Multimed Tools Appl (2015) 74:1397–1421
DOI 10.1007/s11042-014-1955-9

A generic framework for semantic video indexing based on visual concepts/contexts detection

Nizar Elleuch · Anis Ben Ammar · Adel M. Alimi

Published online: 25 April 2014

Abstract Providing a semantic access to video data requires the development of concept detectors. However, semantic concepts detection is a hard task due to the large intra-class and the small inter-class variability of content. Moreover, semantic concepts co-occur together in various contexts and their occurrence may vary from one to another. Thus, it is interesting to exploit this knowledge in order to achieve satisfactory performances. In this paper we present a generic semantic video indexing scheme, called SVI_REGIMVid. It is based on three levels of analysis. The first level (level1) focuses on low-level processing such as video shot boundary/key-frame detection, annotation tools, key-points detection and visual features extraction tools. The second level (level2) aims to build the semantic models for supervised learning of concepts/contexts. The third level (level3) enriches the semantic interpretation of concepts/contexts by exploiting fuzzy knowledge. The obtained experimental results are promising for a semantic concept/context detection process.

Keywords Semantic indexing · Concepts · Contexts · SVI_REGIMVid

1 Introduction

Recent years have witnessed a rapid evolution of technology acquisition and diffusion of multimedia data. The web, television channels and video blogs have led to the explosion of audiovisual databases. This vast amount of audiovisual data currently represents the main challenge research studies on the automatic information processing particularly in the representation and indexing of video contents. The archive professionals need to index their digital video bases for an effective access to video contents.

N. Elleuch (✉) · A. Ben Ammar · A. M. Alimi
REGIM-Lab.: REsearch Groups on Intelligent Machines, University of Sfax, National School of Engineers (ENIS), BP 1173, Sfax 3038, Tunisia
e-mail: nizar.elleuch@ieee.org

A. Ben Ammar
e-mail: anis.benammar@ieee.org

A. M. Alimi
e-mail: adel.alimi@ieee.org

Three main approaches have been proposed for content-based video indexing: text-based approach, feature-based approach, and semantic-based approach. The text-based approach focuses on using keywords or tags to describe the video content. However, these textual descriptions do not reflect objectively the wealth of video content. The feature-based approach aims to extract low-level features such as color, texture, shape and motion from the video data and use them as indexing keys. However, these low-level features are generally complex and not intelligible for ordinary users. To overcome the drawbacks of the text-based approach and the feature-based approach, the semantic-based approach focuses on the automatic video content annotation with their semantic meanings. As a consequence, many efforts have been made in the multimedia field in order to define and unify a large lexicon called semantic concept lexicon [36]. At the Large Scale Concept Ontology for Multimedia (LSCOM) workshop a set of 3156 concepts is proposed. They are related to objects (eg. flower, Airplane), specific objects (eg. Us_Flags, Infant), scenes (eg. Nighttime, Classroom, Doorway, Cityscape) and events (eg. People_Marching, Swimmer).

The ground-truths show that the semantic concepts have a large intra-class and a small inter-class variability of visual content. In addition, the same semantic concept may appear in various contexts and its appearance may vary according to these contexts, on the one hand, and concepts always coexist together, on the other hand. Therefore, the main problem of the semantic-based video indexing is how to categorize effectively the semantic content and exploit the semantic knowledge to overcome the semantic gap between data representation and their interpretation by humans [44].

Nowadays, many efficient approaches for semantic content categorization lean on the visual words and the construction of qualitative and compact codebooks in order to generate a histogram of codeword counts, called bag-of-words (BoW). However, the BoW model cannot describe objectively and discriminatively the content of an image and neglects the spatial distribution of visual words, although it allows significant distinctiveness of the representation. Thus, our approach joins with this tendency in order to overcome these difficulties. We introduce a novel approach for semantic content categorization called visual pseudo-sentences (PS) vocabulary. It invests in the integration of the spatial distribution of the visual elementary words within their own context by grouping them together to provide visual pseudo-sentences which are more informative and discriminative words.

In addition, indexing audiovisual documents based on the concept detector is limited by a fundamental scalability problem. It requires a rather big number of training examples in order to produce a generic indexing system. It also ignores the fact that concepts always coexist together. For example, the concept sky frequently co-occurs with the concept Airplane flying. Thus, the contextual information from Airplane flying is expected to help detect sky. Thus, we try to extract, represent and store such information via a fuzzy ontology. Such ontology represents the fuzzy relationships (roles and rules) among every context and its semantic concepts. We use a deduction engine to handle richer results in our video indexing system by running the proposed fuzzy ontology.

The remaining of our paper is organized as follows: In Section 2, we presented state-of-the-art of the semantic video indexing approach. Section 3 described the proposed framework. In Section 4, extensive experiments were presented and discussed, followed by conclusions in Section 5.

