Visually Supporting Image Annotation based on Visual Features and Ontologies

Jalila Filali, Hajer Zghal, Jean Martinet

Visually Supporting Image Annotation based on Visual Features and Ontologies

Jalila Filali
ENSI, RIADI Laboratory
University of Manouba
Tunisia
Email:jalila.filali@ensi-uma.tn

Hajer Baazaoui Zghal
ENSI, RIADI Laboratory
University of Manouba
Tunisia
Email: hajer_bz@yahoo.fr

Jean Martinet
CRIStAL Laboratory
University of Lille1
France
Email: jean.martinet@univ-lille1.fr

***Abstract*—Automatic Image Annotation (AIA) is a challenging problem in the field of image retrieval, and several methods have been proposed. However, visually supporting this important tasks and reducing the semantic gap between low-level image features and high-level semantic concepts still remains a key issue. In this paper, we propose a visually supporting image annotation framework based on visual features and ontologies. Our framework relies on three main components: (i) extraction and classification of features component, (ii) ontology's building component and (iii) image annotation component. Our goal consists on improving the visual image annotation by:(1) extracting invariant and complex visual features; (2) integrating feature classification results and semantic concepts to build ontology and (3) combining both visual and semantic similarities during the image annotation process.**

***Keywords*—Visualisation, image annotation, visual features, ontologies.**

I. Introduction

The first approaches in image retrieval are the text-based image retrieval, where images are annotated by experts and then retrieved using the annotated keywords. Due to the exponential growth of the quantity of images, assigning relevant text keywords to images is too tedious and time-consuming. To overcome this problem, Content-Based Image Retrieval approaches (CBIR) were developed. In this case, images are represented and retrieved by their visual content, such as color, texture, and shapes. Thus images are extracted without using semantic information describing their contents. However, in CBIR, most systems were only able to interpret images based on low level features and not able to automatically describe images with semantic representation. In addition, recent studies have shown that there is a significant semantic gap between low-level image features and high-level semantic concepts. To overcome this problem and to bridge the semantic gap between low level image features and semantic concepts, the third category of approaches of image retrieval has been introduced. These approaches focused on Automatic Image Annotation. Image annotation consists of automatically assigning relevant text keywords to any given image that reflect its visual content. The main goal of automatic image annotation is to improve image retrieval. Despite the large amount of research works in the image annotation area, two main problems persist: (1) visually supporting the image annotation process and (2) mapping visual and semantic features are still a challenging issues.

In this paper, we propose a visually supporting image annotation framework based on visual features and ontologies. The idea is to visually assist the image annotation process by : (1) extracting and classifying visual features into general categories, (2) building ontologies, and (3) combining the results of classification and relevant concepts associated to the feature classes.

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 our image annotation framework and its different components. In section 4, a case study illustrating the functionality of the proposed framework as well as the application results are presented. Finally, section 5 concludes this paper and proposes directions for future works.

II. Overview and motivations

In image retrieval area, Automatic Image Annotation (AIA) is an important issue aiming to bridge the gap between low level features and high level semantic information. According to the large number of regions in images, techniques for image annotation and retrieval become increasingly important. In this context, several image annotation and retrieval methods based on various learning techniques have been applied [17] and [8]. Moreover, image annotation could be presented as a classification problem when the goal is to improve image classification and annotation accuracy [9], [18] and [19].

Several works have focused on scene recognition and image classification using features learning with Convolutional Neural Networks (CNNs) [20], [5] and [12].

In addition, several works have been focused on comparing features and supporting image retrieval, in [2] an approach based on multidimensional visualisation and coordination techniques is proposed. In this work, the coordination techniques are used to perform image retrieval methods. In [6], a visual analysis tool is developed to support searching and comparing features of multivariate datasets.

