Multimedia Tools and Applications (2021) 80:6823–6851
https://doi.org/10.1007/s11042-020-09864-9

OntoAnnClass: ontology-based image annotation driven by classification using HMAX features

Jalila Filali[1] · **Hajer Baazaoui Zghal**[1,2] · **Jean Martinet**[3]

Received: 20 June 2019 / Revised: 22 April 2020 / Accepted: 9 September 2020 /
Published online: 22 October 2020

Abstract

Several approaches have been proposed in the area of Automatic Image Annotation (AIA) in order to exploit the relationships between words that are extracted from image categories, and to automatically generate annotation words for a given image. Other methods exploit ontologies, where the annotation keywords were derived from ontology to improve image annotation. In this paper, we propose an ontology-based image annotation driven by classification using HMAX features. The idea is (1) to train visual-feature-classifiers and to build an ontology that can finely represent the semantic information associated with training images, and (2) to combine classifier outputs and ontology for image annotation. To annotate images, we define a membership value of words in images. In particular, we propose to evaluate the membership value based on the confidence value of classifiers and the semantic similarity between words. The membership value depends on the word relationships found in the ontology that serve to select annotation words. The obtained experimental results show that the exploitation of both classifier outputs and ontology by evaluating our proposed membership value enables an improvement of image annotation.

Keywords Image annotation · Ontologies · Classification · HMAX features · BoVW model

✉ Jalila Filali
jalila.filali@ensi-uma.tn

Hajer Baazaoui Zghal
hajer.baazaoui@ensea.fr

Jean Martinet
jean.martinet@univ-cotedazur.fr

[1] ENSI, RIADI Laboratory, University of Manouba, Manouba, Tunisia

[2] ETIS UMR 8051, CY University, ENSEA, CNRS, F-59000, Cergy, France

[3] Polytech Nice Sophia Campus SophiaTech, Université Côte d'Azur/I3S/CNRS, Sophia-Antipolis, France

1 Introduction

Automatic Image Annotation (AIA) is one of the most fundamental problems in image retrieval and computer vision. The aim of the AIA is to assign suitable annotation words to any given image, which reflects its content. In general, AIA consists in learning models from a training set of pre-annotated images in order to generate annotation words for unlabeled images. Therefore, due to the link between the visual features and the annotation words, AIA becomes a difficult issue in computer vision. In this context, machine learning approaches are used to learn the mapping between low-level and semantic features, and then generate annotation words for a test image. These include classification approaches that allow classifying images in semantic classes based on their visual features.

In addition, in the literature, several works dealing with image classification and annotation revolve around BoVW method, which consists of building a visual vocabulary from image features [31], [8]. The image features are quantified as visual words to express the image content through the distribution of visual words.

Recently, special attention has been shifted to the use of complex architectures which are characterized by multi-layers. Indeed, the biologically-inspired HMAX model was firstly proposed by [21]. The HMAX model has attracted a great deal of attention in image classification due to its architecture which alternates layers of feature extraction with layers of maximum pooling. The HMAX model was optimized in the work of [24] in order to add multi-scale representation as well as more complex visual features.

In order to achieve a finer representation of the semantic content in images, several annotation approaches based on ontologies have been proposed. The use of ontologies is generally motivated by the need to use semantic relations and describe data at a more semantic level for better annotation. However, such methods do not exploit both visual and semantic features during the image annotation process.

In this paper, we propose an ontology-based image annotation driven by classification using HMAX features. Our idea is to train the classifiers with visual features and to build an ontology that can finely represent the semantic content associated with the training images. Both classifiers and ontology are used for annotating testing images.

Thus, the main contributions consist; 1) in integrating the classifiers and ontology in the training phase; and 2) in evaluating a membership value that serves to select annotation words depending mostly on relationships which are detected in the ontology.

The remainder of this paper is organized as follows. Section 2 presents an overview of the related research, along with our motivations and objectives. Section 3 describes the proposed image annotation approach and its components. In Section 4, we report the experimental results of our approach. Finally, Section 5 concludes this paper and proposes directions for future works.

2 Related work, motivation and objectives

In image retrieval field, two basic image retrieval approaches have been proposed in the literature: 1) content based image retrieval (CBIR) and 2) semantic image indexing and retrieval (SIIR). In this context, most works turn their focus on content based image retrieval that can be considered a principal helps to organize images by their visual content. However, it was shown that CBIR approaches are unable to automatically describe the semantic content of images. As a result, Automatic Image Annotation (AIA) has acquired more attention of researchers in computer vision and multimedia areas. In AIA area, several methods

and approaches have been introduced and applied. In the next subsection, we give a general overview of the main related works.

2.1 Related work

2.1.1 Approaches based on learning techniques

In the AIA area,a large amount of methods based on learning techniques has been applied [10], [19] and [14]. Recently, to annotate images, some researchers have attempted to learn detectors that can localize objects in images. In this context, [10] proposed a weakly supervised part selection method with spatial constraints for fine-grained image classification. The goal of the work is, firstly, to learn a whole-object detector automatically aiming at localizing the object through jointly using saliency extraction and segmentation; secondly, to propose spatial constraints that serve to select the distinguished parts. The spatial constraints define the relationship between an object and its parts and the relationships between the object's parts. The aim is to ensure that the selected parts are located in the object region and are the most distinguishing parts from other categories. The results of this work demonstrate the superiority of this method compared with the methods that used expensive annotations.

In addition, in [36] a fast binary-based HMAX model (B-HMAX) is proposed for object recognition. The goal is to detect corner-based interest points and to extract few features with better distinctiveness. The idea is to use binary strings to describe the image patches extracted around detected corners, and then to use the Hamming distance for matching between two patches.

Moreover, several image annotation approaches based on deep learning models have been proposed. For instance, in [17], two main issues in large-scale image annotation are addressed: 1) how to learn a rich feature representation suitable for predicting a diverse set of visual concepts ranging from object, scene to abstract concept; 2) how to annotate an image with the optimal number of class labels. For the first issue, a novel multi-scale deep model has been proposed, the aim is to extract rich and discriminative features capable of representing a wide range of visual concepts. The deep model is also made multi-modal by taking noisy user-provided tags as model input to complement the image input. For tackling the second issue, a label quantity prediction auxiliary task has been introduced to explicitly estimate the optimal label number for a given image. In this work, extensive experiments are carried out on two large-scale image annotation benchmark datasets and the results show that this method significantly outperforms the state-of-the-art.

In [26], a multi-modal deep learning framework has been introduced, the aim is to optimally integrate multiple deep neural networks pretrained with convolutional neural networks. In particular, the proposed framework explores a unified two-stage learning scheme that consists of learning to fine-tune the parameters of deep neural network with respect to each individual modality, and learning to find the optimal combination of diverse modalities simultaneously in coherent process. The result of this work validate the effectiveness of the proposed framework.

In addition, AIA methods are considered as a kind of efficient schemes to solve the problem of semantic-gap between the original images and their semantic information. In this context, to address this problem, [16] combined the CNN feature of an image into their proposed model which is based on a CNN model-AlexNet. The idea is to extract a CNN feature by removing its final layer. Also, based on the experience of the traditional KNN models, they proposed a model to address the problem of simultaneously addressing the image tag refinement and assignment while maintaining the simplicity of the KNN

model. The proposed model divides the images which have similar features into a semantic neighbor group. Moreover, using a self-defined Bayesian-based model, [16] distributed the tags which belong to the neighbor group to the test images according to the distance between the test image and the neighbors. The experiments of this work show the effectiveness of the proposed model.

2.1.2 Approaches based on ontologies

To improve image annotation ontological techniques have been used for AIA [35],[22] and [30]. For instance, in [22], a complete framework to annotate and categorize images has been proposed. This approach is based on multimedia ontologies organized following a formal model to represent knowledge. In this work, ontologies use multimedia data and linguistic properties to bridge the gap between the target semantic classes and the available low-level multimedia descriptors. The multimedia features are automatically extracted using algorithms based on MPEG-7 standard. The informative image content is annotated with semantic information extracted from the ontologies and the categories are dynamically built by means of a general knowledge base. Experimental results of this work show the efficiency of this method in the annotation and classification tasks using a combination of textual and visual components.

Moreover, in [20], an ontology based supervised learning for multi-label image annotation approach has been proposed, where classifiers' training is conducted using easily gathered web data. This work takes advantage of both low-level visual features and high-level semantic information of given images. The goal is to use ontologies at several phases of supervised learning from large scale noisy training data. Experimental results show the effectiveness of the proposed framework over existing methods.

In [1], an approach based on semantic hierarchies has also been proposed for hierarchical image classification. The goal is to decompose the annotation problem in several independent classification tasks using two methods for computing a hierarchical decision function that serves to annotate images.

In [18] an approach for automatic image annotation has been proposed in order to automatically and efficiently assign linguistic concepts to visual data such as digital images based on both numeric and semantic features. The goal of this approach is to compute a multi-layered active contour and to extract visual features within the regions segmented by these active contours in order to map them into semantic notions. The method relies on decision trees trained using these attributes, and the image is semantically annotated using the resulting decision rules.

Other recent works tackle how coarse and fine labels can be used to improve image classification. In this context, [4] address the problem of classification of coarse and fine grained categories by exploiting semantic relationships. In this work, the idea is to adjust the probabilities of classification according to the semantics of the classes or categories. An algorithm for doing such an adjustment is proposed to show the improvement for both coarse and fine grained classification.

