Construction de hiérarchies sémantiques pour l'annotation d'images

Hichem Bannour, Céline Hudelot

Construction de hiérarchies sémantiques pour l'annotation d'images

Hichem Bannour, Céline Hudelot

École Centrale Paris
Laboratoire de Mathématiques Appliquées aux Systèmes (MAS)
Grande voie des vignes, 92295 Châtenay-Malabry, France
{Hichem.bannour,Celine.hudelot}@ecp.fr

RÉSUMÉ. *Cet article propose une nouvelle méthode pour la construction automatique de hiérarchies sémantiques adaptées à la classification et à l'annotation d'images. La construction de la hiérarchie est basée sur une nouvelle mesure de similarité sémantique qui intègre plusieurs sources d'informations : visuelle, conceptuelle et contextuelle que nous définissons dans cet article. L'objectif est de fournir une mesure qui est plus proche de la sémantique des images. Nous proposons ensuite des règles, basées sur cette mesure, pour la construction de la hiérarchie finale qui encode explicitement les relations hiérarchiques entre les différents concepts. La hiérarchie construite est ensuite utilisée dans un cadre de classification sémantique hiérarchique d'images en concepts visuels. Nos expériences et résultats montrent que la hiérarchie construite permet d'améliorer considérablement les résultats de la classification.*

ABSTRACT. *This paper proposes a new methodology to automatically build semantic hierarchies suitable for image annotation and classification. The building of the hierarchy is based on a new measure of semantic similarity. The proposed measure incorporates several sources of information : visual, conceptual and contextual as we defined in this paper. The aim is to provide a measure that best represents image semantics. We then propose rules based on this measure, for the building of the final hierarchy, and which explicitly encode hierarchical relationships between different concepts. Therefore, the built hierarchy is used in a semantic hierarchical classification framework for image annotation. Our experiments and results show that the built hierarchy improves significantly the classification accuracy.*

MOTS-CLÉS : *construction de hiérarchies sémantiques, sémantique d'images, annotation d'images, mesures de similarité sémantiques, classification hiérarchique d'images.*

KEYWORDS: *semantic hierarchies building, image semantics, image annotation, semantic relatedness measure, hierarchical image classification.*

DOI:10.3166/RIA..1-27

1. Introduction

Avec l'explosion des données images, il devient essentiel de fournir une annotation sémantique de haut niveau à ces images pour satisfaire les attentes des utilisateurs dans un contexte de recherche d'information. Des outils efficaces doivent donc être mis en place pour permettre une description sémantique précise des images. Depuis les dix dernières années, plusieurs approches d'annotation automatique d'images ont été proposées (Barnard *et al.*, 2003 ; Lavrenko *et al.*, 2003 ; Fan *et al.*, 2007 ; 2008 ; Bannour, 2009) pour essayer de réduire le problème bien connu du *fossé sémantique* (Smeulders *et al.*, 2000). Cependant, dans la plupart de ces approches, la sémantique est souvent limitée à sa manifestation perceptuelle, *i.e.* à travers l'apprentissage d'une fonction de correspondance associant les caractéristiques de bas niveau à des concepts visuels de plus haut niveau sémantique (Barnard *et al.*, 2003 ; Lavrenko *et al.*, 2003 ; Carneiro *et al.*, 2007 ; Romdhane *et al.*, 2010). Cependant, malgré une efficacité relative concernant la description du contenu visuel d'une image, ces approches sont incapables de décrire la sémantique d'une image comme le ferait un annotateur humain. Elles sont également confrontées au problème du passage à l'échelle (Liu *et al.*, 2007 ; Deng *et al.*, 2010). En effet, les performances de ces approches varient considérablement en fonction du nombre de concepts et de la nature des données ciblées (Hauptmann *et al.*, 2007). Cette variabilité peut être expliquée d'une part par la large variabilité visuelle intra-concept, et d'autre part, par une grande similarité visuelle inter-concept, qui conduisent souvent à des annotations imparfaites.

Récemment, plusieurs travaux se sont intéressés à l'utilisation de hiérarchies sémantiques pour surmonter ces problèmes. Un état de l'art concernant ce type de travaux est proposé dans (Bannour, Hudelot, 2011 ; Tousch *et al.*, 2012 ; Bannour, Hudelot, 2012b). En effet, l'utilisation de connaissances explicites, telles que les hiérarchies sémantiques, peut améliorer l'annotation en fournissant un cadre formel qui permet d'argumenter sur la cohérence des informations extraites des images. En particulier, les hiérarchies sémantiques se sont avérées être très utiles pour réduire le fossé sémantique (Deng *et al.*, 2010). Trois types de hiérarchies pour l'annotation et la classification d'images ont été récemment explorées : 1) les hiérarchies basées sur des connaissances textuelles (nous ferons référence à ce type de connaissances par information conceptuelle dans le reste du papier) [1] (Marszalek, Schmid, 2007 ; Wei, Ngo, 2007 ; Deng *et al.*, 2009), 2) les hiérarchies basées sur des informations visuelles (ou perceptuelles), *i.e.* caractéristiques de bas niveau de l'image (Sivic *et al.*, 2008 ; Bart *et al.*, 2008 ; Yao *et al.*, 2009), 3) les hiérarchies que nous nommerons sémantiques basées à la fois sur des informations textuelles et visuelles (L.-J. Li *et al.*, 2010 ; Fan *et al.*, 2007 ; Wu *et al.*, 2008 ; Bannour, Hudelot, 2012a). Les deux premières catégories d'approches ont montré un succès limité dans leur usage. En effet, d'un côté l'information conceptuelle seule n'est pas toujours en phase avec la sémantique de l'image, et est alors insuffisante pour construire une hiérarchie adéquate pour l'anno-

1. Exemple d'information textuelle utilisée pour la construction des hiérarchies : les tags, contexte environnant, WordNet, Wikipedia, etc.

tation d'images (Wu *et al.*, 2008). De l'autre côté, l'information perceptuelle ne suffit pas non plus à elle seule pour la construction d'une hiérarchie sémantique adéquate (cf. figure 1). En effet, il est difficile d'interpréter ces hiérarchies dans des niveaux d'abstraction plus élevés. Ainsi, la combinaison de ces deux sources d'information semble donc obligatoire pour construire des hiérarchies sémantiques adaptées à l'annotation d'images.

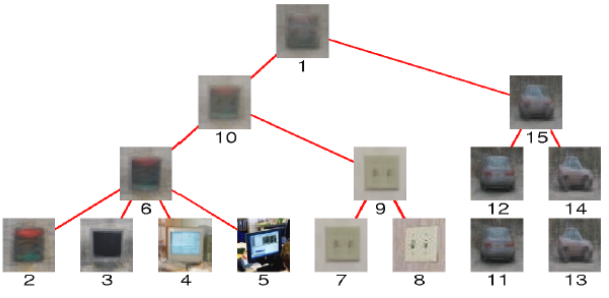


Figure 1. Méthode proposée par (Sivic et al., 2008) pour la construction automatique de hiérarchies de classes. L'exemple illustré est construit sur un sous-ensemble d'images de la base LabelMe (Russell et al., 2008). Les classes générées ne peuvent pas être interprétées dans le niveau sémantique

Dans cet article, nous proposons tout d'abord une nouvelle approche pour la construction automatique de hiérarchies sémantiques adaptées à l'annotation d'images. Notre approche de construction combine les différentes modalités d'une collection d'images pour modéliser au mieux sa sémantique, et permettre de se rapprocher de la perception/représentation que peut avoir un utilisateur face à ces données. Nous proposons par la suite une approche originale, basée sur la hiérarchie construite, pour un apprentissage efficace des classifieurs hiérarchiques. En effet, notre approche d'apprentissage s'appuie sur la structure de la hiérarchie pour décomposer ce problème en plusieurs tâches complémentaires et indépendantes, permettant ainsi le passage à l'échelle de l'approche proposée. Par conséquent, nous proposons deux nouvelles méthodes pour le calcul d'une fonction de décision servant à la classification hiérarchique des images. La première adopte une démarche ascendante, et est calculée par la fusion des scores des classifieurs hiérarchiques afin d'aboutir à la meilleure fonction de décision. La seconde est une approche descendante, et est basée sur les votes des classifieurs pour parcourir la hiérarchie. Les expérimentations faites sur les données du challenge Pascal VOC'2010 ont montré une amélioration considérable de la précision moyenne en utilisant notre méthode de construction de hiérarchies, ainsi que les techniques proposées pour la classification hiérarchique des images.

La suite de cet article est organisée comme suit : dans la section 2 nous présentons les travaux connexes. La section 3 présente la mesure sémantique proposée dans un premier temps, puis les règles utilisées pour la construction de la hiérarchie sémantique. La section 4 présente l'approche proposée pour la classification hiérarchique

des images. Les résultats expérimentaux sont présentés dans la section 5. La section 6 présente nos conclusions et perspectives.

2. État de l'art

Plusieurs méthodes ont été proposées pour la construction de hiérarchies de concepts dédiées à l'annotation d'images (L.-J. Li *et al.*, 2010 ; Fan *et al.*, 2007 ; Marszalek, Schmid, 2007 ; Wei, Ngo, 2007 ; Sivic *et al.*, 2008 ; Bart *et al.*, 2008). Dans cette section nous présentons ces différentes méthodes en suivant la classification en trois types proposée dans l'introduction.