In [18], a novel model based on the Self-Organizing Map (SOM) neural network has been proposed. The aim is to learn features useful for the problem of automatic images classification. In particular, in this work, the SOM model is used to learn single-layer features from the extremely challenging CIFAR-10 data-set. Several other methods based on HMAX architecture have been used for improving image classification [7], [4] and [15]. In [16], a novel method of feature learning based on HMAX architecture for image classification has been

proposed. The purpose of this work is to build complex features with richer information to improve image classification. In addition, a large number of ontological techniques have been applied along with a great deal of research efforts [14], [13] and [10]. For instance, [14] have proposed a novel method to use a hierarchy defined on the annotation words derived from text ontology to improve automatic image annotation and retrieval. The hierarchy is used in the context of generating a visual vocabulary for representing images and as a framework for the proposed hierarchical classification approach for automatic image annotation. In [13], a complete framework to annotate and categorize images has been proposed. This approach is based on multimedia ontology's organized following a formal model to represent knowledge.

In the literature, we remarked that the proposed approaches for image annotation and retrieval have a key problem: the semantic gap between the visual features and the richness of human semantics is not well reduced to improve image annotation and retrieval results. However relations between visual and semantic informations are not well exploited. In particular, most approaches do not exploit both visual and semantic features in a dependence manner. Therefore, they are not expressive and they cannot be efficiently used for automatic image annotation and retrieval. In parallel, several image annotation approaches have been proposed to solve semantic gap problem by using ontologies which could improve image annotation accuracy. Moreover, annotation tasks are generally carried out without any visual support.

Starting from these remarks, we propose a visually supporting image annotation framework that has four main characteristics: (i) it allows extracting invariant and complex visual features; (ii) it allows to learn classification of features; (iii) it is based on ontology which is built integrating feature classification results and semantic concepts; and, (iv) it allows annotating image by combining both visual and semantic similarities.

III. Image Annotation Framework based on Visual Features and Ontologies

In this section, we detail the proposed image annotation framework and its components. As depicted in Figure 1, the visual image annotation is performed on two main phases: training phase and testing phase. The training phase is supported by two components: (1) extraction and classification of features and (2) ontology's building. The testing phase is performed by image annotation component. This phase is composed of three main steps: feature extraction (Figure1 Step (3.1)), image classification (Figure 1 Step (3.2)) and image annotation (Figure 1 Step (3.3)). These components will be detailed in the following subsections.

A. Training phase

1) Extraction and classification of features: The extraction and classification of features component consists, firstly, in extracting features of training images; secondly, in classifying the features. To extract visual features from training images, we use a HMAX architecture; in particular we adopt the HMAX model proposed by [15]. The reason for which we use the HMAX model is to provide complex and invariant visual information and to improve the discrimination of visual features in order to obtain a good classification during the training phase.

The HMAX model follows a general 4 layer architecture. Below we describe the operations of each layer. Simple ("S") layers apply local 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 [15]:
$$L1_{s,o} = |g_{s,o} * I| \quad (1)$$
- **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 partitions each $\mathbf{L1_{s,o}}$ features into small neighborhoods $\mathbf{U_{i,j}}$, and then selects the maximum value inside each $\mathbf{U_{i,j}}$.
$$L2_{s,o} = \max_{U_{i,i} \in L1_{s,o}} * U_{i,j} \quad (2)$$
- **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.
$$L3_{s,m} = \alpha_m * L2_s \quad (3)$$
- **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.
$$L4 = \max_{(x,y),s} L3_S^1(x,y), ..., \max_{(x,y),s} L3_S^M \quad (4)$$

After extracting visual features from training images, the layer 4 feature vector for each image is used to train a classifier.

2) Ontology's building component: The ontology's building component consists, firstly, in selecting the closest concepts of a concept associated to a given feature class, and extracting taxonomic and semantic relationships between them using BabelNet[1], and then, creating the extracted relationships; secondly, in adding features of concepts of the corresponding level to the current feature class, and then, by the use of the concepts, reclassifying all updated features. This process is repeated in a recursive manner until all concepts are treated and all relationships between them are created, where each features class and its corresponding concept are the input of a new ontology level creation process. In last level, relations between concepts and features are added in order to improve the exploitation in both visual and semantic information. In particular, in the extracting taxonomic and semantic relationships phase, if the target concept has many senses according to BabelNet (Babel synset), a semantic disambiguation task is performed using BabelFy[2].

Let Θ be the ontology to be built, Dc denote the original concepts which are extracted from the meta data of our image database.