In [13], a weakly supervised image classification method with coarse and fine labels has been proposed. In this work, they investigated the problem of learning image classification when a subset of the training data is annotated with fine labels, while the rest is annotated with coarse labels. The goal is to use weakly labeled data aiming at learning a classifier to predict the fine labels during testing. To this end, they proposed a CNN- based approach to address this problem, where the commonalities between fine classes in the same coarse class are captured by min-pooling in the CNN architecture. The experimental results of this work show that this method significantly outperforms the work that addresses the same problem.

In addition, [23] addressed the problem of learning subcategory classifiers when only a fraction of the training data is labeled with fine labels while the rest only has labels of coarser categories. In particular, the aim is to adopt the framework of Random Forests [2] and to propose a regularized objective function that takes into account relations between categories and subcategories. The results show that the additional training data with the category-only labels improve the classification of sub-categories.

More closely related is the work of [9]. They proposed a joint framework for describing an image by the proposed context. This approach is based on integrating the multi-layer semantic elements detection and ROI (Region of Interest) identification into one optimization process. The idea is to combine a multi-label regression for hierarchical concept detection and a multi-class SVM for ROI identification in order to better describe the testing images. The experimental results demonstrate the effectiveness of the framework and the output descriptions improve the performance of image retrieval.

To summarize the recent related work, we present in Table 1 a review of the related approaches.

Table 1 Overview of the related image annotation approaches

Approaches	Model	Ontology	Dataset
[6]	HMAX	✓	ImageNet
[17]	CNN	—	NUS-WIDE MSCOCO Corel5k
[16]	CNN	—	EspGame
[4]	CNN	✓	CIFAR-100 CIFAR100
[13]	CNN	✓	ILSVRC2010
[10]	CNN	—	CUB-200-2011
[37]	MVML	—	LabelMe VOC2012
[25]	SVM	—	NUS-WIDE
[29]	SVM	—	dataset
[14]	CNN	—	Image CLEF 2015
[15]	HMAX	–	Caltech 101
[26]	CNN	—	NUS-WIDE IAPRTC-12 Pascal
[9]	SVM	✓	Yahoo
[36]	HMAX	–	GRAZ01
[23]	NCM	—	ILSVRC 2010
[35]	SVM	✓	Image CLEF 2015
[22]	MPEG-7	✓	Caltech 256
[20]	SVM	✓	Image CLEF 2014
[18]	MFVF	—	Corel dataset
[32]	SVM	✓	Berkeey Image Vitterbi USC-SIPI
[1]	SVM	✓	Pascal VOC 2010
[30]	SVM	✓	Web Data

2.2 Motivation and objectives

An image classification and annotation approaches based on visual features and ontologies were proposed in previous works [5, 8] and [7]. However, an improvement of image annotation precision is needed.

In this paper, we propose a novel image annotation method driven by classification and based on HMAX features and ontology.

Our motivation is to exploit both visual and ontological semantic features to improve image annotation.

In particular, we propose an ontology-based image annotation driven by classification using HMAX features. Our method is inspired by the approaches presented above.

Our objective is two-fold, we aim at:

- (1) Training visual-feature-classifiers and building an ontology from image labels that can finely represent the semantic content associated with the training images;
- (2) Exploiting classifier outputs and ontology for image annotation. For this purpose, we need to define a membership value based on both classifiers' confidence value and semantic similarity of words depending on relationships detected in the ontology.

The originality of our proposal lies in the integration of classifiers with ontology that cover the semantic content of images in order to improve image annotation.

3 The proposed image annotation approach

In this section, we describe the architecture of the proposed image annotation approach and detail the different phases and their components. The proposed image annotation approach is composed of two main phases: (1) training phase and (2) image annotation phase. The different components are detailed below.

3.1 Training phase

The training phase includes three components, namely: feature extraction component, classifiers training component, and word extraction and ontology building component.

Firstly, visual features are extracted from the training set (Fig. 1: feature extraction). Our approach uses HMAX features [11, 27, 28], [12], because they are generic, do not require hand-tuning, and can represent well complex features (a detailed description is given below). Secondly, HMAX features are used to train the classifiers. We selected a multi-class linear SVM in order to classify images (Fig. 1: classifiers training).

Finally,image labels from the train set are used to extract words and to build the ontology as a final step, which consists in establishing relationships between words using taxonomic relationships found in WordNet (Fig. 1: word extraction and ontology building).

3.1.1 Feature extraction component

To extract visual features from training images, we used HMAX model; in particular, we adopted the HMAX model to provide complex and invariant visual information and to improve the discrimination of features. The HMAX model follows a general 4 layer architecture. Below we describe the operations of each layer. Simple ("S") layers apply local

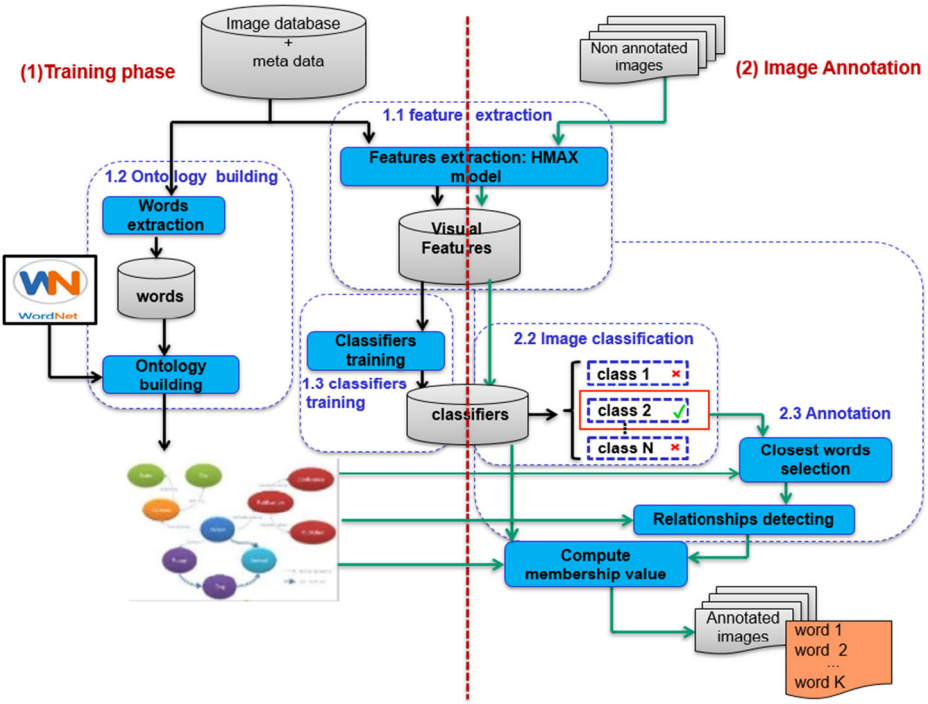


Fig. 1 Architecture and components of the proposed image annotation approach

filters that compute higher-order features and complex ("C") layers increase invariance by pooling units.

– **Layer 1 (S1 Layer):** In this layer, each feature map is obtained by convolution of the input image with a set of Gabor filters $\mathbf{g_{s,o}}$ with orientations **o** and scales **s**. In particular S1 Layer, at orientation **o** and scale **s**, is obtained by the absolute value of the convolution product given an image I:

$$L1_{s,o} = |g_{s,o} * I| \tag{1}$$

– **Layer 2 (C1 Layer):** The C1 layer consists in selecting the local maximum value of each S1 orientation over two adjacent scales. In particular, this layer divides each $\mathbf{L1_{s,o}}$ features into small neighborhoods $\mathbf{U_{i,j}}$, and then selects the maximum value inside each $\mathbf{U_{i,j}}$.

$$L2_{s,o} = \max_{U_{i,i} \in L1_{s,o}} * U_{i,j} \tag{2}$$

– **Layer 3 (S2 Layer):** S2 layer is obtained by convolving filters α^m, which combine low-level Gabor filters of multiple orientations at a given scale.

$$L3_{s,m} = \alpha_m * L2_s \tag{3}$$

– **Layer 4 (C2 Layer):** In this layer, L4 features are computed by selecting the maximum output of $L3_s^m$ across all positions and scales.

$$L4 = \max_{(x,y),s} L3_S^1(x, y), ..., \max_{(x,y),s} L3_S^M \tag{4}$$

The obtained layer 4 vectors define the HMAX features that are the input of the next component.