Marszalek et Schmid (2007) ont proposé de construire une hiérarchie de concepts par l'extraction du graphe pertinent dans WordNet, reliant entre eux l'ensemble des concepts formant le vocabulaire d'annotation. La structure de cette hiérarchie est ensuite utilisée pour construire un ensemble de classifieurs hiérarchiques. Deng *et al.* (2009) ont proposé *ImageNet* [2], une ontologie à grande échelle pour les images qui repose sur la structure de WordNet, et qui vise à peupler les 80 000 synsets de WordNet avec une moyenne de 500 à 1 000 images sélectionnées manuellement. L'ontologie LSCOM (Naphade *et al.*, 2006) vise à concevoir une taxonomie avec une couverture de près de 1 000 concepts pour la recherche de vidéos dans les bases de journaux télévisés. Une méthode pour la construction d'un espace sémantique enrichi par les ontologies est proposée dans (Wei, Ngo, 2007). Bien que ces hiérarchies soient utiles pour fournir une structuration compréhensible des concepts, elles ignorent l'information visuelle qui est une partie importante du contenu des images. Aussi, les hiérarchies de ce type sont généralement profondes, et contiennent beaucoup de concepts intermédiaires qui ne sont pas forcément pertinents dans le domaine de l'image. Il est, par exemple, difficile de différencier, au niveau visuel, les photos des classes suivantes : animaux ruminants vs ceux qui ne le sont pas, ou (animaux) carnivores vs herbivores. De plus, comme nous allons le montrer dans nos expérimentations, la profondeur de la hiérarchie a un impact important sur les performances des classifieurs. La figure 2 illustre une hiérarchie de concepts que nous avons construit par l'extraction du graphe dans WordNet reliant l'ensemble des 20 concepts de la base Pascal VOC'2010. La construction de la hiérarchie est réalisée en reliant d'abord les concepts qui sont les plus similaires en termes de plus court chemin dans WordNet, puis en prenant tous leurs hypernymes jusqu'à atteindre la racine de WordNet. Une étape de désambiguïsation est notamment réalisée pour aboutir au bon sens (synset) qui correspond à chaque mot.

D'autres travaux se sont basés sur l'information visuelle pour la construction de hiérarchies (Griffin, Perona, 2008 ; Sivic *et al.*, 2008 ; Bart *et al.*, 2008 ; Marszalek, Schmid, 2008 ; Yao *et al.*, 2009 ; Gao, Koller, 2011). Une plateforme (I2T) dédiée à la génération automatique de descriptions textuelles pour les images et les vidéos est proposée dans (Yao *et al.*, 2009). I2T est basée principalement sur un graphe AND-

2. http://www.image-net.org

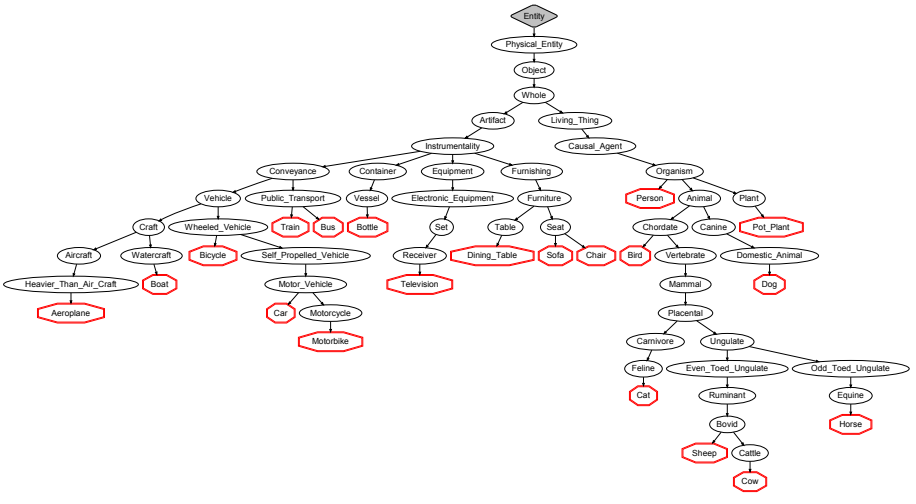


Figure 2. Hiérarchie de concepts construite par l'extraction du graphe pertinent dans WordNet reliant les 20 concepts de la base VOC'2010. Les nœuds en double octogone sont les concepts de départ

OR pour la représentation des connaissances visuelles. Sivic *et al.* (2008) ont proposé de regrouper les objets dans une hiérarchie (visuelle) en fonction de leurs similarités visuelles. Le regroupement est obtenu en adaptant, pour le domaine de l'image, le modèle d'Allocation Dirichlet Latente hiérarchique (hLDA) (Blei *et al.*, 2004). Bart *et al.* (2008) ont proposé une méthode bayésienne pour organiser une collection d'images dans une arborescence en forme d'arbre hiérarchique. Dans (Griffin, Perona, 2008), une méthode pour construire automatiquement une taxonomie pour la classification d'images est proposée. Les auteurs suggèrent d'utiliser cette taxonomie afin d'augmenter la rapidité de la classification au lieu d'utiliser un classifieur multiclasse sur toutes les catégories. Cependant, ces hiérarchies visuelles présentent un handicap majeur qui est le manque d'expressivité au niveau sémantique, *i.e.* l'incapacité de les exploiter, ni de les interpréter au niveau sémantique. Ainsi, une bonne alternative pour la construction d'une hiérarchie sémantique compréhensible et adéquate pour l'annotation d'images serait de tenir compte à la fois de l'information conceptuelle et de l'information visuelle lors du processus de construction.

Parmi les approches pour la construction de hiérarchies sémantiques, (L.-J. Li *et al.*, 2010) ont présenté une méthode basée à la fois sur des informations visuelles et textuelles (les étiquettes associées aux images) pour construire automatiquement une hiérarchie, appelée «semantivisual», selon le modèle hLDA. Une troisième source d'information que nous nommons information contextuelle est aussi utilisée pour la construction de telles hiérarchies. Nous discutons plus précisément de ce type d'information dans la section suivante. Fan *et al.* (2009) ont proposé un algorithme qui intègre la similarité visuelle et la similarité contextuelle entre les concepts. Ces si-

milarités sont utilisées pour la construction d'un réseau de concepts utilisé pour la désambiguïsation des annotations. Une méthode pour la construction de hiérarchies basées sur la similarité contextuelle et visuelle est proposée dans (Fan *et al.*, 2007). La «distance de Flickr» est proposée dans (Wu *et al.*, 2008). Elle représente une nouvelle mesure de similarité entre les concepts dans le domaine visuel. Un réseau de concepts visuels (VCNet) basé sur cette distance est également proposé dans (Wu *et al.*, 2008). Ces hiérarchies sémantiques semblent avoir un potentiel intéressant pour améliorer l'annotation d'images.

Discussion

Comme nous venons de le voir, plusieurs approches se sont basées sur WordNet pour la construction des hiérarchies (Marszalek, Schmid, 2007 ; Deng *et al.*, 2009). Toutefois, WordNet n'est pas complètement approprié à la modélisation de la sémantique des images. En effet, l'organisation des concepts dans WordNet suit une structure psycholinguistique qui peut être utile pour raisonner sur les concepts et comprendre leur signification, mais elle est limitée et inefficace pour raisonner sur le contexte de l'image ou sur son contenu. Ainsi, les distances entre les concepts similaires dans WordNet ne reflètent pas nécessairement la proximité des concepts dans un cadre d'annotation d'images. Par exemple, selon la distance du plus court chemin dans WordNet, la distance entre les concepts "Requin" et "Baleine" est de 11 (nœuds), et celle entre "Humain" et "Baleine" est de 7. Cela signifie que le concept "Baleine" est plus proche (similaire) de "Humain" que de "Requin". Ceci est tout à fait cohérent d'un point de vue biologique, parce que "Baleine" et "Humain" sont des mammifères tandis que "Requin" ne l'est pas. Cependant, dans le domaine de l'image il est plus intéressant d'avoir une similarité plus élevée entre "Requin" et "Baleine", puisqu'ils vivent dans le même environnement, partagent de nombreuses caractéristiques visuelles, et il est donc plus fréquent qu'on les retrouve conjointement dans une même image ou un même type d'images (ils partagent un même contexte). Par conséquent, une hiérarchie sémantique appropriée devrait représenter cette information ou permettre de la déduire, pour aider à comprendre la sémantique de l'image.

3. Mesure proposée pour estimer la similarité sémantico-visuelle entre concepts

En se basant sur la discussion précédente, nous définissons les hypothèses suivantes sur lesquelles repose notre approche :

Une hiérarchie sémantique appropriée pour l'annotation d'images doit :

1. *modéliser le contexte des images (comme défini dans la section précédente),*

2. *refléter la sémantique des images,* i.e. *l'organisation des concepts dans la hiérarchie et leurs relations sémantiques est fidèle à la sémantique d'images,*

3. *permettre de regrouper des concepts selon leurs caractéristiques visuelles et sémantiques.*

Nous proposons dans cet article une nouvelle méthode pour la construction de hiérarchies sémantiques appropriées à l'annotation d'images. Notre méthode se base sur une nouvelle mesure pour estimer les relations sémantiques entre concepts. Cette mesure intègre les trois sources d'information que nous avons décrites précédemment. Elle est basée sur 1) une similarité visuelle qui représente la correspondance visuelle entre les concepts, 2) une similarité conceptuelle qui définit un degré de similarité entre les concepts cibles en se basant sur leur définition dans WordNet, et 3) une similarité contextuelle qui mesure la dépendance statistique entre chaque paire de concepts dans un corpus donné (cf. figure 3). Ensuite cette mesure est utilisée dans des règles qui permettent de statuer sur la vraisemblance des relations de parenté entre les concepts, et permettent de construire une hiérarchie.

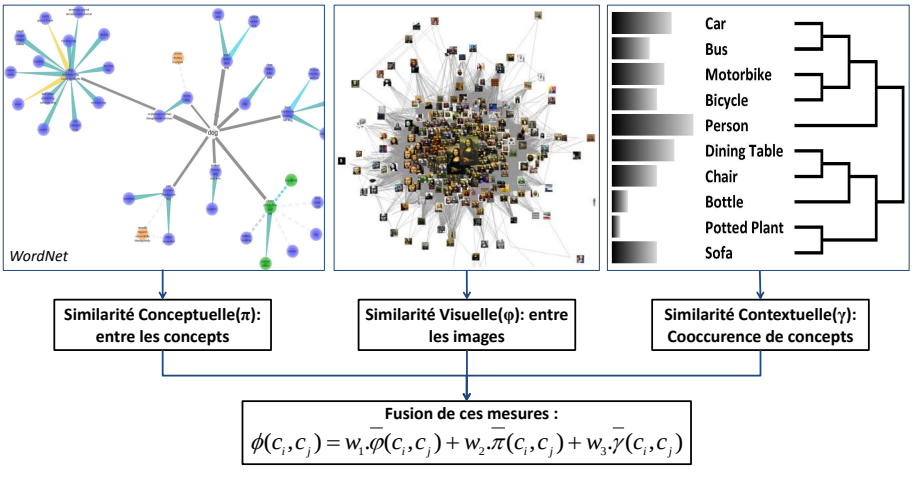


Figure 3. Illustration de la mesure proposée basée sur les similarités normalisées : visuelle $\overline{\varphi}$, conceptuelle $\overline{\pi}$ et contextuelle $\overline{\gamma}$ entre les concepts

Étant donné un ensemble de couples image/annotation, où chaque annotation décrit un ensemble de concepts associés à l'image, notre approche permet de créer automatiquement une hiérarchie sémantique adaptée à l'annotation d'images. Plus formellement, nous considérons $I = \langle i_1, i_2, \cdots, i_{\mathcal{L}} \rangle$ l'ensemble des images de la base considérée, et $C = \langle c_1, c_2, \cdots, c_{\mathcal{N}} \rangle$ le vocabulaire d'annotation de ces images, *i.e.* l'ensemble de concepts associés à ces images. L'approche que nous proposons consiste alors à identifier $\mathcal{M}$ nouveaux concepts qui permettent de relier tous les concepts de C dans une structure hiérarchique représentant au mieux la sémantique d'images.