Let consider:

- $Dc = Dc_1, Dc_2, ..., Dc_M$: a set of the original concepts;
- Lr : a lexical resource ;

[1] http://babelnet.org/about

[2] http://babelfy.org/about

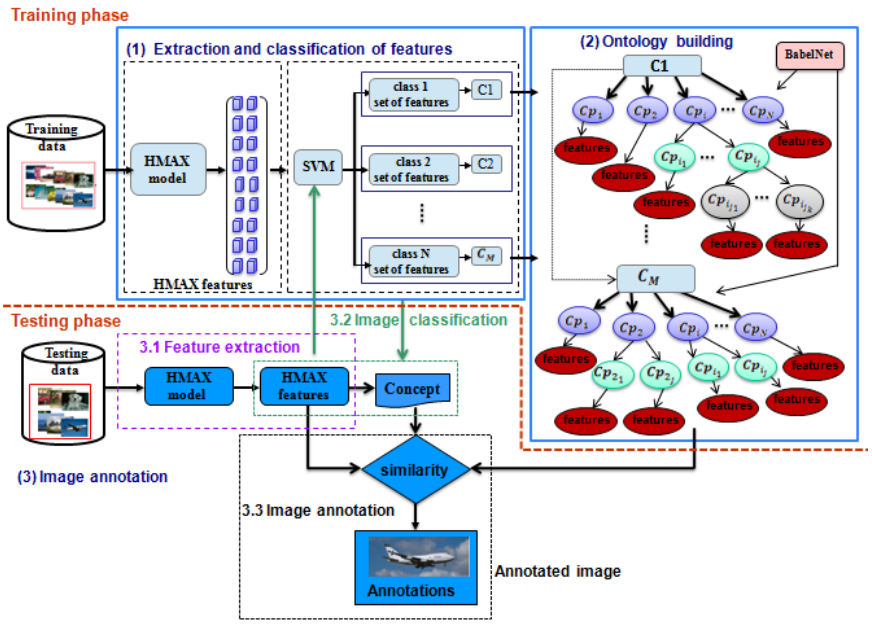

Fig. 1: Framework's Architecture

- Fc : is the feature class related to the current concept;
- C : is the current concept that is the root of the ontology which we will build.

We summarize the ontology building process according to the algorithm 1.

After building ontology of each concept and its corresponding features class, semantic relationships between all root nodes of each ontology, are extracted, and created in order to provide an integrated exploitation and view of both the related semantic concepts and visual features.

B. Testing phase

The testing phase includes three main steps which are: features extraction, image classification and image annotation (Figure1). To achieve the testing phase we need to use similarity measures in order to select the annotation concepts that represent the given image.

1) Similarity measures : To perform the image annotation process, both visual and semantic similarities between concepts are computed.

Visual similarity between concepts: the visual similarity between concepts consists on estimating the visual correlation between two concepts. In the field of image annotation and retrieval, recent approaches have been proposed to measure this similarity between concepts using several methods like the confusion matrix [1], Kullback-Leibler (KL) divergence [3] and the common features shared across the classes. In our work, we propose a simple method for estimating the visual correlation between concepts using the euclidean distance between feature vectors associated to each concept.

The visual similarity between two adjacent concepts C_i and C_j is:

$$VisualSim(C_i, C_j) = d(v(C_i), v(C_j)) \quad (5)$$

The visual similarity between two not adjacent concepts C_i and C_j is:

Algorithm 1 Summarized Algorithm of building ontology

Input: Fc: feature class, C: Concept, ID: image database
Output: ontology

Initialization (Θ:ontology, C: root concept)
$Dc \leftarrow ExtractOriginalConcepts(ID)$
$Cp \leftarrow ExtractRelatedConcepts(C, Lr)$
for each Cp_i **in** Cp **do**
  TaxonomicRelations←ExtractTaxRelations(C, Cp_i)
  SemanticRelations←ExtractSemRelations(C, Cp_i)
  $Update\Theta : add(TaxonomicRelations)$
  $Update\Theta : add(SemanticRelations)$
  if $(Cp_i \notin Dc)$ **then**
    $UpdateFc : add$(features of nearset concept of Cp_i)
  end if
end for
$SubFeaturesClasses \leftarrow classify(Fc, Cp)$
for each $Cp_i \in Cp$ **do**
  if $(Cp_i isLeafNode())$ **then**
    ConFeatRelation←CFRelation$(Cp_i, SubFeaturesClasses)$
  end if
end for
$Update\Theta : add(ConFeatRelation)$
return *ontology*