2 Related works and background

Semantic concept detection has captured extensive research attention in multimedia indexing and retrieval, thanks to its promising capabilities to bridge the semantic gap. Many techniques

have been developed and several image/video indexing systems have been built [1, 13, 38, 46]. Early systems, such as Informedia [13, 14] and MARVEL multimedia analysis engine [37], proposed to combine the video modalities in order to enhance the video interpretation. However, the majority of these systems have mainly focused on visual content processing. The semantic pathfinder [46] explored different paths through three consecutive analysis steps: content analysis, style analysis, and context analysis. In addition, other works have focused on including spatio-temporal information into visual content processing in order to detect moving objects and events [23, 53, 56]. However, other alternatives have focused on using semantic knowledge to enhance the accuracy of concept detection [47].

These systems have a common approach: understanding the semantic in a multimedia document is basically performed by building visual concept detectors in order to annotate automatically the video shots with respect to a set of semantic concepts.

Building concept detectors is assimilated to the representation of the whole key frame of each video shot and concepts detection is performed by conducting a learning process. In fact, the semantic concept detection in an image/a video is perceived as a problem of content representation and of a pattern classification. For instance, given pattern x, part of a key-frame of a video shot Vs_k, the aim is to obtain a probability measure $P(Vs_k|c_x)$, which indicates whether the corresponding semantic concept c_x is present in the key frame of each video shot. So, such concepts are portrayed by a composition of the image as a whole, rather than characterized by one specific part in the image. Furthermore, the background context of a semantic concept can be more informative than the semantic concept itself and can provide significant detection solutions. Therefore, it is very interesting to model the whole image for the concepts categorization.

The codebook, or bag-of-visual-words, is one of the most useful techniques for the modeling of image contents. It can be described as follows. First, a set of stable points or regions are sampled from the input images. These stable points or regions which carry important information are repeatedly found under transformations, including scale [26, 29], rotation, and affine [28, 30] transformations. Next, feature descriptors are constructed using local image information in the neighborhood of the feature points (or regions). The set of features collected from each image is then clustered into k visual words. Thus, an image can be represented as a k-dimensional histogram by counting the visual words occurrence in the image. This fixed size vector representation of an image then serves as input to a classifier.

Accordingly, the performance of video indexation systems by visual concepts depends essentially on the construction of the dictionary, on the one hand, and the structure that represents the content on the other. Hence, several methods have been developed to construct a small, compact, vocabulary that discriminates well between different concept categories. These techniques can roughly be divided into two categories, namely a semantic approach which relies on annotation and a data driven approach characterized by an unsupervised clustering.

The first category is based on the principle of compositionality. An image can be derived from the meaning of its constituent named patches. So, a semantic vocabulary is obtained by manually assigning the meaningful labels to image patches [35, 38].

The second category tries to discriminate the continuous high-dimensional features space. To achieve this goal, several algorithms and approaches are available in the literature. In fact, among these approaches, the unsupervised clustering by K-means algorithm is the most popular one which attempts to determine K partitions by minimizing the variance between them. However, one of the main disadvantages of K-means is the fact of specifying the number of clusters K as an input to the algorithm. Moreover, in K-means clustering the centers are almost exclusively around the denser regions in high dimensional feature space and thus fail to

decode other informative regions. For the above mentioned reasons, some other proposed approaches recommend the use of RAdius-Based Clustering (RABC) to generate discrete visual codebook [11, 18]. RABC sequentially seeks to detect new clusters by finding local maxima of the density and clustering all data points within a fixed radius *r*. Thereby, the clusters centroid represents the local maximum density and all clustered points are assigned to the new cluster. Hence, the RABC algorithm yields an even distribution of visual words over the continuous high dimensional feature space.

After discretization of the mentioned local features, the obtained codebook allows a visual categorization by representing an image via a histogram of codeword counts. Therefore, an image can be described as a BoW, but it suffers from some limitations. The main disadvantage of BoW is the fact that it discards information about the spatial relations between visual words in an image. So, the bags of words representation cannot describe objectively and discriminate the content of an image. In fact, the ground truth shows that most visual words have some mutual spatial dependencies in different degree. Therefore, a few recent works have proposed methods to extend bags of words taking into consideration spatial information about the visual words. The common approach is to employ graph-based representations that deal separately with local patches (nodes of the graph) and spatial relations (arcs of the graph) [6, 15].