3.1. Similarité visuelle

Soit x_i^v une représentation visuelle quelconque de l'image i_i (vecteur de caractéristiques visuelles), on apprend pour chaque concept c_j un classifieur qui permet d'associer ce concept à ses caractéristiques visuelles. Pour cela, nous utilisons $\mathcal{N}$ ma-

chines à vecteurs de support (SVM) (Cortes, Vapnik, 1995) binaires (un-contre-tous) avec une fonction de décision $\mathcal{G}(x_i^v)$:

$$\mathcal{G}(x_i^v) = \sum_k \alpha_k y_k \mathbf{K}(x_k^v, x_i^v) + b \tag{1}$$

où : $\mathbf{K}(x_k^v, x_i^v)$ est la valeur d'une fonction noyau pour l'échantillon d'apprentissage x_k^v et l'échantillon de test x_i^v, $y_k \in \{1, -1\}$ est l'étiquette de la classe de x_k^v, α_k est le poids appris de l'échantillon d'apprentissage x_k^v, et b est un paramètre seuil appris. Il est à noter que les échantillons d'apprentissage x_k^v avec leurs poids $\alpha_k > 0$ forment *les vecteurs de support*.

Après avoir testé différentes fonctions noyau sur notre ensemble d'apprentissage, nous avons choisi en fonction des performances obtenues d'utiliser une fonction noyau à base radiale :

$$\mathbf{K}(x_k^v, x_i^v) = exp\Big(\frac{\|x_k^v - x_i^v\|^2}{\sigma^2}\Big) \tag{2}$$

Maintenant, compte tenu de ces $\mathcal{N}$ SVM appris où les représentations visuelles des images sont les entrées et les concepts (classes d'images) sont les sorties, nous voulons définir pour chaque classe de concepts un centroïde $\vartheta(c_i)$ qui soit représentatif du concept c_i. Les centroïdes définis doivent alors minimiser la somme des carrés à l'intérieur de chaque ensemble S_i :

$$\underset{S}{\operatorname{argmin}} \sum_{i=1}^{\mathcal{N}} \sum_{x_j^v \in S_i} \|x_j^v - \mu_i\|^2 \tag{3}$$

où S_i est l'ensemble de *vecteurs de support* de la classe c_i, $S = \{S_1, S_2, \cdots, S_{\mathcal{N}}\}$, et μ_i est la moyenne des points dans S_i.

L'objectif étant d'estimer une distance entre ces classes afin d'évaluer leurs similarités visuelles, nous calculons le centroïde $\vartheta(c_i)$ de chaque concept visuel c_i en utilisant :

$$\vartheta(c_i) = \frac{1}{|S_i|} \sum_{x_j^v \in S_i} x_j^v \tag{4}$$

La similarité visuelle entre deux concepts c_i et c_j, est alors inversement proportionnelle à la distance entre leurs centroïdes respectifs $\vartheta(c_i)$ et $\vartheta(c_j)$:

$$\varphi(c_i, c_j) = \frac{1}{1 + d(\vartheta(c_i), \vartheta(c_j))} \tag{5}$$

où $d(\vartheta(c_i), \vartheta(c_j))$ est la distance euclidienne entre les deux vecteurs $\vartheta(c_i)$ et $\vartheta(c_j)$ définie dans l'espace des caractéristiques visuelles.

3.2. Similarité conceptuelle

La similarité conceptuelle reflète la relation sémantique entre deux concepts d'un point de vue linguistique et taxonomique. Plusieurs mesures de similarité ont été proposées dans la littérature (Budanitsky, Hirst, 2006 ; Resnik, 1995 ; Banerjee, Pedersen, 2003). La plupart sont basées sur une ressource lexicale, comme WordNet (Fellbaum, 1998). Une première famille d'approches se base sur la structure de cette ressource externe (souvent un réseau sémantique ou un graphe orienté) et la similarité est alors calculée en fonction des distances des chemins reliant les concepts dans cette structure (Budanitsky, Hirst, 2006). Cependant, comme nous l'avons déjà dit précédemment, la structure de ces ressources ne reflète pas forcément la sémantique des images, et ce type de mesures ne semble donc pas adapté à notre problématique. Une approche alternative pour mesurer le degré de similarité sémantique entre deux concepts est d'utiliser la définition textuelle associée à ces concepts. Dans le cas de WordNet, ces définitions sont connues sous le nom de glosses. Par exemple, Banerjee et Pedersen (Banerjee, Pedersen, 2003) ont proposé une mesure de proximité sémantique entre deux concepts qui est basée sur le nombre de mots communs (chevauchements) dans leurs définitions (glosses).

Dans notre approche, nous avons utilisé la mesure de similarité proposée par (Patwardhan, Pedersen, 2006), qui se base sur WordNet et l'exploitation des vecteurs de co-occurrences du second ordre entre les glosses. Plus précisément, dans une première étape un espace de mots de taille $\mathcal{P}$ est construit en prenant l'ensemble des mots significatifs utilisés pour définir l'ensemble des synsets [3] de WordNet. Ensuite, chaque concept c_i est représenté par un vecteur $\overrightarrow{w}_{c_i}$ de taille $\mathcal{P}$, où chaque *ième* élément de ce vecteur représente le nombre d'occurrences du *ième* mot de l'espace des mots dans la définition de c_i. La similarité sémantique entre deux concepts c_i et c_j est alors mesurée en utilisant la similarité cosinus entre $\overrightarrow{w}_{c_i}$ et $\overrightarrow{w}_{c_j}$:

$$\eta(c_i, c_j) = \frac{\overrightarrow{w}_{c_i} \cdot \overrightarrow{w}_{c_j}}{|\overrightarrow{w}_{c_i}||\overrightarrow{w}_{c_j}|} \tag{6}$$

Certaines définitions de concepts dans WordNet sont très concises et rendent cette mesure peu fiable. En conséquence, Patwardhan et Pedersen (2006) ont proposé d'étendre les glosses des concepts avec les glosses des concepts situés dans leur voisinage d'ordre 1. Ainsi, pour chaque concept c_i l'ensemble Ψ_{c_i} est défini comme l'ensemble des glosses adjacents connectés au concept c_i (Ψ_{c_i}={gloss(c_i), gloss(hyponyms(c_i)), gloss(meronyms(c_i)), etc.}). Ensuite pour chaque élément x (gloss) de Ψ_{c_i}, sa représentation $\overrightarrow{w}_x$ est construite comme expliqué ci-dessus. La mesure de similarité

3. Synonym set : composante atomique sur laquelle repose WordNet, composée d'un groupe de mots interchangeables dénotant un sens ou un usage particulier. A un concept correspond un ou plusieurs synsets.

entre deux concepts c_i et c_j est alors définie comme la somme des cosinus individuels des vecteurs correspondants :

$$\theta(c_i, c_j) = \frac{1}{|\Psi|} \sum_{x \in \Psi_{c_i}, y \in \Psi_{c_j}} \frac{\overrightarrow{w}_x \cdot \overrightarrow{w}_y}{|\overrightarrow{w}_x||\overrightarrow{w}_y|} \quad (7)$$

où : $|\Psi| = |\Psi_{c_i}| = |\Psi_{c_j}|$.

Enfin, chaque concept dans WordNet peut correspondre à plusieurs sens (synsets) qui diffèrent les uns des autres dans leur position dans la hiérarchie et leur définition. Une étape de désambiguïsation est donc nécessaire pour l'identification du bon synset. Par exemple, la similarité entre "Souris" (animal) et "Clavier" (périphérique) diffère largement de celle entre "Souris" (périphérique) et "Clavier" (périphérique). Ainsi, nous calculons d'abord la similarité conceptuelle entre les différents sens (synset) de c_i et c_j. La valeur maximale de similarité est ensuite utilisée pour identifier le sens le plus probable de ces deux concepts, *i.e.* désambigüiser c_i et c_j. La similarité conceptuelle est alors calculée par la formule suivante :

$$\pi(c_i, c_j) = \underset{\delta_i \in s(c_i), \delta_j \in s(c_j)}{\operatorname{argmax}} \theta(\delta_i, \delta_j) \quad (8)$$

où $s(c_x)$ est l'ensemble des synsets qu'il est possible d'associer aux différents sens du concept c_x.

3.3. Similarité contextuelle

Comme cela a été expliqué dans la section 2, l'information liée au contexte d'apparition des concepts est très importante dans un cadre d'annotation d'images. En effet, cette information, dite contextuelle, permet de relier des concepts qui apparaissent souvent ensemble dans des images ou des mêmes types d'images, bien que sémantiquement éloignés du point de vue taxonomique. De plus, cette information contextuelle peut aussi permettre d'inférer des connaissances de plus haut niveau sur l'image. Par exemple, si une photo contient "Mer" et "Sable", il est probable que la scène représentée sur cette photo est celle de la plage. Il semble donc important de pouvoir mesurer la similarité contextuelle entre deux concepts. Contrairement aux deux mesures de similarité précédentes, la mesure de similarité contextuelle dépend du corpus, ou plus précisément dépend de la répartition des concepts dans le corpus.