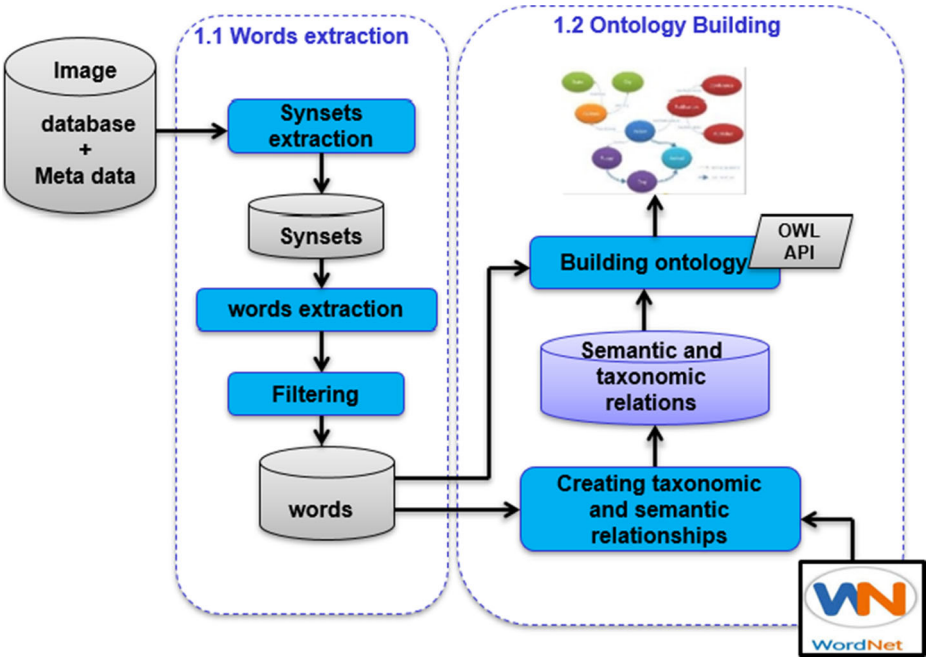


Fig. 2 Word extraction and ontology building components

3.1.2 Classifiers training component

SVMs are mainly designed for the discrimination of two classes. Also, they can be adapted to multi-class problems where a multi-class SVM classifier can be obtained by training several classifiers. In our work, the aim is to learn a discriminative model for each "*class*" in order to predict the visual features membership. To achieve this goal, we focus on linear SVM classifiers since the diversity of image categories makes using nonlinear models is impractical.

In particular, given the visual features (HMAX features) of the training images, we train a One-vs-All SVM classifier [3] for each class to discriminate between this class and the other classes.

3.1.3 Word extraction and ontology building component

As depicted in Fig. 2, the ontology building component consists of two main steps, namely, word extraction and ontology building.

Word extraction Let us consider that the meta-data of the train images consists of the synset IDs (synsets are called a "*synonym set*" or "*synset*" or concepts and described by one or multiple words) which are defined by the WordNet.[1] As depicted in Fig. 2, word extraction component consists firstly, in determining synsets which are associated with the train images; secondly, in extracting words from the obtained synsets. Finally, we need to

[1] https://wordnet.princeton.edu/

filter all words after extracting them from the obtained synsets. The filtering of words is performed by removing the words that are not defined by the WordNet dictionary.

Ontology building Let us consider an image database DB consisting of a set of pairs (images, synsets) and each synset is composed of words where:

- I=$i_1, i_2, ..., i_L$ is the set of all images in DB,
- L is the number of images in the DB,
- S=$S_1, S_2, ..., S_M$ is the synsets which are associated with the train images in DB,
- W=$w_1, w_2, ..., w_N$ is the words that are extracted from S
- N is the size of the words set,
- M is the number of the synsets associated with the train images in DB,
- LD is a lexical database of nouns, verbs, adjectives and adverbs which are grouped into sets of cognitive synonyms (synsets). Synsets are interlinked by means of semantic and lexical relations.

Given the previous parameters, the aim is to build an ontology, consisting of a set of words W dedicated to this specific annotation problem and depending on the annotation vocabulary.

At the beginning, we define a set of main symbols which are necessary for the definition of our ontology.

Definition 1 Ontology

We define an ontology, denoted in the sequel θ, by words W and relations R among words. The ontology θ relies mainly on the two following concepts: "*Thing*" represents the top concept of the ontology and "*Word*" represents the word in our ontology, is any word from the annotation vocabulary used to describe the content of images.

Formally, θ is a triplet defined as follows:

$$\theta = \left\{Root, W, R_{TS}(W_i, W_j)\right\} \tag{5}$$

where:

- $Root =$ "$Thing$" is the top concept of the ontology;
- R_{TS} represent the relationships among words $W_i W_j$;
- $W_i W_j \in W$ and $i, j = 1, ..., N$ with $i \neq j$.

To extract relationships between words, we used LD. We are interested in relationships that are detailed in Table 2.

The type of relationships can be classified into hyponymy or specialization relationships generally known as *kindOf* or *isA*, hypernymy or generalization relationships known as "hasKind", partitive or meronymy relationships called *partOf* which describe words that are parts of other words and holonymy relationships known as *hasPart* which defines the whole-to-part relationships, and synonym relationships.

The ontology is defined by considering only the words extracted from the image database itself. The resulting sets of taxonomic and semantic relationships as well as the resulting set of words are the basis of our ontology.

Once the taxonomic and semantic relationship extraction is carried out, ontology building is then performed. To construct successfully the ontology, we used the OWL API.[2] Some rules are applied in order to transform the extracted relationships in OWL language.

[2] http://owlapi.sourceforge.net/

Table 2 Words relationships used in our ontology

Relations	Definition	Meaning
isA	$R_{isA}(W_i, W_j)$	The word w_i is a hyponym word of w_j.
HasKind	$R_{hasKind}(W_i, W_j)$	The word w_i is a hypernym word of the word w_j.
PartOf	$R_{partOf}(W_i, W_j)$	Meronymy relationship where w_i is a part of w_j.
HasPart	$R_{hasPart}(W_i, W_j)$	Holonymy relationship where w_i is a holonym of w_j.
Synonym	$R_{synonym}(W_i, W_j)$	The words w_i and w_j have the same meaning.

3.2 Image annotation phase

The image annotation phase includes three main components which are: feature extraction, image classification and image annotation (Fig. 1: Image annotation). Firstly, features of the testing images are extracted (Fig. 1: feature extraction). Secondly, the image is classified (Fig. 1: image classification). Thirdly, a membership value is computed using both the outputs of classifiers and ontology. In particular, the membership value is computed using the confidence value of the classes and the semantic similarity between the ontology. The membership value depends mostly on relationships found in the ontology. Annotation words are ranked according to their membership values, in order to assign a set of annotation words to the query image (Fig. 1: image annotation). In the following subsections, we describe the formalization of our annotation problem and we detail the image annotation phase.

3.2.1 Problem formalization

We work in multi-class classification images, so for each "*word*" in our ontology a related classifier is trained. We consider θ the ontology which is built and N is the number of classifiers associated to the "*words*" in θ . Let us present the following definitions to explain the image annotation problem:

- w_t is the obtained class or "*word*" from the best classifier of the test image I.
- C_{w_i} is the classifier of the word w_i with $w_i \in \theta$.
- A_I is the annotation words for image I , consisting of the words $\{w_j \in W, j = 1, ..., K\}$ that will be assigned to the test image I .

Given the θ ontology, w_t and C_{w_i} classifiers, the aim is to assign K annotation words to the test image where $A_I = \{w_1, w_2, ..., w_K\}$, the assignment of K annotation words depends on the membership value between the test image and the related words in θ.

3.2.2 Ontology-driven image annotation using classification

To annotate images, we focus on proposing a membership value of the closest words to the test image. To this end, firstly we propose to assign to each closest word a semantic weight. The semantic weight depend on the neighborhood degree of the closest word to the target word w_t , and the semantic similarity of the pair (w_t , closest word). Secondly, according to the relationships related to w_t, the confidence values of the closest words and their semantic weights are used for computing the membership values.

Let us present the following definitions to explain the functions that we used for computing the membership degree of the related words in the ontology to the test image according to the ontological relationships:

- $R_{TS}(w_i, w_j)$= ($R_{isA}(w_i, w_j)$, $R_{hasKind}(w_i, w_j)$, $R_{partOf}(w_i, w_j)$, $R_{hasPart}(w_i, w_j)$, $R_{synonym}(w_i, w_j)$) represents the set of relation types existing in θ with w_i and $w_j \in W$;
- ClosestWords(w_t, L_{Max})=Clw= $\{w_1, .., w_M\}$ is a function allowing to find the closest words of w_t in θ according to a length L_{Max}, with M as the number of the closest words and L_{Max} as the maximum path length between w_t and the words returned by the ClosestWords(w_t, L_{Max});
- getWordsDirectLink(w_t, θ) = $\beta = \{\beta_1, ..., \beta_b\}$ is a function allowing to return the words that have a direct link with the target word w_t;
- CV(w_i) is the confidence value of the word w_i that is obtained by its own classifier C_{w_i};
- L(w_t, w_i) is a function that returns the shortest path length between w_t and w_i;
- SW(w_i) is the semantic weight of w_i in the closest semantic space of w_t with $w_i \in Clw$;
- $MV_{R_{TS}}(I, w_i)$ is the membership function to compute the membership degree of word w_i to the test image I according to the relationship $R_{TS}(w_t, w_i)$;

In our image annotation method, to compute the membership value, we are interested in the set of the closest words of w_t that are returned by the ClosestWords(w_t, L_{Max}) function. For this purpose, we propose to assign a semantic weight to each closest word. The semantic weight depends, firstly, on the neighborhood degree of the closest word w_i to w_t, secondly, on the semantic similarity of the (w_t, w_i) pair. We assign the maximum semantic weight to w_t so SW (w_t)=1.