$$\text{VisualSim}(C_i, C_j) = d(v(C_i), v(CM_1)) * \prod_{k=1}^{n} d(v(CM_k), v(CM_{k+1})) * d(v(CM_k), v(C_j))$$

where

- $V(C)$: is the feature vector related to the concept C.
- $d(v(C_i), v(C_j))$: is the euclidean distance between $v(C_i)$ and $v(C_j)$).
- C_M : is the intermediate concept between C_i and C_j in the ontology.

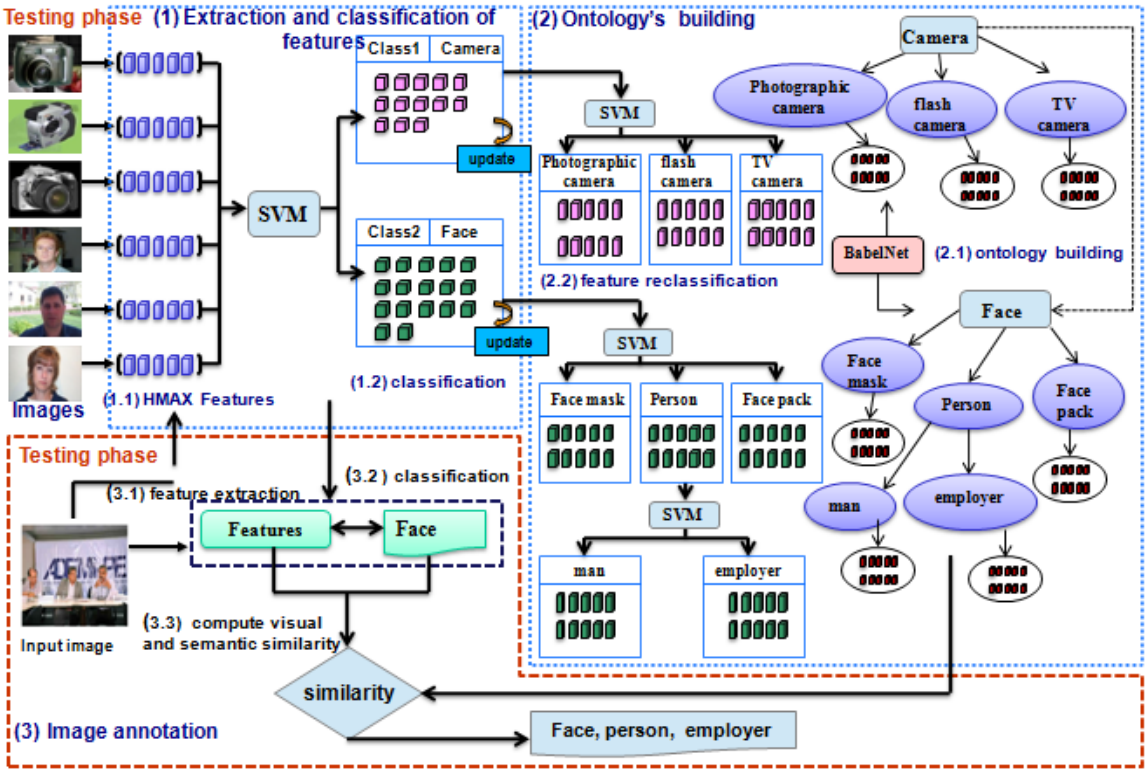

Fig. 2: Case study: illustrated process for a given image query

- n : is the number of intermediate concept between C_i and C_j.

Semantic similarity between concepts: In ontology-based image annotation and retrieval, many semantic similarity measures can be used. In this context, several studies used semantic similarity measures to improve the image annotation and retrieval [11]. In our context, semantic similarity between concepts is computed according to the following formula.

$$SemanticSim(C_j, C_k) = \eta(C_j, C_k) = \frac{w\vec{C_j} w\vec{C_k}}{||w\vec{C_j}||.||w\vec{C_k}||} \quad (6)$$

where wC_k : is the concept C_k vector defined in the words space.