Dans notre approche, nous modélisons la similarité contextuelle entre deux concepts c_i et c_j par l'information mutuelle PMI (Church, Hanks, 1990) (Pointwise mutual information) $\rho(c_i, c_j)$:

$$\rho(c_i, c_j) = \log \frac{P(c_i, c_j)}{P(c_i)P(c_j)} \quad (9)$$

où, $P(c_i)$ est la probabilité d'apparition de c_i, et $P(c_i, c_j)$ est la probabilité jointe de c_i et de c_j. Ces probabilités sont estimées en calculant les fréquences d'occurrence et de co-occurrence des concepts c_i et c_j dans la base d'images.

Étant donné $\mathcal{N}$ le nombre total de concepts dans notre base d'images, $\mathcal{L}$ le nombre total d'images, n_i le nombre d'images annotées par c_i (fréquence d'occurrence de c_i) et n_{ij} le nombre d'images co-annotées par c_i et c_j, les probabilités précédentes peuvent être estimées par :

$$\widehat{P(c_i)} = \frac{n_i}{\mathcal{L}}, \quad \widehat{P(c_i, c_j)} = \frac{n_{ij}}{\mathcal{L}} \tag{10}$$

Ainsi :

$$\rho(c_i, c_j) = \log \frac{\mathcal{L} * n_{ij}}{n_i * n_j} \tag{11}$$

$\rho(c_i, c_j)$ quantifie la quantité d'information partagée entre les deux concepts c_i et c_j. Ainsi, si c_i et c_j sont des concepts indépendants, alors $P(c_i, c_j) = P(c_i) \cdot P(c_j)$ et donc $\rho(c_i, c_j) = log\ 1 = 0$. $\rho(c_i, c_j)$ peut aussi être négative si c_i et c_j sont corrélés négativement. Sinon, $\rho(c_i, c_j) > 0$ et quantifie le degré de dépendance entre ces deux concepts. Dans ce travail, nous cherchons uniquement à mesurer la dépendance positive entre les concepts et donc les valeurs négatives de $\rho(c_i, c_j)$ sont ramenées à 0.

Enfin, afin de la normaliser dans l'intervalle [0,1], nous calculons la similarité contextuelle entre deux concepts c_i et c_j dans notre approche par :

$$\gamma(c_i, c_j) = \frac{\rho(c_i, c_j)}{-\log[\max(P(c_i), P(c_j))]} \tag{12}$$

Il est à noter que la mesure PMI dépend de la distribution des concepts dans la base. Plus un concept est rare plus sa PMI est grande. Donc si la distribution des concepts dans la base n'est pas uniforme, il est préférable de calculer ρ par :

$$\rho(c_i, c_j) = P(c_i, c_j) \log \frac{P(c_i, c_j)}{P(c_i)P(c_j)} \tag{13}$$

3.4. Mesure de similarité proposée

Pour deux concepts donnés, les mesures de similarité visuelle, conceptuelle et contextuelle sont d'abord normalisées dans le même intervalle. La normalisation est faite par la normalisation Min-Max. Puis en combinant les mesures précédentes, nous obtenons la mesure de similarité sémantique adaptée à l'annotation suivante :

$$\phi(c_i, c_j) = \omega_1 \cdot \overline{\varphi}(c_i, c_j) + \omega_2 \cdot \overline{\pi}(c_i, c_j) + \omega_3 \cdot \overline{\gamma}(c_i, c_j) \tag{14}$$

où : $\sum_{i=1}^{3} \omega_i = 1$; $\overline{\varphi}(c_i, c_j)$, $\overline{\pi}(c_i, c_j)$ et $\overline{\gamma}(c_i, c_j)$ sont respectivement la similarité visuelle, la similarité conceptuelle et la similarité contextuelle normalisées.

Le choix des pondérations ω_i est très important. En effet, selon l'application ciblée, certains préféreront construire une hiérarchie spécifique à un domaine (qui représente le mieux une particularité d'un domaine ou d'un corpus), et pourront donc attribuer

un plus fort poids à la similarité contextuelle ($\omega_3 \nearrow$). D'autres pourront vouloir créer une hiérarchie générique, et devront par conséquent donner plus de poids à la similarité conceptuelle ($\omega_2 \nearrow$). Toutefois, si le but de la hiérarchie est plutôt de construire une plateforme pour la classification de concepts visuels, il est peut-être avantageux de donner plus de poids à la similarité visuelle ($\omega_1 \nearrow$).

4. Construction de la hiérarchie sémantique

La mesure proposée précédemment ne permet que de donner une information sur la similarité entre les concepts deux à deux. Notre objectif est de regrouper ces différents concepts dans une structure hiérarchique. Pour cela, nous définissons un ensemble de règles qui permettent d'inférer les relations d'hypernymie entre les concepts. Nous définissons d'abord les fonctions suivantes sur lesquelles se basent nos règles de raisonnement :

– $Closest(c_i)$ qui retourne le concept le plus proche de c_i selon notre mesure :

$$Closest(c_i) = \underset{c_k \in \mathcal{C} \setminus \{c_i\}}{\operatorname{argmax}} \phi(c_i, c_k) \tag{15}$$

– $LCS(c_i, c_j)$ permet de trouver l'ancêtre commun le plus proche (*Least Common Subsumer*) de c_i et c_j dans WordNet :

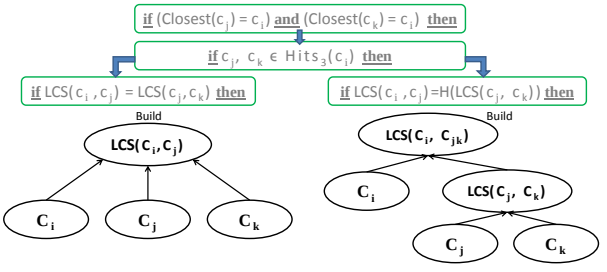
$$LCS(c_i, c_j) = \underset{c_l \in \{H(c_i) \cap H(c_j)\}}{\operatorname{argmax}} len(c_l, root) \tag{16}$$

où : $H(c_i)$ permet de trouver l'ensemble des hypernymes de c_i dans la ressource WordNet, $root$ représente la racine de la hiérarchie WordNet et $len(c_x, root)$ renvoie la longueur du plus court chemin entre c_x et $root$ dans WordNet.

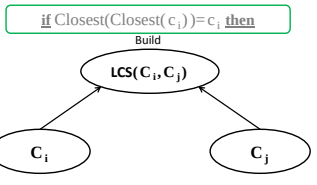
– $Hits_3(c_i)$ renvoie les 3 concepts les plus proches de c_i au sens de la fonction $Closest(c_i)$.

Nous définissons ensuite trois règles qui permettent d'inférer les liens de parenté entre les différents concepts. Ces différentes règles sont représentées graphiquement sur la figure 4, et sont exécutées selon le même ordre que celui dans cette figure. La première règle vérifie si un concept c_i est classé comme le plus proche par rapport à plusieurs concepts ($(Closest(c_j) = c_i), \forall j \in \{1, 2, \cdots\}$). Si oui et si ces concepts $\{c_j\}, \forall j \in \{1, 2, \cdots\}$, sont réciproquement dans $Hits_3(c_i)$, alors en fonction de leur *ancêtre commun le plus proche* (LCS) ils seront soit reliés directement à leur LCS ou dans une structure à 2 niveaux, comme illustré dans la figure 4a. Dans la seconde règle, si $(Closest(c_i) = c_j)$ et $(Closest(c_j) = c_i)$ (peut aussi être écrite $Closest(Closest(c_i)) = c_i$) alors c_i et c_j sont fortement apparentés et seront reliés à leur LCS. La troisième règle concerne le cas où $(Closest(c_i) = c_j)$ et $(Closest(c_j) = c_k)$ - voir figure 4c.

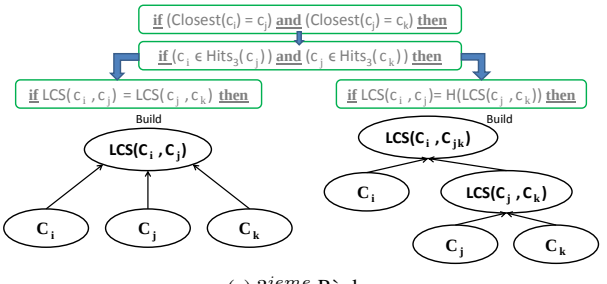
La construction de la hiérarchie suit une approche ascendante (*i.e.* commence à partir des concepts feuilles) et utilise un algorithme itératif jusqu'à atteindre le nœud



(a) 1^{ere} Règle



(b) 2^{ieme} Règle



(c) 3^{ieme} Règle.

Figure 4. Règles pour inférer les liens de parenté entre les différents concepts. En gris les préconditions devant être satisfaites, en noir les actions de création de nœuds dans la hiérarchie

racine. Étant donné un ensemble de concepts associés aux images dans un ensemble d'apprentissage, notre méthode calcule la similarité $\phi(c_i, c_j)$ entre toutes les paires de concepts, puis relie les concepts les plus apparentés tout en respectant les règles définies précédemment. La construction de la hiérarchie se fait donc pas-à-pas en ajoutant un ensemble de concepts inférés des concepts du niveau inférieur. On itère le processus jusqu'à ce que tous les concepts soient liés à un seul nœud racine. La complexité de notre algorithme de construction de hiérarchies est donc de l'ordre de $\mathcal{O}(\mathcal{N}^2)$, avec $\mathcal{N}$ la taille du vocabulaire d'annotation initiale. En effet, l'algorithme nécessite deux étapes : i) le calcul de la similarité entre chaque paire de concepts qui se fait en $\mathcal{O}(\mathcal{N}^2)$, et ii) relier chaque paire de concepts à son ancêtre commun le plus proche en utilisant les règles illustrées dans la figure 4, étape qui se fait dans le pire des cas en $\mathcal{O}(\mathcal{N}log(\mathcal{N}))$.