Thus, we define the following function to compute the semantic weight of each closest word w_i:

$$SW(w_i) = Nd_L(w_t, w_i)(w_i) * Sim(w_t, w_i) \tag{6}$$

Where:

- $Nd(w_i)$ is the neighborhood degree of w_i to w_t according to the length between w_t and w_i : in our case, we define the $Nd(w_i)$ as follows:

$$Nd(w_i) = \frac{1}{L(w_t, w_i)} \tag{7}$$

  where $L(w_t, w_i)$ is the path length between w_t and w_i.
- Sim(w_t, w_i) is the semantic similarity between w_t and w_i based on the Wu-Palmer metric (WUP) [34]:

$$Sim(w_t, w_i) = \frac{2 \cdot depth(LCS(w_t, w_i))}{depth(w_t) + depth(w_i)} \tag{8}$$

  with LCS is the Lowest Common Subsumer(s).

Subsequently, for any relation type found between w_t and the words that are returned by the getWordsDirectLink(w_t, θ) function, we start from w_t node and then we follow the path according to the relation types until reaching a path length equal to L_{Max}. We are generally interested only in the relationship sense that started from w_t. Thus, for each relation type, we propose a method to compute a membership value of the closest words to the test image I.

In the following, we detail how to compute the proposed membership value for each relation type.

Proposed membership value according to *isA/hasKind* relationships In the case where w_t is linked to a word from the word set β by the *isA* relation type (w_t is a hyponym of the β_j), semantically, we could be sure that w_t is a β_j, since it is clear that w_t is necessarily a β_j.

For example, we could be sure that a "*tree*" is necessarily a "*woody plant*", also a mammal is necessarily an animal. To confirm this certitude, the minimum membership value of the hypernym words of w_t must be equal to the w_t membership value.

Thus, to compute the membership value for this case, we are defined the flowing function:

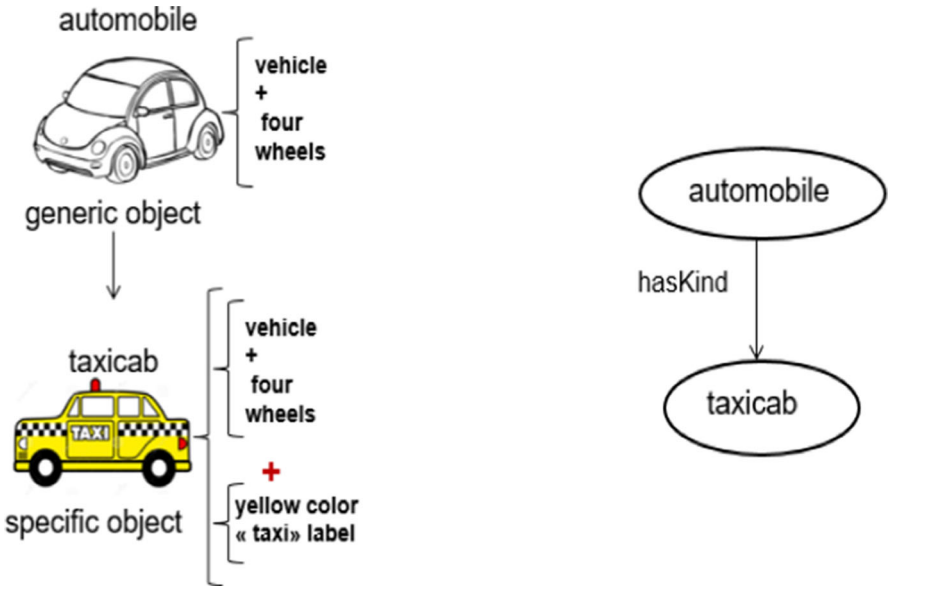
$$MV_{R_{isA}}(I, w_i) = SW(w_t) * CV(w_t) = CV(w_t) \tag{9}$$

Where w_i is a hypernym word of w_t and $w_i \in Clw$, $CV(w_t)$ is the confidence value of the target word w_t and $SW(w_t)$ is the semantic weight of w_t.

The hypernym words of w_t that are included in the closest words set Clw are added to the annotation words.

In case that the w_t is linked to the β_j word by the hasKind relationship type (w_t is a hypernym of the β_j), the hasKind relation conveys the specific information from the upper class (class of w_t) to its children classes (the w_t hyponym word classes). Also, starting from the upper word node (w_t) and following the *hasKind* relationships, this relation type keeps the generic information of the wt class in their hyponym word classes.

For example, as depicted in Fig. 3, if we have w_t = "*automobile*" and it has a hyponym word "*taxicab*", following the *hasKind* relation from the "automobile" to "*taxicab*", a specific information allowing to define "*taxicab*" object, is added to the generic information



(a) Visual representation (b) Semantic representation

Fig. 3 Visual and semantic representation of *hasKind* relationship

which presents the "*automobile*". Thus, the representation of the object ("*taxicab*") can be defined by the generic object representation ("*automobile*") and the specific detailed information of the object ("*taxicab*").

So, the percentage of having a taxicab on an image is the percentage of having an automobile added to the percentage of having a taxicab on the image. To this end, to estimate the membership value of the hyponym words, we merged the confidence values of the hyponym words and thew_t, that are weighted by their semantic weights. The goal is to improve the accuracy of image annotation.

In this case, the final function to compute the membership value is:

$$MV_{R_{hasKind}}(I, w_i) = \frac{CV(w_t) + \sum_{j=1}^{|P|} SW(P_j) * CV(P_j)}{1 + \sum_{j=1}^{|P|} SW(P_j)} \tag{10}$$

Where:

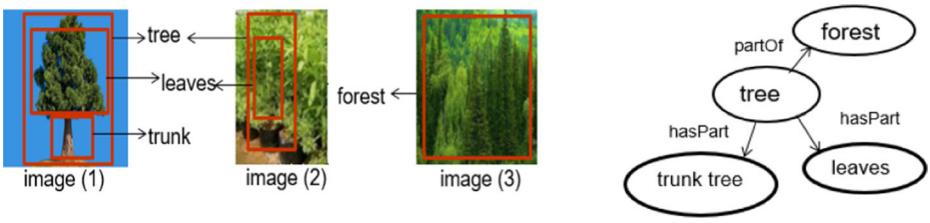
- CV(w_t) is the confidence value of the w_t;
- P is the set of words in the path P_{w_t,w_i} from w_t to w_i following the hyponym relationship of w_t;
- SW(P_j) is the semantic weight of the hyponym word which exists in P.

For this aim, starting from the w_t and following the *hasKind* relationship until reaching the maximum path length L_{Max}, we compute the membership value of the hyponym words according to the previous function.

Proposed membership value according to *hasPart/partOf* relationships In the case that the w_t is linked to the β_j word by the holonymy (w_t is a holonym of the β_j or w_t has partβ_j) / meronymy (w_t is a meronym of theβ_j or w_t is a part of β_j) relationships, it is clear that the w_t class represents the whole information parts of meronym word classes or the parts of the whole information of holonym word classes. Thus, it is obvious that the relationship between the classes represents composition relationships.

For example, as depicted in Fig. 4 (a), let us suppose that we have three test images (1), (2) and (3). We suppose that they are assigned by "*tree*" word as a target word w_t, which is obtained by the best classifier. The three images are composed of a *tree* object. The content of the first image is represented by a *tree* object, also, by both "*trunk*" and "*leaves*" sub-objects with a significant appearance. In the second image, only the *tree* object and the *leaves* sub-object appeared. However the third image is composed of only the forest object.

In the semantic representation, as depicted in Fig. 4 (b), to annotate the test image, we process to follow from the holonym word ("*tree*") to the meronym nodes, and we select the word that has the highest confidence value. But, we could not be sure that the selected



(a) Visual representation (b) Semantic representation

Fig. 4 Visual and semantic representation of *hasPart/partOf* relations

meronym word will appear in the top annotation words sorted according to their confidence value. Also, in the opposite case, walking from the target word "*tree*" to the holonym words "*forest*" and "*wood*", although the word that has the greater confidence value is selected, we could not be sure if the holonym word object appeared in the test image.

Thus, the main problem, in this case, is how to predict if the "*trunk*" object (meronym word) or "*forest*" object (holonym word) belong to the content of a test image or not.

To overcome this problem, we propose to estimate a visual similarity between holonym and meronym words. The aim is to estimate a distance between the word classes. Thus, the visual similarity between the words is inversely proportional to the distance between their visual classes. The function to estimate the visual similarity between the words is:

$$VisSim(w_t, w_i) = \frac{1}{1 + d(C_{w_t}, C_{w_i})} \tag{11}$$

Where:

$d(C_(w_t), C_(w_i))$ is the Euclidean distance between the classifiers of the words w_t and w_i, and w_i is a holonym or meronym word of w_t.

The objective being to combine the visual similarity of words and their semantic weights in order to improve the annotation accuracy. Thus, to compute the membership value, we proposed the following function:

$$MV_{R_{hasPart/partOf}}(I, w_i) = \frac{CV(w_t) + \sum_{j=1}^{|P|} SW(P_j) * visSim(w_t, P_j)}{1 + \sum_{j=1}^{|P|} SW(P_j)} \tag{12}$$

Where:

- w_i is a meronym word of w_t ;
- P is the set of words in the path P_{w_t,w_i} from w_t to w_i following the meronym relationship;

  $visSim(w_t, P_j)$ is the visual similarity of the pair (w_t, P_j);

For the *partOf* relation, we compute the membership value with the same manner as the previous function, but we are interested in the holonym words instead of the meronym words.