2) Image annotation: First, a test image is introduced and features are extracted using HMAX architecture. Then, the class with maximum probability, which is generated by the classifier, is selected as the concept of the test image. Finally, giving this concept, we find the closest concepts in the ontology. The annotation concepts are found using both visual and semantic similarities between concept that is found in the image classification step (Figure 1: 3.2) and nearest concepts that are defined in the ontology. The annotation is generated using the concepts that have a high annotation score which is computed by combining both visual and semantic similarities. The steps of image annotation process are detailed in the algorithm 2. During the annotation image process (Figure 1 Step (3.2)), visual and semantic similarities are used in order to compute annotation score between the given image and the closest concepts that are defined in ontology. The annotation score is performed by a formula that combines the visual and semantic similarities. The both visual and semantic similarity measures that are used will be presented in the next subsection.

Algorithm 2 Image annotation

Input: I: test image, Θ : ontology
Output: annotation concepts

```
Initialisation: annotation concepts vector Vc = ∅
FeaturesVector ← feature extraction(I)
Classify(featuresVector)
C ←concept of result class
Ct ← ExtractClosestConceptOfC(C,Θ)
Generate a empty vector of annotation score Vs
for each Ct_i ∈ Ct do
    AnnoScore(I,Ct_i) =visSim(C,Ct_i) * semSim(C, Ct_i)
    Add AnnoScore (I, Ct_i) in Vs
end for
bestConcepts←Select the top-k concepts from Vs
Add bestConcepts in Vc
return VC
```

IV. Experimental evaluation

In order to show that our proposal framework can have a great interest and that can contribute to improve the image annotation task, we are interested, firstly, in conducting case study; secondly, in developing a prototype to show that the proposed framework can improve the performance image of annotation. We implemented the prototype using Matlab and JAVA programming language. In this section, we first present

the experimental setup, then we present the case study and finally we show the obtained experimental results.

A. Experimental setup

To evaluate our approach, we used the Image Caltech 101 data-set. This data-set contains 9144 images from 102 categories. The number of images per category varies from 31 to 800. Most images are of medium resolution (about 300 x 300 pixels) and well aligned with some variability.
In our case, we use 7 categories (faces, ibis, car, airplanes,camera, lotus, elephant) with 20 images for each categories. Also we use 10 and 15 training examples for each category and 35 images are used as testing set. In order to evaluate the proposed image annotation method, we used the average precision as evaluation metric.

B. Case study

Throughout this section, we illustrate a case study of the proposal framework. Figure 3 illustrates the different steps with a specific example related to the given image. Let's consider a test image composed of the three "man". So this image represents three "faces" of three "persons".
As depicted in Figure 3, during the testing phase, when a test image is submitted, a HMAX features are extracted (Figure 3 Step (3.1)), after that, this image is classified aiming at affecting a concept to the input image (Figure 3 Step (3.2)). When our system detects the concept that represents the input image, visual and semantic similarities between this concept and all the closest concepts that are defined in the corresponding ontology, are combined and computed in order to determine score annotation for each pair "input image-closest concept" (Figure 3 Step (3.3)). As shown in Figure 3, the concept "face" is detected when the classification of the input image is achieved. Using this concept "face" and its corresponding features, the system extracts the top ranked concepts that can represent the input image according to their score annotation.
During the training phase, image features are extracted using HMAX architecture and a learning classification consists in training one-vs.-all SVM to operate in the feature space. In particular, features is classified into two classes and the concepts "camera" and "face" are associated to them. The extracted visual features and their classes are visualized (Figure 3 Step (1)). After that, for each features class and its corresponding concept, ontology is built. In particular, when relationships between the concept which is associated to the features class and their closest concepts, are extracted and created, the class of features is updated and reclassified using SVM classifier (Figure 3 Step (2.1) and Step (2.2)). As shown in Figure 3, concepts "photographic camera", "flash camera" and "TV camera" are found, thus taxonomic relationships between these concepts and "camera" are created. In parallel, concepts "face mask", "person" and "face pack", are also selected and taxonomic relationships between them are created. Thus, features of concept "person" are reclassified into two classes that represent concepts "employer" and "man". Part of the built ontology is represented in Figure 4.
When our input image is submitted to our system, visual features are extracted and concept "face" is detected. After that, annotation score is computed by combining both visual