5. Méthode proposée pour la classification hiérarchique

L'annotation d'images a été considérée, au cours de la dernière décennie, comme un problème de classification multiclasse. De nombreuses approches ont proposé de combiner des structures hiérarchiques avec des classifieurs SVMs pour faire face à un grand nombre de catégories (concepts) (Marszalek, Schmid, 2007 ; Griffin, Perona, 2008 ; L.-J. Li *et al.*, 2010 ; Fan *et al.*, 2008 ; Cevikalp, 2010 ; Bannour, Hudelot, 2012c). Ces approches peuvent être qualifiées de méthodes descendantes, *i.e.* la hiérarchie de classes est construite par un partitionnement récursif de l'ensemble des classes (Griffin, Perona, 2008 ; Cevikalp, 2010 ; Gao, Koller, 2011), ou de méthodes ascendantes, *i.e.* la hiérarchie de classes est construite par un regroupement agglomératif de l'ensemble des classes (Marszalek, Schmid, 2007 ; Fan *et al.*, 2008 ; L.-J. Li *et al.*, 2010 ; Bannour, Hudelot, 2012a). Deux directions ont été explorées pour la classification hiérarchique d'images : 1) en utilisant des graphes de décision acycliques orientés (GDAO) (Platt *et al.*, 2000 ; Marszalek, Schmid, 2007 ; Gao, Koller, 2011), et 2) en utilisant des arbres de décision binaires hiérarchiques (ADBH) (Griffin, Perona, 2008 ; Cevikalp, 2010). Etant donné, le vocabulaire d'annotation $C = \langle c_1, c_2, \cdots, c_{\mathcal{N}} \rangle$ de la base, les approches basées sur les GDAO nécessitent l'apprentissage de $N(N-1)/2$ classifieurs binaires et utilisent un graphe acyclique orienté (GAO) pour décider de l'appartenance d'une image i_i à une classe $c_j \in C$. Ces méthodes permettent d'éliminer d concepts candidats à chaque nœud à une distance d de la racine du GAO, et donc $N-1$ nœuds de décision doivent être évalués pour l'annotation d'une nouvelle image. Par contre, les approches basées sur les ADBH construisent et utilisent les hiérarchies sous formes d'arbres binaires, *i.e.* les données sont divisées hiérarchiquement en deux sous-ensembles jusqu'à ce que chaque sous-ensemble est constitué d'une seule classe. Cette partition des données est souvent réalisée en utilisant un algorithme de clustering. Ainsi, un SVM est formé pour chaque nœud de l'arbre, résultant en une exécution de $\log_2 N$ SVM pour annoter une nouvelle image. Les approches basées sur les ADBH visent à optimiser l'efficacité des classifieurs SVM en réduisant les comparaisons inutiles tout en conservant une bonne précision de la classification (Cevikalp, 2010).

Cependant, les approches basées sur les ADBH et GDAO se focalisent sur l'optimisation de la classification hiérarchique et ne se préoccupent en aucune façon de modéliser la sémantique de l'image. Bien que ces approches permettent d'améliorer la précision de la classification, elles limitent la construction de leurs hiérarchies à des structures binaires, ce qui a pour conséquence de les limiter à un vocabulaire d'annotation restreint. Par exemple, la méthode de (Marszalek, Schmid, 2007) ne permet plus de construire des hiérarchies dès que le nombre de concepts dépasse 40, puisque les concepts intermédiaires sont extraits de WordNet en fonction de la relation de hypernymie, alors que la profondeur de WordNet est limitée à 20 niveaux.

Récemment, d'autres approches ont proposé l'utilisation des relations sémantiques entre les concepts pour la construction des hiérarchies. Fan *et al.* (2008) ont proposé d'intégrer une ontologie de concepts et un algorithme d'apprentissage «multitâches» pour l'apprentissage hiérarchique des concepts. L'annotation d'une nouvelle image est

obtenue par une procédure de vote à tous les niveaux de la hiérarchie, *i.e.* une exécution de $|C + C'|$ SVM est nécessaire pour l'étiquetage de l'image. Deng *et al.* (2009) ont proposé une méthode de classification hiérarchique, nommée «tree-max classifier», qui se base sur la structure de ImageNet pour l'apprentissage des classifieurs et le calcul de la fonction de décision. Dans la suite de cette section, nous proposons deux nouvelles méthodes de classification hiérarchique d'images. L'apprentissage des classifieurs hiérarchiques s'appuie sur la structure de la hiérarchie pour décomposer le problème en plusieurs tâches indépendantes et complémentaires, ce qui va permettre le passage à l'échelle de nos méthodes. La fonction de décision pour chacune de nos méthodes est également calculée en fonction de la structure de la hiérarchie afin d'aboutir à une meilleure annotation de l'image.

5.1. Apprentissage des classifieurs hiérarchiques : «un-contre-nœuds-opposés»

Pour introduire notre méthode de classification hiérarchique, nous basons notre explication sur la structure de la hiérarchie sémantique illustrée dans la figure 6. Ainsi, la classification hiérarchique est effectuée par l'apprentissage d'un ensemble de ($\mathcal{N}$+$\mathcal{M}$) classifieurs hiérarchiques conformes à la structure de la hiérarchie, où $\mathcal{M} = |C'|$ est le nombre de nouveaux concepts inférés lors de la construction de la hiérarchie. En effet, en se basant sur la structure de la hiérarchie, nous proposons d'apprendre plusieurs classifieurs hiérarchiques qui représentent le même concept dans des niveaux d'abstraction différents. Ces classifieurs sont consistants les uns avec les autres puisqu'ils sont liés par la relation de subsomption, et représentent donc la même information avec différents niveaux de détails. Par conséquent, les résultats de ces classifieurs peuvent être fusionnés afin de parvenir à une décision pertinente sur l'appartenance d'une image à une classe donnée (Bannour, Hudelot, 2012c).

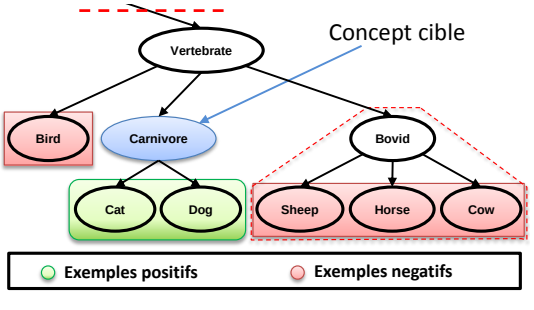


Figure 5. Apprentissage des classifieurs hiérarchiques «un-contre-nœuds-opposés»

Concrètement, un classifieur est entraîné pour chaque nœud concept de la hiérarchie en utilisant un modèle SVM qu'on va nommer *un-contre-nœuds-opposés* - cf. figure 5. En effet, afin de proposer une méthode de classification qui passe à l'échelle, nous allons décomposer le problème en plusieurs tâches indépendantes en fonction de la structure de la hiérarchie. Ainsi, au lieu de considérer toutes les images de la base pour la formation des classifieurs, nous utilisons seulement les images des nœuds fils

d'un concept cible donné. Ceci est similaire à ne considérer que le sous-arbre dont la racine est le nœud père du concept cible pour lequel on veut associer un classifieur. Par conséquent, pour l'apprentissage du classifieur d'un nœud cible, nous utiliserons toutes les images de ses nœuds fils comme échantillons positifs. Les échantillons négatifs sont toutes les images des nœuds fils de son ancêtre immédiat. Par exemple, pour apprendre un classifieur pour le concept "Carnivore", les images de "Dog" et "Cat" seront prises comme exemples positifs et les images de "Bird", "Sheep", "Horse" et "Cow" comme exemples négatifs - cf. figure 5.

5.2. Fonction de décision hiérarchique ascendante : approche par fusion des scores des classifieurs hiérarchiques (AFSCH)

Le calcul de la fonction de décision hiérarchique s'effectue avec une approche ascendante. À partir des nœuds feuilles et en suivant les relations de subsomption, nous calculons la moyenne des scores de confiance de tous les chemins dans la hiérarchie. La fonction de décision hiérarchique est alors obtenue en fonction du signe de cette moyenne de score. Une propriété importante de cette méthode est que les résultats de la classification de ces SVM hiérarchiques sont indépendants. Par conséquent, il est également possible de calculer au préalable la fonction de décision de chaque classifieur, puis de fusionner, en fonction de la structure de la hiérarchie, les scores des classifieurs correspondant à un chemin donné dans la hiérarchie. Ainsi, la complexité de l'annotation d'une image donnée est $\leq (2N-1)$ SVM.

Finalement, la fonction de décision hiérarchique qui permet de calculer le degré d'appartenance d'une image i_i (avec une représentation visuelle x_i^v) à une classe $c_j \in C$ est la suivante :

$$\overline{f_{c_j}(x_i^v)} = sign\Big(\frac{1}{|\mathcal{A}|}\sum_{c_l \in \mathcal{A}} \mathcal{G}_{c_l}(x_i^v)\Big) \tag{17}$$

où, $\mathcal{A}$ est l'ensemble des ancêtres de c_j. $\mathcal{G}_{c_l}(x_i^v)$ est la fonction de décision du classifieur associé au concept c_l - cf. Equation 1.

D'un point de vue statistique, la fonction de décision hiérarchique $\overline{f_{c_j}(x_i^v)}$ peut être considérée comme la réalisation de n mesures du même événement (n=$|\mathcal{A}|$, est la profondeur de la hiérarchie). Ainsi, l'incertitude sur $\overline{f_{c_j}(x_i^v)}$ peut être calculée comme l'écart-type $\sigma_{\overline{f_{c_j}(x_i^v)}} = \frac{\sigma}{\sqrt{n}}$. Par conséquent, la fonction de décision hiérarchique est $\sqrt{n}$ fois plus précise que la fonction de décision obtenue à partir d'un seul classifieur (classification plate).

5.3. Fonction de décision hiérarchique descendante : approche par vote des classifieurs hiérarchiques (AVCH)

La méthode *AVCH* vise à décomposer le problème de classification d'image en plusieurs sous-tâches complémentaires. Elle consiste alors à construire plusieurs classifieurs hiérarchiques capables de discriminer une classe (un concept) parmi d'autres

sous un nœud parent donné. Ainsi, pour parvenir à la décision finale concernant l'appartenance d'une image à une classe, il est essentiel de parcourir la hiérarchie en fonction des votes (réponses) des classifieurs. La méthode *AVCH* est efficace en termes de complexité, puisqu'elle nécessite l'apprentissage de moins de $2N - 1$ classifieurs pour la classification hiérarchique, et l'évaluation de moins de $\log_2 N$ nœuds de décision pour l'annotation d'une nouvelle image - cf. tableau 1. Cependant, cette méthode est sensible à la classification initiale, *i.e.* les classifieurs aux niveaux inférieurs ne peuvent pas se remettre d'une erreur de classification qui peut se produire à un niveau plus élevé. Ainsi, cette erreur de classification va se propager vers les nœuds fils. Néanmoins, la précision moyenne est très élevée pour les nœuds/concepts se trouvant dans les niveaux intermédiaires les plus élevés de la hiérarchie, et par conséquent la propagation des erreurs reste relativement faible - cf. figure 10.