Illustrative example of our image annotation method Let us suppose that we have an ontology part, as depicted in Fig. 5, the target word of the test image is "*tree*". In this case, we suppose that L_{Max}= 4 to select the closest word of the w_t. Therefore, each closest word has a confidence value (the value in blue), and a semantic weight value (the value in green) which is obtained according to the function (6). To annotate the test image, initially the annotation words contained only the w_t ="*tree*" as a target word. Then, starting from the target word and following each relationship with w_t until reaching the maximum path length L_{Max}. Then, according to the relation type, membership value of the words is computed using the functions defined previously. After computing the membership values, all closest words are ranked and the top $K(= 10)$ words are assigned to the test image. The algorithm of our image annotation model is presented in Algorithm 1.

Algorithm 1 Summarized algorithm of image annotation.

Input : θ:ontology, C: classifiers of words, I: test image, w_t: target word, L_{Max}: maximum path length traveled in θ

Output: A_I

Initialization: $A_I \leftarrow w_t$
$TemAnWords \leftarrow \emptyset$
$Clw \leftarrow closestWords(w_t, L_{Max}, \theta)$
Foreach $w_t \in Clw$ **do**
    $SW(w) \leftarrow computeWeight(w)$
EndForeach
$\beta \leftarrow getWordsDirectLink(w_t, \theta)$
$ComputeMV(\beta, L_{Max}, w_t, TemAnWords, Clw)$
$TemAnWords \leftarrow RankedWords(TemAnWords)$
$AnWords \leftarrow selectTopWords(K, TemAnWords)$
$A_I \leftarrow A_I + AnWords$
Return A_I

4 Experimental results and discussion

Throughout this section, we illustrate the experimental results of our work. We start with the experimental setup, then, we present the evaluation of our approach by introducing image classification and annotation performance.

4.1 Experimental setup

To evaluate our approach, we used ImageNet [3] and OpenImages datasets [4]:

- ImageNet dataset: images are organized according to the WordNet hierarchy. This database contains about 1,281,167 images from 1000 synsets. The number of images for each synset (category) ranges from 732 to 1300 and all images are in JPEG format. Images are heterogeneous and represent diverse themes. In our work, we used about 200K images from 1000 categories or synset, there are 190K as a training set and 10K images as a testing set.
- OpenImages dataset: where images have been annotated with labels spanning over 600 categories, there are 1,743,042 images as a training set and 125,436 as a testing set. In our work, we used about 200 images for each categories.

In order to evaluate our proposed approach, we used, as evaluation metrics, accuracy for image classification, and precision for image annotation. We provide precision at the top K (P@K) of the annotation words.

For annotation results obtained by ImageNet dataset, to evaluate the ability of our model to annotate correctly the test images, we studied the capability of our approach to detect the semantic relations between the annotation words which are assigned to the test images. For

[3] http://www.image-net.org/
[4] https://storage.googleapis.com/openimages/web/download.html

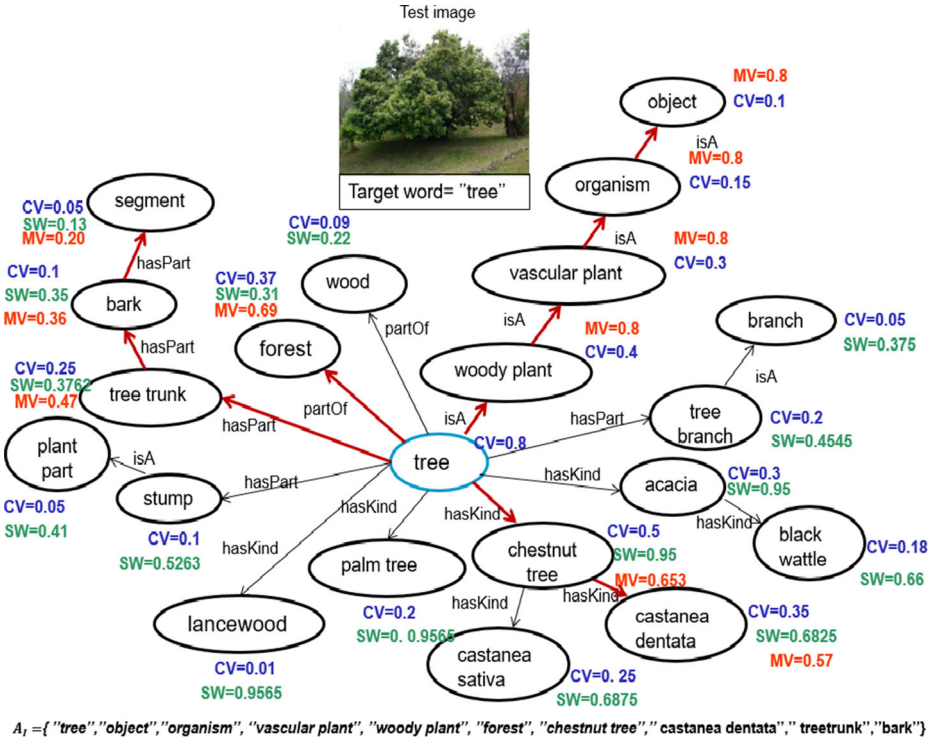


Fig. 5 Detailed example of image annotation

this purpose, we proposed a novel metric, called Target Precision, that is inspired from the method of [33].

In particular, we suppose that we have some ground-truth in the form of a matrix S, where the value of $S_{i,j} = 1$ if w_i and w_j are equivalent or there exists a synonym relationship between w_i and w_j, or the word w_j is a hyperonym of w_j and $S_{i,j} = 0$ otherwise.

In order to build the matrix S, we defined S as follows:

$$S_{i,j} = \begin{cases} 1, & \text{if i=j } \vee \exists R_{synonym}(w_i, w_j) \vee \exists R_{isA}(w_i, w_j). \\ 0, & \text{otherwise.} \end{cases} \tag{13}$$

Where:

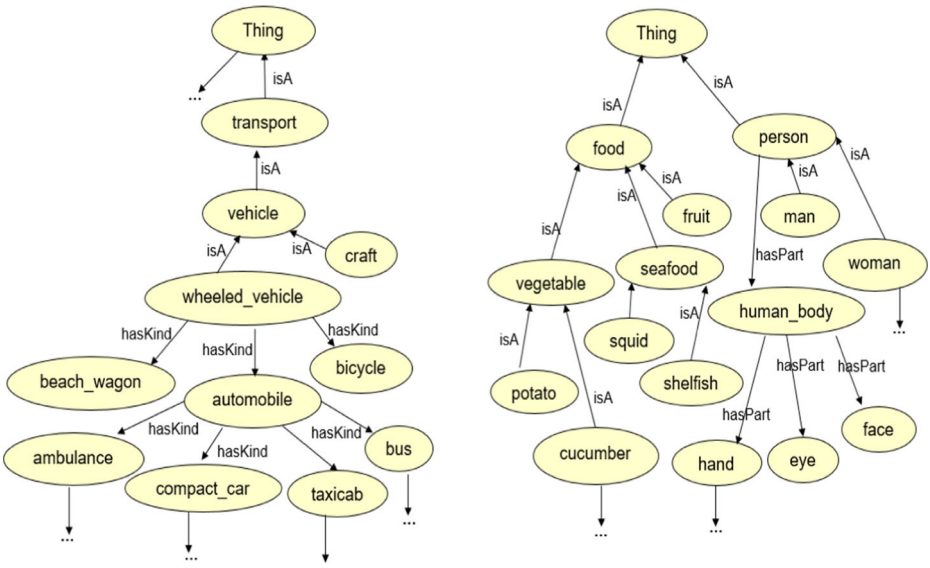
- $R_{synonym}(w_i, w_j)$ is a synonym relation between the word i and the word j.
- $R_{isA}(w_i, w_j)$ is a *isA* relation between the word i and the word j.

For this purpose, we provide Target Precision at the top K (TP@K) of the annotation words. In particular, for a ranking K annotation words= $Aw_1, Aw_2, ..., Aw_k$, TP@k is defined as:

$$TP@K(I) = \frac{\sum_{i=1}^{K} S_{w_I, Aw_i}}{K} \tag{14}$$

Where:

- I: is the test image;
- K: is the number of the annotation words that are assigned to I;



(a) ImageNet (b) Open Images

Fig. 6 A part of each ontology that has been built using ImageNet and OpenImages datasets

Each image data set has its own annotation vocabulary that is used for annotating images. For that, we need to build ontology for each image collection.

Using ImageNet, we started by extracting about 1648 words from 1000 synsets (that were related to the database images). After filtering words, there were 1400 words that are used to build ontology.

For OpenImages, using the 600 classes (that are related to this database), we extracted about 1134 related words using WordNet.

For the two cases, relationships between words are extracted using WordNet. To successfully construct each ontology, we used the OWL API.[5] Some rules are applied in order to transform the extracted relationships in OWL language.

Figure 6 represents a part of each ontology that has been built using ImageNet and OpenImages datasets.