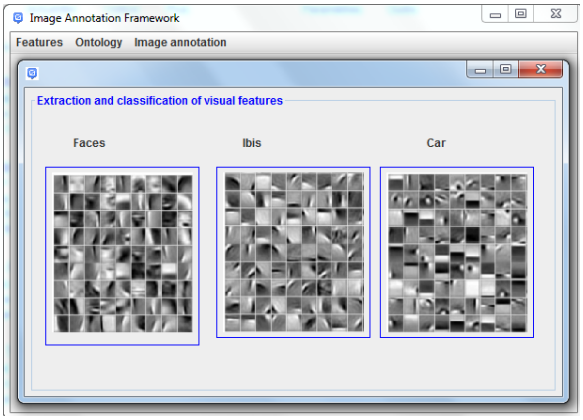

Fig. 3: Visualization of extracted features and their classes

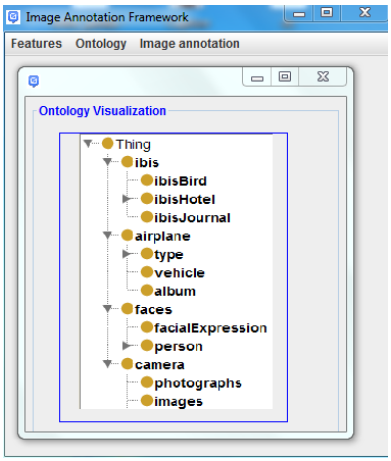

Fig. 4: Part of the built ontology

and semantic similarities between this concept and the nearest concept that are extracted from the ontology. Thus, as depicted in Figure 3, according to the annotation score, relevant annotation concepts which represent semantic content of the test image like "person", and "employer" are returned. An example of image annotation results is visualized in Figure 5.

C. Experimental results

In order to evaluate the results improvement of image annotation we define two image annotation strategies. The first strategy is based on extraction and classification of visual features (ECVF) and the second strategy is focused on combining the visual features classification with ontologies (VFCO) and using visual and semantic similarities.
The evaluation results shown in Table 1 represent the average precision obtained according to the strategies. The strategies VFCO and ECVF use the HAMAX features to annotate images. Compared to a classical image annotation strategy, we remarked that our strategy VFCO outperforms them. We observe that combining visual features classification and ontologies improves the average precision by 5.55% and 4.41% when using respectively 10 training images per category and 15 training images per category.

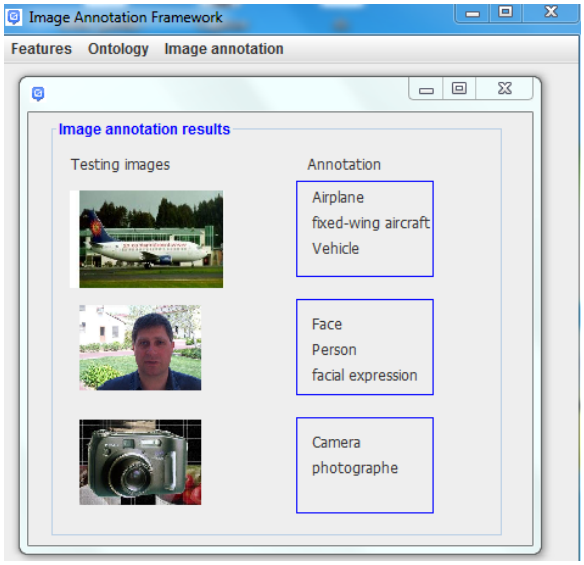

Fig. 5: Image annotation results

TABLE I: Image annotation results evaluation

Strategy	Hmax parameters		Average precision	
	scale	**orientation**	10 training images	15 training images
ECVF	8	12	0.54	0.68
VFCO	8	12	0.57(+5.55%)	0.71(+4.41%)

V. Conclusion

This paper presents a visually supporting image annotation framework based on visual features and ontologies. Our contribution can be summarized in: (1) visually support image annotation process, (2) combining feature classification results to semantics in order to build ontologies, (3) providing an integrated exploitation and view of semantics and Hmax features. The case study and the experiments that have been carried out highlight an improvement of the image annotation results.
Currently, we integrate big data technologies to the framework in order to experiment it on ImageNet.