Algorithme 1 : approche par vote des classifieurs hiérarchiques (AVCH)

Entrées : La hiérarchie sémantique, l'image à annoter
Sorties : Image annotée
début
  $\Omega \leftarrow$ Nœuds fils direct de la racine de la hiérarchie
  tant que $(|\Omega| > 0)$ **faire**
    $\Upsilon \leftarrow \emptyset$
    pour chaque $(c_l \in \Omega)$ **faire**
      si $(\mathcal{G}_{c_l}(x_i^v) > 0)$ **alors**
        $\Upsilon \leftarrow \Upsilon + c_l$
    si $(|\Upsilon| = 0)$ **alors**
      $\Upsilon \leftarrow \underset{c_l \in \Omega}{\operatorname{argmax}}\ \mathcal{G}_{c_l}(x_i^v)$
    $\Omega \leftarrow$ Enfants immédiats des noeuds $\in \Upsilon$
  retourner Υ

La classification des images est effectuée dans une approche descendante, comme illustrée dans l'algorithme 1. En partant du nœud racine, les fonctions de décision des nœuds se trouvant dans le niveau inférieur sont évaluées. Les nœuds ayant une valeur de confiance positive sont récursivement explorés jusqu'à atteindre les nœuds feuilles. Plusieurs chemins de la hiérarchie peuvent être explorés, et ainsi l'image en entrée peut être associée à plusieurs classes/concepts. Notons finalement, que si un chemin est exploré sans aucune réponse positive de la part de ses nœuds feuilles, le concept avec la plus grande valeur de confiance sera associé à l'image.

6. Résultats expérimentaux

Pour valider notre approche, nous utilisons les données du challenge Pascal VOC' 2010 (11 321 images, 20 concepts). Les images et leurs annotations sont utilisées pour la construction de la hiérarchie sémantique et pour l'évaluation des performances de la classification d'images.

6.1. Représentation visuelle

Pour calculer la similarité visuelle des concepts, nous avons utilisé dans notre approche le modèle de sac-de-mots visuels (Bag-of-Features) (BoF)(F.-F. Li, Perona, 2005). Le modèle BoF utilisé est construit comme suit : détection de caractéristiques visuelles à l'aide des détecteurs DoG de Lowe (Lowe, 1999), description de ces caractéristiques visuelles en utilisant le descripteur SIFT (Lowe, 1999), puis génération du dictionnaire de mots visuels en utilisant un K-Means. Le dictionnaire généré est un ensemble de caractéristiques supposées être représentatives de toutes les caractéristiques visuelles de la base. Étant donnée la collection de patches (points d'intérêts) détectés dans les images de l'ensemble d'apprentissage, nous générons un dictionnaire de taille $D = 1000$ en utilisant l'algorithme k-Means. Ensuite, chaque patch dans une image est associé au mot visuel le plus similaire dans le dictionnaire en utilisant un arbre KD. Chaque image est alors représentée par un histogramme de 1000 mots visuels (1000 étant la taille du dictionnaire), où chaque bin dans l'histogramme correspond au nombre d'occurrences d'un mot visuel dans cette image.

6.2. Pondération

Dans le cadre de ce travail, nous allons évaluer la qualité de la hiérarchie produite en comparant la précision moyenne de la classification hiérarchique d'images en utilisant notre hiérarchie à celle de la classification plate d'images. Par conséquent, notre objectif est de construire une hiérarchie adaptée à l'annotation/classification d'images. Nous avons donc fixé les facteurs de pondération de manière *expérimentale* comme suit : $\omega_1 = 0,4$, $\omega_2 = 0,3$, et $\omega_3 = 0,3$. Ces facteurs de pondération ont été trouvés par un processus de validation croisée visant à maximiser la précision de la classification hiérarchique d'images, et en utilisant des pas d'ajustement des poids de $0,1$. En effet, nous avons remarqué que l'utilisation de pas d'ajustement des poids plus faible ($\triangle_{\omega_i} < 0,1$) n'a pas d'incidence importante sur la hiérarchie produite, mais en contrepartie engendre un temps d'exécution très important. Nous rappelons aussi que les résultats de la classification d'images dépendent de la structure de la hiérarchie et non pas des poids utilisés. Nos expérimentations sur l'impact des poids (ω_i) ont également montré que la similarité visuelle est plus représentative de la similarité sémantique des concepts, comme cela est illustré sur la figure 6 avec la hiérarchie produite. Cette hiérarchie est construite sur les données de Pascal VOC'2010.

Dans la figure 7, nous illustrons les matrices d'affinité sémantique entre les concepts de VOC'2010 en fonction des différentes modalités de l'image, *i.e.* la modalité visuelle, conceptuelle, contextuelle et la modalité sémantico-visuelle que nous proposons. Pour chaque modalité, nous représentons la matrice d'affinité sémantique (appelée aussi matrice de distance) entre les concepts avec une carte de chaleur [4]. Nous

4. Une carte de chaleur est une représentation graphique des données où les valeurs individuelles contenues dans la matrice sont représentées comme des couleurs. Ces cartes permettent de visualiser la proximité des individus en fonction d'une gamme de couleurs prédéfinies.

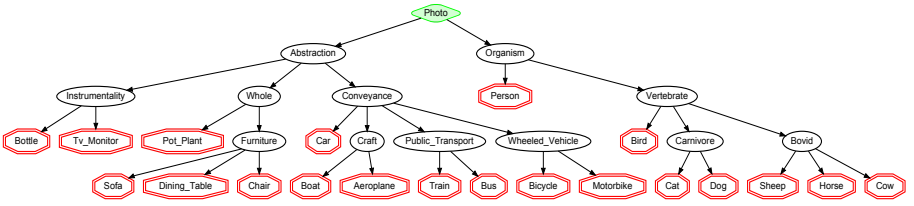
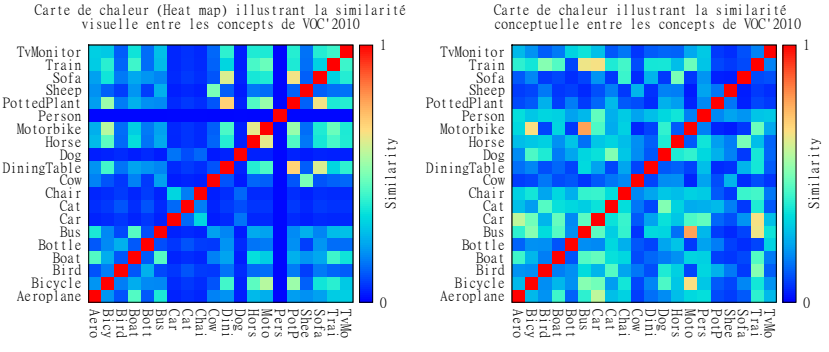


Figure 6. La hiérarchie sémantique construite sur les données de Pascal VOC'2010 en utilisant la mesure proposée et les règles de construction. Les nœuds en double octogone sont les concepts de départ, le nœud en diamant est la racine de la hiérarchie construite et les autres sont les nœuds inférés



(a) Similarité visuelle

(b) Similarité conceptuelle

(c) Similarité contextuelle

(d) Notre mesure de similarité

Figure 7. Cartes de chaleur (Heat maps) illustrant les matrices d'affinité sémantique entre les concepts de VOC'2010 en fonction de la mesure de similarité utilisée. (a) illustre la similarité visuelle, (b) la similarité conceptuelle, (c) la similarité contextuelle et (d) notre mesure de similarité sémantico-visuelle

pouvons remarquer sur cette figure que la corrélation entre les différentes paires de concepts varie largement en fonction de la modalité de l'image considérée, et respectivement la mesure utilisée. Par conséquent, nous pouvons conclure que la distribution de chaque similarité, calculée sur une des modalités de l'image, est indépendante des autres. Ce résultat est tout à fait justifiable. En effet, la similarité visuelle ne reflète que la similarité perceptuelle entre les concepts, la similarité contextuelle est une mesure qui dépend du corpus (dépend de la distribution des concepts dans le corpus), et la similarité conceptuelle est sensible au contexte. Les résultats présentés dans la figure 7, quoiqu'ils contredisent les hypothèses de Deselaers et Ferrari (2011), fournissent une preuve convaincante que la sémantique visuelle, la sémantique conceptuelle et la sémantique contextuelle ne sont pas toujours corrélées, et présentent des distributions différentes et indépendantes (en raison des problèmes mentionnés ci-dessus).

6.3. Évaluation

Pour évaluer notre approche, nous avons utilisé $50\ \%$ des images du challenge Pascal VOC'2010 pour l'apprentissage des classifieurs et les autres images pour les tests. Chaque image peut appartenir à une ou plusieurs des 20 classes (concepts) existantes. La classification plate est effectuée par l'apprentissage de $\mathcal{N}$ SVM binaires un-contre-tous, où les entrées sont les représentations en BoF des images de la base et les sorties sont les réponses du SVM pour chaque image (1 ou -1) - pour plus de détails voir la section 3.1. Un problème important dans les données de Pascal VOC est que les données ne sont pas équilibrées, *i.e.* plusieurs classes ne contiennent qu'une centaine d'images positives parmi les 11 321 images de la base. Pour remédier à ce problème, nous avons utilisé la validation croisée d'ordre 5 en prenant à chaque fois autant d'images positives que négatives. La classification hiérarchique est effectuée par l'apprentissage d'un ensemble de $(\mathcal{N} + \mathcal{M})$ classifieurs hiérarchiques, comme présenté dans la section 5. La méthode de référence est construite en prenant la moyenne des résultats de soumissions au challenge VOC'2010. Dans la suite, les évaluations sont effectuées en utilisant les courbes de rappel/précision et le score de précision moyenne (AP).

