, pp. 1150-1157, 1999.

[32] Bay, H., Tuytelaars, T., Gool, L. V., Surf : speeded up robust features. Computer Vision and Image Understanding 110(3), pp. 346-359, 2008.

[33] Ojala, T., Pietikinen, M., Harwood, D., A comparative study of texture measures a with classification based on feature distributions. Pattern Recognition 29, pp. 51-59, 1996.

[34] Mahjoub, M. A., Bouzaiene, A., Ghanmy, N., Tutorial and selected approaches on parameter learning in Bayesian network with incomplete data. In Advances in Neural Networks-ISNN, pp. 478-488, 2012

[35] Dempster A. P., Laird N. M., Rubin D. B., Maximum likelihood from incomplete data via the EM algorithm, Journal of the Royal Statistical Society, vol. 39, no 1, pp. 1-38, 1977.

[36] Kim J. H., Pearl J., A computational model for combined causal and diagnostic reasoning in inference systems, International Joint Conference on Artificial Intelligence, pp. 190-193, 1983.