References

[1] Jia Deng et al. "What does classifying more than 10,000 image categories tell us?" In: *Computer Vision–ECCV 2010* (2010), pp. 71–84.

[2] Danilo Medeiros Eler et al. "Coordinated multiple views to support image retrieval". In: *Information Visualisation (IV), 2014 18th International Conference on*. IEEE. 2014, pp. 139–144.

[3] Jianping Fan et al. "Mining multilevel image semantics via hierarchical classification". In: *IEEE Transactions on Multimedia* 10.2 (2008), pp. 167–187.

[4] Xiaolin Hu et al. "Sparsity-regularized HMAX for visual recognition". In: *PloS one* 9.1 (2014), e81813.

[5] Kai Kang et al. "Object detection from video tubelets with convolutional neural networks". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 817–825.

[6] Hiroaki Kobayashi, Hiroko Suzuki, and Kazuo Misue. "A Visualization Technique to Support Searching and Comparing Features of Multivariate Datasets". In: *Information Visualisation (iV), 2015 19th International Conference on*. IEEE. 2015, pp. 310–315.

[7] Kean Hong Lau, Yong Haur Tay, and Fook Loong Lo. "A HMAX with LLC for visual recognition". In: *arXiv preprint arXiv:1502.02772* (2015).

[8] Yong Li et al. "Hybrid Learning Framework for Large-Scale Web Image Annotation and Localization." In: *CLEF (Working Notes)*. 2015.

[9] Julien Mairal et al. "Convolutional kernel networks". In: *Advances in Neural Information Processing Systems*. 2014, pp. 2627–2635.

[10] Joanna Isabelle Olszewska. "Semantic, Automatic Image Annotation Based On Multi-Layered Active Contours and Decision Trees". In: *International Journal of Advanced Computer Science and Applications* 4.8 (2013), pp. 201–208.

[11] Siddharth Patwardhan and Ted Pedersen. "Using WordNet-based context vectors to estimate the semantic relatedness of concepts". In: *Proceedings of the eacl 2006 workshop making sense of sense-bringing computational linguistics and psycholinguistics together*. Vol. 1501. 2006, pp. 1–8.

[12] Mattis Paulin et al. "Local convolutional features with unsupervised training for image retrieval". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2015, pp. 91–99.

[13] Antonio M Rinaldi. "Using Multimedia Ontologies for Automatic Image Annotation and Classification". In: *Big Data (BigData Congress), 2014 IEEE International Congress on*. IEEE. 2014, pp. 242–249.

[14] Munirathnam Srikanth et al. "Exploiting ontologies for automatic image annotation". In: *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM. 2005, pp. 552–558.

[15] Christian Theriault, Nicolas Thome, and Matthieu Cord. "Extended coding and pooling in the hmax model". In: *IEEE Transactions on Image Processing* 22.2 (2013), pp. 764–777.

[16] Christian Theriault, Nicolas Thome, and Matthieu Cord. "HMAX-S: deep scale representation for biologically inspired image categorization". In: *Image Processing (ICIP), 2011 18th IEEE International Conference on*. IEEE. 2011, pp. 1261–1264.

[17] Dongping Tian. "Support Vector Machine for Automatic Image Annotation". In: *International Journal of Hybrid Information Technology* 8.11 (2015), pp. 435–446.

[18] Marco Vanetti. "Unsupervised Feature Learning using Self-Organizing Maps". In: *International Conference on Computer Vision Theory and Applications (VISAPP)*. 2013.

[19] Xiaosong Wang et al. "Unsupervised Joint Mining of Deep Features and Image Labels for Large-scale Radiology Image Categorization and Scene Recognition". In: *arXiv preprint arXiv:1701.06599* (2017).

[20] Bolei Zhou et al. "Learning deep features for scene recognition using places database". In: *Advances in neural information processing systems*. 2014, pp. 487–495.