Dans la figure 8, nous avons comparé notre méthode *un-contre-nœuds-opposés* pour l'apprentissage de classifieurs hiérarchiques à la méthode *un-contre-tous*. Notre méthode permet d'obtenir un meilleur résultat que celui obtenu par la méthode *un-contre-tous*, avec une précision moyenne de $63,25\ \%$ contre $56,42\ \%$ pour la classification hiérarchique *un-contre-tous*.

La figure 9 illustre les performances de nos méthodes de classification hiérarchique d'images en comparaison avec d'autres, *i.e.* une méthode de classification plate décrite dans le début de cette section, la méthode *H-SVM* de (Marszalek, Schmid, 2007) et la méthode de référence qui représente la moyenne des résultats de soumissions au challenge VOC'2010. Concernant l'implémentation de la méthode *H-SVM*, nous avons construit une hiérarchie de concepts à partir de WordNet en utilisant la base d'images Pascal VOC'2010. La construction de la hiérarchie ainsi que l'apprentissage

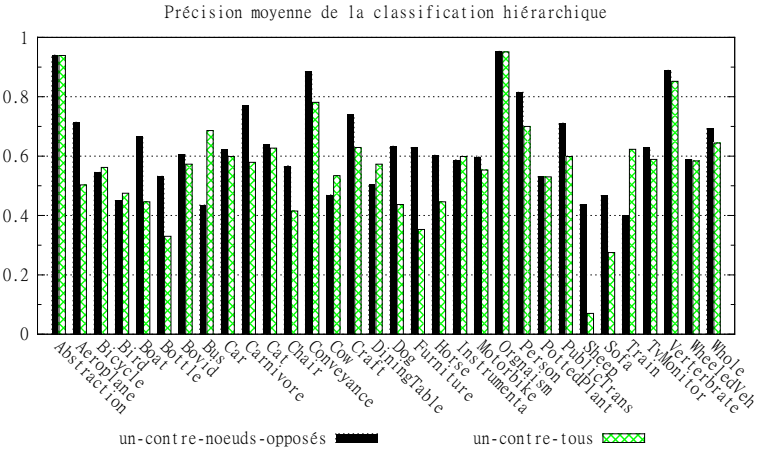


Figure 8. Comparaison des performances de la classification hiérarchique par les méthodes «un-contre-nœuds-opposés» *et* «un-contre-tous» *sur les données du challenge Pascal VOC'2010*

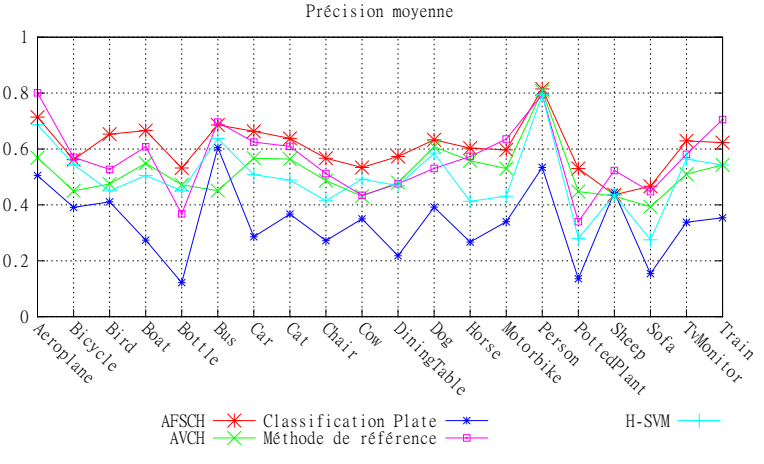
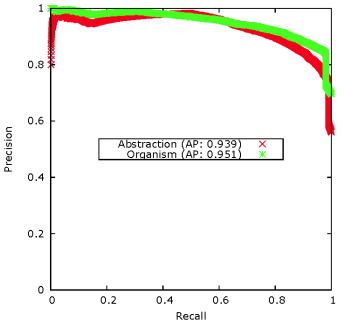


Figure 9. Comparaison des performances de nos méthodes de classification hiérarchique (AFSCH et AVCH) avec les méthodes suivantes : classification plate, H-SVM (Marszalek et al., 2007) et la méthode de référence. Ces résultats sont obtenus sur les images du chalenge VOC'2010

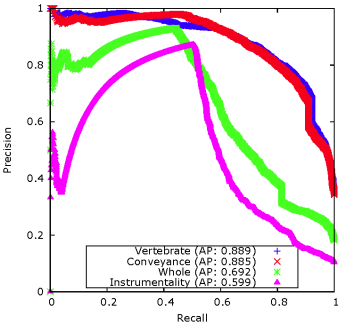
des classifieurs sont faits de la même manière décrite dans (Marszalek, Schmid, 2007). Comme nous pouvons le constater sur cette figure, l'utilisation de la hiérarchie proposée dans un cadre de classification hiérarchique assure des meilleures performances qu'une classification plate, avec une amélioration moyenne de $+26,8\ \%$ pour la méthode *AFSCH* et une amélioration de $+16,04$ pour la méthode *AVCH*. En comparaison avec la méthode de référence, nos approches de classification affichent une petite

amélioration en termes de précision moyenne. Cependant, ce résultat reste prometteur puisque nous avons utilisé un descripteur d'images assez basique, alors que les méthodes soumises au challenge Pascal VOC utilisent des descripteurs d'images beaucoup plus performants en termes de dimension, *i.e.* des descripteurs qui regroupent plusieurs attributs de bas niveau de l'image (par exemple, SIFT+ caractéristiques de couleur ou de texture). Notons en plus, que nos résultats sont obtenus en n'utilisant que la moitié des images du jeu d'apprentissage de Pascal VOC. En effet, en l'absence des images de test utilisées dans le challenge, nous avons utilisé la deuxième moitié de l'ensemble d'apprentissage pour faire les tests. Nous avons aussi inclus les images marquées comme difficiles dans les évaluations de notre méthode. Une comparaison de nos méthodes de classification hiérarchique d'images à la méthode de (Marszalek, Schmid, 2007) est aussi illustrée dans la figure 9. Notre méthode *AFSCH* affiche une meilleure précision moyenne avec un gain $+8,99$ par rapport à la méthode *AVCH*, et de $+10,67$ % en comparaison avec la méthode de (Marszalek, Schmid, 2007). La précision moyenne de nos méthodes de classification hiérarchique était de $60,6$ % pour *AFSCH* et de $51,61$ % pour *AVCH*, alors que celle de la classification plate reste à $33,8$ %. La précision moyenne de la méthode de référence était à $56,79$ %, et celle de (Marszalek, Schmid, 2007) était à $49,84$ %. On peut donc conclure qu'il y a une nette amélioration des performances avec l'utilisation de la hiérarchie proposée et de nos méthodes de classification hiérarchique.

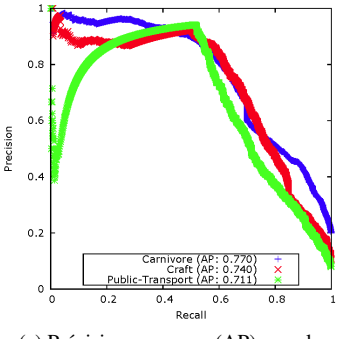
La figure 10 illustre la précision moyenne des différents concepts aux niveaux intermédiaires de la hiérarchie. On peut remarquer que la précision moyenne diminue à mesure qu'on descend les niveaux de la hiérarchie. Ceci s'explique par le fait que les concepts situés dans les niveaux les plus élevés de la hiérarchie sont suffisamment différents visuellement, et donc il est plus facile de trouver une frontière qui les sépare. Ces concepts (classes) peuvent aussi être considérés comme équilibrés. Par exemple, le ratio d'échantillons positifs/négatifs dans les données de VOC'2010 est d'environ 5 %. Notre méthode de classification hiérarchique (*un-contre-nœuds-opposés*) permet de surmonter ce problème puisqu'elle décompose la procédure d'apprentissage en plusieurs sous-tâches. Le ratio d'échantillons positifs/négatifs est de $35,6$ % pour notre méthode, *i.e.* les classes sont assez équilibrées et il n'est pas nécessaire d'utiliser des techniques tels que le suréchantillonnage ou le sous-échantillonnage pour régler le problème des données non équilibrées. Ce résultat démontre aussi l'importance de notre méthode de construction de hiérarchie par rapport aux méthodes de construction de hiérarchies inférées à partir de connaissances textuelles. En effet, la figure 2 illustre une hiérarchie inférée à partir de connaissances textuelles, *i.e.* construite par l'extraction du graphe pertinent dans WordNet reliant les 20 concepts de VOC'2010. La profondeur de cette hiérarchie textuelle est de 17 niveaux. Comme le prouvent les résultats illustrés dans la figure 10, on peut s'attendre à une dégradation importante des résultats de la classification hiérarchique à mesure qu'on avance en profondeur dans cette hiérarchie. Par conséquent, les résultats pour les nœuds (concepts) feuilles seront moins pertinents en termes de précision moyenne par rapport à la méthode que l'on propose.

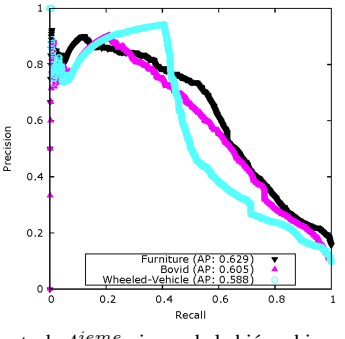


(a) AP pour les concepts du 2^{ieme} niveau de la hiérarchie.



(b) AP pour les concepts du 3^{ieme} niveau de la hiérarchie.





(c) Précision moyenne (AP) pour les concepts du 4^{ieme} niveau de la hiérarchie.