4.2 Experimental results

4.2.1 Classification results

In this section, we are interested in showing and analyzing the image classification performance. We use different image classification strategies. We introduce the proposed strategies below:

1. **HMAX-SVM:** HMAX features are extracted and classified with SVM.
2. **BoVW-SVM:** classical BoVW model is used with SVM.

[5] http://owlapi.sourceforge.net/

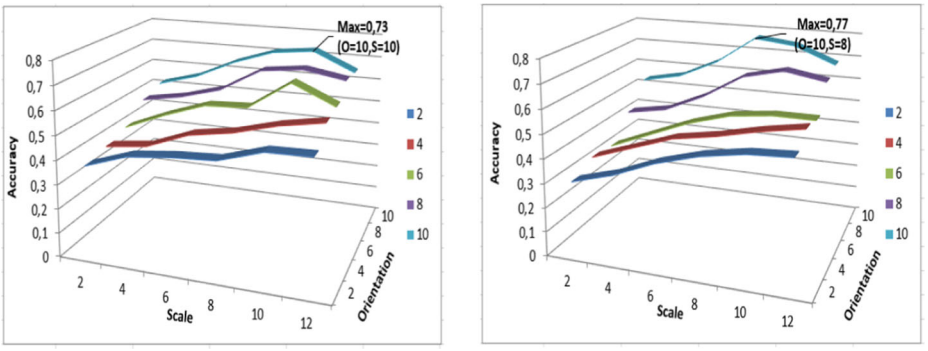
In the case of the classification method that based on HMAX model (HMAX-SVM), HMAX features are extracted as detailed in Section 3.1.1 and they are used to train SVM classifiers. The size of the final features is, in the BoVW model, given by the size of the vocabulary. However, in the HMAX model, the size is given by the number of the C2 features. For the classification method based on BoVW model (BoVW-SVM), SIFT features are extracted and quantized with KMeans and histograms of visual words are used to train SVM classifiers. For both methods, multi-class classification is done using one-versus-all SVM.

In this study, firstly we focus on evaluating the HMAX-SVM method, for this purpose, we study the variation of orientations *(O)* and scales *(S)* that are used to convolve images with Gabor filters. The goal of this part is to analyze how many orientations and scales are fit enough to improve the classification accuracy. Secondly, we report the performance of image classification depending on the number of features for each classification strategy using ImageNet and OpenImages datasets. Finally, we compare image classification accuracy obtained by both HMAX-SVM and BoVW-SVM methods.

To evaluate image accuracy of HMAX-SVM method depending on the variation of the number of orientations and scales, we tested the HMAX-SVM strategy using six different scales 2,4,6,8,10, 12 and five different orientations 2,4,6,8,10. The obtained classification results using ImageNet and OpenImages are presented in Fig. 7.

As depicted in Fig. 7, it can be seen that the best accuracy achieved 0,73 with 10 orientations and 10 scales using ImageNet dataset. We notice that the accuracy value increases with the increase of the scale until reaching 10 scales and 10 orientations. Moreover, we observe the same evaluation when using OpenImages dataset, the best accuracy achieved 0,77 with 10 orientations and 8 scales.

For both ImageNet and OpenImages, this increase of the classification accuracy could be explained by the impact of the amount of data that are extracted. However, when the number of scales tends to 12, the accuracy value decreases to 0,65 using ImageNet and to 0.68 using OpenImages. The degradation in the classification accuracy can be explained by the lack of additional data to be extracted. There is no more data to be exploited for computing response Gabor filters. So, the redundancy of the same information that are extracted with 10 scales (8 scales), and also with 12 scales (10 scales) using ImageNet (OpenImages) dataset, respectively, can decreases the accuracy value.



(a) ImageNet (b) Open Images

Fig. 7 Accuracy results obtained by HMAX-SVM depending on the variation of scale and orientation numbers using ImageNet and OpenImages datasets

Table 3 Accuracy results for HMAX-SVM and BoVW-SVM methods on ImageNet dataset

Methods \ N.features	500	1000	1500	2500	3000	3500	4000	Low-Gain	Best-Gain
BoVW-SVM	0,42	0,48	0,55	0,59	0,64	**0,69**	0,68	–	–
HMAX-SVM	0,47	0,52	0,59	0,67	0,71	**0,73**	0,71	+4,41%	+13,55%

To compare the classification performance between both HMAX-SVM and BoVW-SVM strategies, we focus on the influence of the number of features on classification accuracy. For this purpose, different vocabulary sizes are applied to experiment and a comparison of classification accuracy is shown in Tables 3 and 4.

We observe that the classification method which is based on the HMAX model (HMAX-SVM) provides a better performance than the classification method which is based on the BoVW model (BoVW-SVM) for both ImageNet and OpenImages datasets. The best improvement reaches 13,55% for ImageNet (cf. Table 3) and 19,14% for OpenImages (cf. Table 4).

Table 3 shows that the best accuracy for the HMAX-SVM method is obtained with a dictionary of 3500 features for ImageNet dataset (0,73). But, by setting the dictionary size to 4000 and using OpenImages dataset, the HMAX-SVM achieves the best performance as 0,79 (cf. Table 4).

The differences in performance between the HMAX-SVM and BoVW-SVM classification strategies can be explained by considering that the HAMX model build complexes visual features with richer information using multiple orientations and scales of image structures, however BOW model select only the interest points that are detected by SIFT detectors, and represent images by only the distribution of features, as histogram which reflects the clusters frequency of occurrence. We conclude that the classification method based on HMAX model provides a better performance than the classification method based on BoVW model on a large image database.

4.2.2 Image annotation results

To show a better performance of the proposed image annotation approach, we define different image annotation scenarios. We introduce the proposed annotation scenarios as follows:

1. **BoVW-SVM:** to perform this strategy, we annotate testing image by keeping the top K words of the best classifiers;
2. **BoVW-SVM-ONTO:** image annotation strategy based the previous strategy and the ontology: to perform this strategy, we annotate testing image by applying our method;
3. **HMAX-SVM:** image annotation strategy based on the HMAX-SVM classification strategy: to perform this strategy, we annotate testing image by keeping the top K words of the best classifiers;

Table 4 Accuracy results for HMAX-SVM and BoVW-SVM methods on OpenImages dataset

Methods \ N.features	500	1000	1500	2500	3000	3500	4000	Low-Gain	Best-Gain
BoVW-SVM	0,45	0,47	0,53	0,57	0,63	0,67	**0,68**	–	–
HMAX-SVM	0,48	0,56	0,62	0,69	0,72	0,77	**0,79**	+6,66%	+19,14%

4. **HMAX-SVM-ONTO:** image annotation strategy is based on the previous strategy and the ontology: to perform this strategy, we annotate testing image by applying our method.

The goal, in this experiment, is to show the effect and the advantages of integrating ontology with the output of the training classifiers by exploiting the ontological relationships and the confidence value of classifiers, on the performance of image annotation.

For this purpose, we compare our approach, based on the exploitation of the ontology and the classifier's confidence value by computing their membership values according to the ontological relationships, with the baseline method that consists of annotating images based only on SVM classification.

In particular, our image annotation model is based on two parameters: 1) the L_{Max} value (that can be equal to 1,2,3, 4,5 and 6) and 2) the K parameter which presents the number of annotation words.

To achieve the goal of this experiment, we set the L_{max} value of our image annotation model to 6, and we analyze the image annotation results of the strategies that were presented previously.

Table 5 shows the comparison image annotation results in terms of the Target Precision metric (TP) for the different proposed strategies on both ImageNet and OpenImages datasets.

As depicted in Table 5, using ImageNet, we observe that the annotation results of the strategy that's based on the BoVW model and ontology (BoVW-SVM-ONTO) are clearly higher than the strategy that's based on the BoVW model without ontology (BoVW-SVM). Also, we observe the same comparison when using OpenImages dataset. The best target precision obtained by BoVW-SVM-ONTO for both ImageNet and OpenImages datasets, were performed with TP@3. The best TP@3 achieves 0.68 for ImageNet and 0,69 for OpenImages (cf. Table 5). In addition, we observe the same comparison for HMAX-SVM and HMAX-SVM-ONTO. It highlights a similar increase in Target Precision when the ontology is used (HMAX-SVM-ONTO) for both ImageNet and OpenImages datasets. The best target precision (TP@3) achieves 0,73 for ImageNet and 0,75 for OpenImages (cf. Table 5).

In fact, using BoVW-SVM-ONTO, the best improvement of P@10 reaches 77,77% for ImageNet and 70,96% for OpenImages. Using our method HMAX-SVM-ONTO, the best improvement of P@10 reaches 47,05% for ImageNet and 38,46% for OpenImages (cf. Table 5).

According to the results, we conclude that our ontology-based annotation method increases the annotation results for both HMAX and BoVW features. This explains that exploiting ontological relationships with output classifiers, can improve the image annotation results.

To show a better performance of our proposed image annotation approach, we introduce a comparison of image annotation results in terms of the precision metric. Table 6 shows the comparison results of image annotation in term of precision for BoVW-SVM vs BoVW-SVM-ONTO and HMAX-SVM vs HMAX-SVM-ONTO methods on both ImageNet and OpenImages datasets.

The obtained image annotation results using BoVW model and ontology (BoVW-SVM-ONTO) is clearly higher than the strategy that's based on the BoVW model without ontology (BoVW-SVM) using both ImageNet and OpenImages datasets. The best precision improvement achieves 17,14% for P@10 using ImageNet and 35,48% for P@10 using OpenImages (cf. Table 6).