Figure 10. Courbes Rappel/Précision pour les concepts de chaque niveau de la hiérarchie

Tableau 1. Complexité des méthodes proposées pour la classification hiérarchique d'images en comparaison avec les méthodes basées sur les GDAO et sur les ADBH. t, t' : représentent 1 unité de temps, t : pour l'apprentissage d'un classifieur, et t' : le temps de réponse d'un classifieur

	Apprentissage	Étiquetage	Apprentissage sur VOC'10	Étiquetage sur VOC'10
GDAO	$(N^2 - N)/2$	$N - 1$	190 t	19 t'
ADBH	$2N - 1$	$\log_2 N$	39 t	5 t'
AVCH	$\leq 2N - 1$	$\leq \log_2 N$	32 t	4 t'
AFSCH	$\leq 2N - 1$	$\leq 2N - 1$	32 t	32 t'

Finalement, nous illustrons dans le tableau 1 la complexité de nos méthodes de classification hiérarchique basées sur la hiérarchie sémantique construite - cf. figure 6. Nos méthodes de classification hiérarchique présentent une meilleure complexité en temps de calcul pour l'apprentissage des classifieurs par comparaison avec les approches basées sur les arbres de décision binaires hiérarchiques et les approches basées sur les graphes de décision acycliques orientés. En termes de temps nécessaire pour l'annotation (étiquetage) d'une nouvelle image, notre méthode de classification hiérarchique «*approche par fusion des scores des classifieurs hiérarchiques*» (AFSCH) présente une complexité supérieure par rapport aux autres méthodes, mais offre en contrepartie une meilleure précision moyenne que toutes ces méthodes. Notre méthode «*approche par vote des classifieurs hiérarchiques*» (AVCH) est par contre plus efficace que toutes les autres en termes de complexité pour l'étiquetage des nouvelles images.

7. Conclusion

Cet article présente une nouvelle approche pour construire automatiquement des hiérarchies adaptées à l'annotation sémantique d'images. Notre approche est basée sur une nouvelle mesure de similarité sémantique qui prend en compte la similarité visuelle, conceptuelle et contextuelle. Cette mesure permet d'estimer une similarité sémantique entre concepts adaptée à la problématique de l'annotation. Un ensemble de règles est proposé pour ensuite effectivement relier les concepts entre eux selon la précédente mesure et leur ancêtre commun le plus proche dans WordNet. Ces concepts sont ensuite structurés en hiérarchie. Nos expériences ont montré que notre méthode fournit une bonne mesure pour estimer la similarité des concepts, qui peut aussi être utilisée pour la classification d'images et/ou pour raisonner sur le contenu d'images. Nos recherches futures porteront sur la construction automatique d'ontologies adaptées à l'annotation d'images. En effet, nous pensons que la hiérarchie sémantique construite dans le cadre de cet article peut servir comme structure de base pour une ontologie multimédia, et où nous nous proposons d'intégrer l'information spatiale et contextuelle pour construire une ontologie riche et expressive, permettant de raisonner sur le contenu et les annotations des images.

Bibliographie

Banerjee S., Pedersen T. (2003). Extended gloss overlaps as a measure of semantic relatedness. In *International Joint Conference on Artificial Intelligence (IJCAI'03).*

Bannour H. (2009). Une approche sémantique basée sur l'apprentissage pour la recherche d'image par contenu. In *COnférence en Recherche d'Infomations et Applications (CORIA'09)*, p. 471-478.

Bannour H., Hudelot C. (2011). Towards ontologies for image interpretation and annotation. In *Content-Based Multimedia Indexing (CBMI'11)*, p. 211 -216.

Bannour H., Hudelot C. (2012a). Building semantic hierarchies faithful to image semantics. In *advances in Multimedia Modeling (MMM'12)*, vol. 7131, p. 4–15. Springer.

Bannour H., Hudelot C. (2012b). Combinaison d'information visuelle, conceptuelle, et contextuelle pour la construction automatique de hiérarchies sémantiques adaptées à l'annotation d'images. In *actes de la conférence Reconnaissance des Formes et Intelligence Artificielle (RFIA'12)*, p. 462–469. Lyon, France.

Bannour H., Hudelot C. (2012c). Hierarchical image annotation using semantic hierarchies. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM'12)*, p. 2431-2434.

Barnard K., Duygulu P., Forsyth D., Freitas N. de, Blei D. M., Jordan M. I. (2003). Matching words and pictures. *Journal of Machine Learning Research*, vol. 3, p. 1107–1135.

Bart E., Porteous I., Perona P., Welling M. (2008). Unsupervised learning of visual taxonomies. In *Computer Vision and Pattern Recognition (CVPR'08)*.

Blei D. M., Griffiths T. L., Jordan M. I., Tenenbaum J. B. (2004). Hierarchical topic models and the nested chinese restaurant process. In *Neural Information Processing Systems (NIPS'04)*.

Budanitsky A., Hirst G. (2006). Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics*, vol. 32, p. 13–47.

Carneiro G., Chan A. B., Moreno P. J., Vasconcelos N. (2007). Supervised learning of semantic classes for image annotation and retrieval. *IEEE Transaction Pattern Analysis and Machine Intelligence*, vol. 29, p. 394–410.

Cevikalp H. (2010). New clustering algorithms for the support vector machine based hierarchical classification. *Pattern Recognition Letters*, vol. 31, n° 11, p. 1285 - 1291.

Church K. W., Hanks P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, vol. 16, p. 22–29.

Cortes C., Vapnik V. (1995). Support-vector networks. *Machine Learning*, vol. 20.

Deng J., Berg A. C., Li K., Fei-Fei L. (2010). What does classifying more than 10,000 image categories tell us? In *European conference on computer vision (eccv'10)*.

Deng J., Dong W., Socher R., Li L.-J., Li K., Fei-Fei L. (2009). Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition (CVPR'09)*.

Deselaers T., Ferrari V. (2011). Visual and semantic similarity in imagenet. In *Computer Vision and Pattern Recognition (CVPR'11)*, p. 1777 -1784.

Fan J., Gao Y., Luo H. (2007). Hierarchical classification for automatic image annotation. In *Conference on research and development in information retrieval (SIGIR'07)*, p. 111–118.

Fan J., Gao Y., Luo H. (2008). Integrating concept ontology and multitask learning to achieve more effective classifier training for multilevel image annotation. *IEEE Transaction on Image Processing*, vol. 17, n° 3.

Fan J., Luo H., Shen Y., Yang C. (2009). Integrating visual and semantic contexts for topic network generation and word sense disambiguation. In *ACM international Conference on Image and Video Retrieval (CIVR'09)*.

Fellbaum C. (1998). *WordNet: An electronic lexical database*. Cambridge, MA, MIT Press.

Gao T., Koller D. (2011). Discriminative learning of relaxed hierarchy for large-scale visual recognition. In *International Conference on Computer Vision (ICCV'11)*, p. 2072-2079.

Griffin G., Perona P. (2008). Learning and using taxonomies for fast visual categorization. In *Computer Vision and Pattern Recognition (CVPR'08).*

Hauptmann A., Yan R., Lin W.-H. (2007). How many high-level concepts will fill the semantic gap in news video retrieval? In *ACM international Conference on Image and Video Retrieval (CIVR'07)*, p. 627–634.

Lavrenko V., Manmatha R., Jeon J. (2003). A model for learning the semantics of pictures. In *Neural Information Processing Systems (NIPS'03).*

Li F.-F., Perona P. (2005). A bayesian hierarchical model for learning natural scene categories. In *Computer Vision and Pattern Recognition (CVPR'05)*, vol. 2, p. 524–531. Washington, DC, USA.

Li L.-J., Wang C., Lim Y., Blei D. M., Li F.-F. (2010). Building and using a semantivisual image hierarchy. In *Computer Vision and Pattern Recognition (CVPR'10).*

Liu Y., Zhang D., Lu G., Ma W.-Y. (2007). A survey of content-based image retrieval with high-level semantics. *Pattern Recognition*, vol. 40, n° 1, p. 262–282.

Lowe D. G. (1999). Object recognition from local scale-invariant features. In *International Conference on Computer Vision (ICCV'99).*

Marszalek M., Schmid C. (2007). Semantic hierarchies for visual object recognition. In *Computer Vision and Pattern Recognition (CVPR'07)*, p. 1-7.

Marszalek M., Schmid C. (2008). Constructing category hierarchies for visual recognition. In *European Conference on Computer Vision (ECCV'08)*, p. 479–491.

Naphade M., Smith J. R., Tesic J., Chang S.-F., Hsu W., Kennedy L. (2006). Large-scale concept ontology for multimedia. *IEEE MultiMedia*, vol. 13, p. 86–91.

Patwardhan S., Pedersen T. (2006, April). Using wordnet-based context vectors to estimate the semantic relatedness of concepts. In *Proceedings of the EACL 2006 Workshop on Making Sense of Sense: Bringing Computational Linguistics and Psycholinguistics Together*, p. 1–8.

Platt J. C., Cristianini N., Shawe-taylor J. (2000). Large margin dag for multiclass classification. In *Advances in Neural Information Processing Systems (NIPS'00).*

Resnik P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In *International Joint Conferences on Artificial Intelligence (IJCAI'95).*

Romdhane L. B., Bannour H., Ayeb B. el. (2010). Imiol: a system for indexing images by their semantic content based on possibilistic fuzzy clustering and adaptive resonance theory neural networks learning. *Applied Artificial Intelligence*, vol. 24, n° 9, p. 821-846.

Russell B. C., Torralba A., Murphy K. P., Freeman W. T. (2008). LabelMe: A Database and Web-Based Tool for Image Annotation. *International Journal of Computer Vision*, vol. 77, n° 1-3, p. 157–173.

Sivic J., Russell B. C., Zisserman A., Freeman W. T., Efros A. A. (2008). Unsupervised discovery of visual object class hierarchies. In *Computer Vision and Pattern Recognition (CVPR'08).*

Smeulders A. W. M., Worring M., Santini S., Gupta A., Jain R. (2000). Content-based image retrieval at the end of the early years. *IEEE Transaction Pattern Analysis and Machine Intelligence*, vol. 22, p. 1349–1380.

Tousch A.-M., Herbin S., Audibert J.-Y. (2012). Semantic hierarchies for image annotation: A survey. *Pattern Recognition*, vol. 45, p. 333–345.

Wei X.-Y., Ngo C.-W. (2007). Ontology-enriched semantic space for video search. In *ACM Multimedia (MM'07)*, p. 981–990.

Wu L., Hua X.-S., Yu N., Ma W.-Y., Li S. (2008). Flickr distance. In *ACM Multimedia (MM'08)*, p. 31–40.

Yao B., Yang X., Lin L., Lee M. W., Zhu S. C. (2009). I2t: Image parsing to text description. In *Proceedings of IEEE.*