Table 5 Image annotation results evaluation with L_{Max}= 6 in the terms of Target Precision TP on ImageNet and Open Images datasets

	ImageNet						Open Images					
Methods	P@3	P@6	P@10	G-P@3	G-P@6	G-P@10	P@3	P@6	P@10	G-P@3	G-P@6	G-P@10
BoVW-SVM	0,53	0,43	0,27	–	–	–	0,55	0,45	0,31	–	–	–
BoVW-SVM-ONTO	0,68	0,59	0,48	28,30%	37,20%	77,77%	0,69	0,61	0,53	25,45%	35,55%	70,96%
HMAX-SVM	0,61	0,52	0,34	–	–	–	0,59	0,54	0,39	–	–	–
HMAX-SVM-ONTO	0,73	0,62	0,5	19,67%	19,23%	47,05%	0,75	0,65	0,54	27,11%	20,37%	38,46%

Table 6 Image annotation results evaluation with L_{Max}= 6 in the terms of Precision on ImageNet and Open Images datasets

	ImageNet						Open Images					
Methods	P@3	P@6	P@10	G-P@3	G-P@6	G-P@10	P@3	P@6	P@10	G-P@3	G-P@6	G-P@10
BoVW-SVM	0,45	0,41	0,35	–	–	–	0,47	0,43	0,30	–	–	–
BoVW-SVM-ONTO	0,51	0,46	0,41	13,33%	12,19%	17,14%	0,53	0,49	0,42	13,61%	14,41%	35,48%
HMAX-SVM	0,52	0,43	0,38	–	–	–	0,54	0,45	0,34	–	–	–
HMAX-SVM-ONTO	0,63	0,58	0,51	21,15%	34,88%	34,21%	0,65	0,56	0,52	20,37%	22,80%	52,94%

We observe the same evaluation results for HMAX-SVM and HMAX-SVM-ONTO methods. It highlights a similar increase in precision when the ontology is used (HMAX-SVM-ONTO) for both ImageNet and OpenImages datasets. The best precision (P@3) achieves 0,63 for ImageNet and 0,65 for OpenImages (cf. Table 6). Using HMAX-SVM-ONTO, the best improvement for P@10 reaches 34,21% for ImageNet and 52,94% for OpenImages (cf. Table 6).

This results indicate that our proposed method, brings an increase in image annotation precision, independently of the selected features. Moreover, we can explain that the adoption of the ontological relationships with output classifiers improves the image annotation results due to using the proposed membership value that combine classification results and ontology.

4.2.3 Impact of variation of L_{Max} value on the image annotation performance

In this section, we study the impact of the variation of the L_{Max} parameter value. The L_{Max} parameter presents the path maximum length which is used in the ontology between the w_t and the closest words that are returned by the closesetWords(w_t, L_{Max}) function (section 3.4.2) during the image annotation process.

To this end, we assigned the L_{Max} value to 1, 2, 3,4,5 and 6. Then, we tested the image annotation results by applying our method which is detailed in the section 3. In particular, we apply the image annotation process by assigning 6 values to the L_{Max} parameter (L_{Max} = 1,2,3,4,5 and 6).

For this purpose, we introduced the impact of L_{Max} value on improving the image annotation performance in the terms of P@3, P@6 and P@10. In particular, to measure the significance of the impact on improving the image annotation, we carried out 6 different runs on the P@3, P@6 and P@10 of our method (HMAX-SVM-ONTO). The annotation results according to the variation of L_{Max} parameter value for both ImageNet and OpenImages are shown in Table 7.

As depicted in Table 7, using ImageNet dataset, the value of P@3 of our method is better with $L_{Max} = 3$ (0,78), in case where L_{Max}, the P@3 decreases to 0,63. Thus, we observe that if we increase the value of L_{Max} to 6, the number of the closest words of w_t also increases. This influences on the precision of the image annotation. Also, for the P@6, as depicted in the Table 7, the image annotation precision is better when we have $L_{Max} = 3$. However, the low value of P@6 is obtained with $L_{Max} = 6$.

Table 7 Annotation results in terms of precision for our method according to the variation of L_{Max} value on ImageNet and OpenImages datasets

	ImageNet			Open Images		
Lmax	P@3	P@6	P@10	P@3	P@6	P@10
1	0,73	0,68	0,51	0,74	0,66	0,51
2	0,75	0,69	0,56	0,76	0,7	0,58
3	**0,78**	**0,71**	**0,64**	0,78	0,73	0,61
4	0,72	0,67	0,59	**0,8**	**0,76**	**0,65**
5	0,68	0,61	0,52	0,73	0,65	0,58
6	0,63	0,58	0,51	0,65	0,56	0,52

In terms of P@10, we observe the same impact of the variation of L_{Max} on improving the image annotation performance.

For OpenImages, we observe that the value of P@3 is better with $L_{Max} = 4$ (0,8) (cf. Table 7), in case where $L_{Max} = 6$, the P@3 decrease to 0,65 . Thus, we observe the same comparison for P@6 and P@10.

According to the analysis of the experimental results, we conclude that, if the L_{Max} value increases to 3 and 4, we obtained an important impact on improving the image annotation performance, but when increasing the L_{Max} value to 6, the precision values of image annotation decrease. This explains that increasing the closest words number of w_t reduces the annotation precision. This can be due to the appearance of irrelevant words in the closed semantic space of w_t in the ontology. Thus, the words can affect the performance of image annotation.

4.3 Comparison of our method with a deep learning model: inception-V3

To better evaluate our proposed approach, we illustrate a comparison of our method with a deep learning model, Inception-V3, which is a widely used for image classification and annotation.

For this purpose, we perform an annotation strategy using the Inception-V3 model, in particular, we annotate images using words that have the best scores which are obtained by this model. We compare the annotation results obtained by this strategy to our proposed strategies that are introduced in the previous subsection: BoVW-SVM-ONTO and HMAX-SVM-ONTO.

Table 8 shows the image annotation results comparison in terms of the precision for our method and Inception-V3 method on both ImageNet and OpenImages datasets.

We remarked that our method outperforms the inception-v3 method for P@6 and P@10 (cf. Table 8). Especially, the HMAX-SVM-ONTO method leads to an increase in the P@6 and P@10. In fact, the best P@6 improvement (+14.28%) and the best P@10 improvement (+36,84%) was performed when annotating images based on our method using OpenImage dataset (cf. Table 8). However, for P@3, we observe that the difference in performance of annotation results is much smaller.

We conclude that the adoption of the ontology to annotate images improves the results due to the combination of the semantic level introduced using ontologies with the classifier outputs by computing our proposed membership value.

4.4 Discussion

In this paper, we introduced our ontology- based image annotation driven by classification using HMAX features.

Our main contributions concern training visual-feature-classifiers, building an ontology that can finely represent the semantic content of images, and evaluating a membership value for each relation type found in the ontology based on both classifiers' confidence value and the semantic similarity of words. The membership value serves to rank annotation words that are assigned to a test image. The main goal is to improve the image annotation results. The experimental results show the interest of the proposed approach.

We point out that, the built ontologies, which cover about 1000 concepts and represent a rich semantic content may be seen as a reusable component of image annotation and retrieval tasks; hence, the originality of our work concerns the automatic construction of the ontology.

Table 8 Annotation results comparison in terms of precision for our method to Inception-V3 method on ImageNet and Open Images datasets

	ImageNet						Open Images					
Methods	P@3	P@6	P@10	G-P@3	G-P@6	G-P@10	P@3	P@6	P@10	G-P@3	G-P@6	G-P@10
Inception-V3	0,65	0,52	0,4	–	–	–	0,66	0,49	0,38	–	–	–
HMAX-SVM-ONTO	0,63	0,58	0,51	−3,07%	+11,53%	+27,5	0,65	0,56	0,52	−1,51%	+14,28%	+36,84%

In addition, in order to measure the significance of the improvement obtained by our approach, we carried out several tests on the image annotation precision of the different proposed scenarios.

The experimental results show that the improvements of image annotation obtained by our approach are statistically significant. Therefore, the results indicate that the gain between our approach and the baseline methods is significant.

Our proposed approach can have a great interest in Automatic Image Annotation and it can contribute to improving the performance of image annotation.

5 Conclusion

This paper describes an ontology-based image annotation using HMAX features classification. Our goal is to improve image annotation results.

Our contribution is, firstly, to extract invariant and complex visual features from training images and training classifiers, and then to automatically build the ontology that can finely represent the semantic information associated with the training images. Secondly, to combine both the classifiers' confidence values and the ontology for annotating test images. To this end, we proposed and we evaluated a membership value that is depended on each relationship that is found in the ontology. During the image annotation process, the membership value serves to select k annotation words, which are assigned to testing images.

The experiments that have been carried out highlight an improvement in image annotation results compared to baseline methods. Indeed, our proposal contributes to significantly increase the relevance of annotation results, by enhancing the precision of annotation. This improvement confirms our proposal about using ontology and visual features by exploiting both relationships and classifier outputs.

In a future work, we intend to expand our approach by exploiting other semantic relationships in order to enrich the annotation vocabulary. We also intend to improve our approach by combining a deep learning model with the ontology.

References

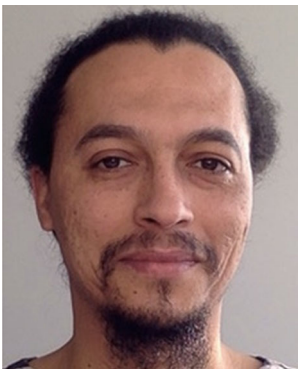
1. Bannour H, Hudelot C (2012) Hierarchical image annotation using semantic hierarchies. In: 21st ACM international conference on information and knowledge management, CIKM'12, Maui, HI, USA, October 29 - November 02, 2012. ACM, pp 2431–2434
2. Breiman L (2001) Random forests. Machine Learn 45(1):5–32
3. Cortes C, Vapnik V (1995) Support-vector networks. Machine Learn 20(3):273–297
4. Dutt A, Pellerin D, Quénot G. (2017) Improving image classification using coarse and fine labels. In: Proceedings of the 2017 ACM on international conference on multimedia retrieval, ICMR 2017, Bucharest, Romania, June 6-9, 2017. ACM, pp 438–442
5. Filali J, Zghal HB, Martinet J (2017) Visually supporting image annotation based on visual features and ontologies. In: 21st international conference information visualisation, IV 2017, London, United Kingdom, July 11–14, 2017. IEEE Computer Society, pp 182–187
6. Filali J, Zghal HB, Martinet J (2019) Ontology and HMAX features-based image classification using merged classifiers. In: Proceedings of the 14th international joint conference on computer vision, imaging and computer graphics theory and applications, VISIGRAPP 2019, vol 5. SciTePress, pp 124–134. VISAPP, Prague, Czech Republic, February 25-27, 2019
7. Filali J, Zghal HB, Martinet J (2020) Ontology-based image classification and annotation. International Journal of Pattern Recognition and Artificial Intelligence 34(11)
8. Gao H, Dou L, Chen W, Sun J (2013) Image classification with bag-of-words model based on improved SIFT algorithm. In: 9th Asian Control Conference, ASCC 2013, Istanbul, Turkey, June 23-26, 2013. IEEE, pp 1–6

9. Han Y, Li G (2015) Describing images with hierarchical concepts and object class localization. In: Proceedings of the 5th ACM on international conference on multimedia retrieval, Shanghai, China, June 23-26, 2015. ACM, pp 251–258
10. He X, Peng Y (2017) Weakly supervised learning of part selection model with spatial constraints for fine-grained image classification. In: Proceedings of the thirty-first AAAI conference on artificial intelligence, February 4-9, 2017, San Francisco, California, USA. AAAI Press, pp 4075–4081
11. Hu X, Zhang J, Li J, Zhang B (2014) Sparsity-regularized hmax for visual recognition. PloS one 9(1)
12. Lau KH, Tay YH, Lo FL (2015) A HMAX with LLC for visual recognition. arXiv:1502.02772
13. Lei J, Guo Z, Wang Y (2017) Weakly supervised image classification with coarse and fine labels. In: 14th conference on computer and robot vision, CRV 2017, Edmonton, AB, Canada, May 16-19, 2017. IEEE Computer Society, pp 240–247
14. Li Y, Liu J, Wang Y, Liu B, Fu J, Gao Y, Wu H, Song H, Ying P, Lu H (2015a) Hybrid learning framework for large-scale web image annotation and localization. In: CLEF (working notes)
15. Li Y, Wu W, Zhang B, Li F (2015b) Enhanced HMAX model with feedforward feature learning for multiclass categorization. Front Comput Neurosci 9:123
16. Ma Y, Liu Y, Xie Q, Li L (2019) Cnn-feature based automatic image annotation method. Multimedia Tools Appl 78(3):3767–3780
17. Niu Y, Lu Z, Wen J, Xiang T, Chang S (2019) Multi-modal multi-scale deep learning for large-scale image annotation. IEEE Trans Image Process 28(4):1720–1731
18. Olszewska JI (2013) Semantic, automatic image annotation based on multi-layered active contours and decision trees. Int J Adv Comput Sci Appl 4(8):201–208
19. Priyadarshini A et al (2015) A map reduce based support vector machine for big data classification. Int J Database Theory Appl 8(5):77–98
20. Reshma IA, Ullah MZ, Aono M (2014) Ontology based classification for multi-label image annotation. In: Advanced Informatics: Concept, Theory and Application (ICAICTA), 2014 international conference of. IEEE, pp 226–231
21. Riesenhuber M, Poggio T (1999) Hierarchical models of object recognition in cortex. Nature Neurosc 2(11):1019
22. Rinaldi AM (2014) Using multimedia ontologies for automatic image annotation and classification. In: 2014 IEEE International Congress on Big Data, Anchorage, AK, USA, June 27 - July 2, 2014. IEEE Computer Society, pp 242–249
23. Ristin M, Gall J, Guillaumin M, Gool LV (2015) From categories to subcategories: Large-scale image classification with partial class label refinement. In: IEEE conference on computer vision and pattern recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015. IEEE Computer Society, pp 231–239
24. Serre T, Wolf L, Bileschi S, Riesenhuber M, Poggio T (2007) Robust object recognition with cortex-like mechanisms. IEEE Trans Pattern Anal Mach Intell 3:411–426
25. Sun F, Xu Y, Zhou J (2016) Active learning svm with regularization path for image classification. Multimed Tools Appl 75(3):1427–1442
26. Sun C, Zhu S, Shi Z (2015) Image annotation via deep neural network. In: 14th IAPR international conference on machine vision applications, MVA 2015, Miraikan, Tokyo, Japan, 18-22 May, 2015. IEEE, pp 518–521
27. Theriault C, Thome N, Cord M (2011) HMAX-S: deep scale representation for biologically inspired image categorization. In: 18th IEEE International Conference on Image Processing, ICIP 2011, Brussels, Belgium, September 11-14, 2011. IEEE, pp 1261–1264
28. Theriault C, Thome N, Cord M (2013) Extended coding and pooling in the HMAX model. IEEE Trans Image Process 22(2):764–777
29. Tian D (2015) Support vector machine for automatic image annotation. Int J Hybrid Inf Technol 8(11):435–446
30. Tsai D, Jing Y, Liu Y, Rowley HA, Ioffe S, Rehg JM (2011) Large-scale image annotation using visual synset. In: IEEE International Conference on Computer Vision, ICCV 2011, Barcelona, Spain, November 6-13, 2011, pp 611–618. IEEE Computer Society
31. Wang C, Huang K (2015) How to use bag-of-words model better for image classification. Image Vis Comput 38:65–74
32. Wei Z, Luo X, Zhou F (2013) Ontology based automatic image annotation using multi-class SVM. In: Proceedings of the seventh international conference on image and graphics, ICIG 2013, Qingdao, China, July 26-28, 2013, pp 434–438. IEEE Computer Society
33. Weston J, Bengio S, Usunier N (2010) Large scale image annotation: learning to rank with joint word-image embeddings. Mach Learn 81(1):21–35
34. Wu Z, Palmer M (1994) Verbs semantics and lexical selection. In: Proceedings of the 32nd annual meeting on association for computational linguistics, pp 133–138

35. Zarka M, Ammar AB, Alimi AM (1391) Regimvid at imageclef 2015 scalable concept image annotation task: Ontology based hierarchical image annotation
36. Zhang H, Lu Y, Kang T, Lim M (2016) B-HMAX: a fast binary biologically inspired model for object recognition. Neurocomputing 218:242–250
37. Zou F, Liu Y, Wang H, Song J, Shao J, Zhou K, Zheng S (2016) Multi-view multi-label learning for image annotation. Multimed Tools Appl 75(20):12627–12644

Jalila Filali obtained her Ph.D degree in computer science from the National School of computer sciences (University of Manouba) and she is a member of RIADI laboratory at ENSI. She obtained her computer engineer degree from ISAMM in 2013. Her current research interests focus on computer vision, machine learning, deep learning, image classification and annotation using machine learning techniques and semantic features, as well as image retrieval.

Hajer Baazaoui Zghal obtained her PhD and accreditation to supervise research (HDR) in computer science from the National School of computer sciences - University of Manouba. She is actually Professor at CY TECH and permanent researcher at ETIS laboratory - CY University. She has been a holding a Professorship position at University of Manouba and senior researcher at Riadi laboratory. Her research interests mainly include knowledge-based systems and predictive systems with applications to different fields: information and image retrieval, medical and environmental big datasets.



Jean Martinet Since 2019, Jean Martinet is a Full Professor of Computer Science at Université Côte d'Azur(France), attached to the graduate engineering school Polytech Nice Sophia, and the UCACNRS joint I3S ("Informatique, Signaux et Systémes de Sophia Antipolis") research lab. His research interests include bio-inspired machine learning, computer vision, and data science. He was formerly with the University of Lille (France) during 12 years, where he was a teacher at the Technology Institute, and the head of a research group in Computer Vision at CRIStAL lab between 2013 and 2019. During 2005-2007, Jean Martinet visited the National Institute of Informatics in Tokyo (Japan) for a postdoctoral project regarding Image/video mining and annotation, and he finished a PhD in Computer Science in 2004 at Joseph Fourier University in Grenoble, on the topic of image retrieval.