In [4], R. Behmo presented a new compact, self-contained representation of image content. It consists in computing properties of a graph built on interest points of the image; interest points are first collected in the image to constitute the set of nodes of what we call a "feature graph". Then the corresponding "collapsed graph" is built by associating each node of the feature graph to an entry of a descriptor codebook. Finally the symmetric matrix of commute times between the nodes of this collapsed graph is computed to obtain the final image representation which encodes both the relative frequency of the codebook entries to which the features are associated, as in the case of the bag of features representation. However, in addition to the quite high computation time, the performance of the graph model depends on the image semantic context. Therefore, we need to weight effectively the connections in the feature graph.

In [24], S. Lazebnik proposes a multiscale bag of features construction to capture the spatial structure of features. An image is partitioned into increasingly fine sub-regions and for each one the histogram of local features is computed. However, region-based representation can generate problems of mismatching features in case the concept objects in the training images are in different regions from those in the testing images. One more problem is the fact that many objects may cross region boundaries. For resolving this issue, M. Marszalek employed spatial relationships between features to reduce background clutter [27]. He strengthens the weights of features that agree on the position and shape of the object and eliminates the weights of the background features. However, the background context of an object can be more informative than the object itself.

In [12], K. Grauman used the Earth Mover's Distance (EMD) [41], which incorporates spatial distances, for comparing images based on their discrete distributions (bags) of distinctive local invariant features, without clustering descriptors. However, using the spatial information only, the contents semantic is removed. Furthermore, the clustering process may considerably reduce the computation time and treatment complexity.

In [2], J. Amores used a novel type of image representation: the Generalized Correlogram (GC). For the image representation, the object is described as a constellation of GCs where each one encodes information about some local patches and the spatial relations between these patches to the others. Nevertheless, the performance of GC declines due to low-resolution images and content transformation.

Though the concept detectors approach for semantic video indexing provides satisfactory performance for some concepts, most of these concepts are still not easily detected. Thus, the

typical concept detector approach alone is not efficient for multimedia processing, especially for video indexing. In fact, the major drawback of this approach is the choice of the learning machine and its parameters. Another drawback of this approach is that the concept detectors are often developed independently, ignoring the fact that concepts always coexist together and the training samples are naturally multi-labeled. Therefore, much new research has involved the exploration of the semantic knowledge among concepts for video indexing. They particularly aim to develop a context-based concept fusion (CBCF) framework to enhance the concept detection results [16, 17, 20, 45, 54, 57]. These approaches fall into two categories.

The first category is based on the exploration of pair-wise concept correlation [17, 54, 57] that is generally determined by observation (e.g., from manual annotations or machine tagging of the training data set). Based on the statistical principles and the manual annotation used to approximate pair-wise concept relations, these previous approaches suffer from two major drawbacks. First, such approximations may not be generally consistent when we have a limited training data. Furthermore, it is difficult to obtain accurate statistics involving different generic concepts in general video collections. Therefore, the relationship to other concepts is generally ignored. Second, the manual annotation methods used for labeling semantic concepts are most often incomplete. Thus, such missing information can lead to inaccurate approximations and to misleading statistics.

Using multimedia ontologies based on logical roles is the most appropriate solution for such drawbacks. Indeed, they formally allow the specification of the relationships between the different concepts. So far, multiple ontologies have been proposed. The Large Scale Concept Ontology for Multimedia (LSCOM) [17] proposes a hierarchical inter-concept relationships such as “IS-A”. However, ImageNet [7] proposes several useful relations, which are inspired from WordNet [31], such as “opposes”, “is_a_member_of”, “is_a_part_of”, etc. Since the appearance of contexts as meta-information to describe a group of conceptual entities and to partition a knowledge base into manageable sets, various conceptual ontologies have been suggested. In fact, in [49], Spyrou et al. have introduced a visual context ontology which contains various relations among different types of content entities, such as images, regions, region types and high-level concepts. To model the topological relations between concepts, the authors proposed “Adjacent”, “Inside”, “Outside”, “Above”, “Below”, “Left” and “Right”. In [33], Mylonas et al. have proposed Fuzzy topological relations defined by domain expert in order to model real-life information such as “Part”, “Specialization”, “Example”, “Instrument”, “Location”, “Patient” and “Property”. However, in [35], the authors defined other relationships incorporating fuzziness in their definition. They utilized a set of relations derived from MPEG-7 such as “Similar”, “Accompanier”, “Part”, “Component”, “Specialization”, “Generalization”, “Example”, “Location” and “Property”. For each relation, they have associated a degree of confidence.

The second category is based on learning techniques [16, 20, 45]. In [45], the contextual relationship is modeled by SVM. Firstly, the Discriminative Model Fusion (DMF) method generates a model vector aggregating the detection score of all the individual detectors. After that, an SVM is learned in order to refine the detection of the original concepts. Although some performance improvement is reported in [16, 45], there are major drawbacks. First, these approaches are fully supervised and require explicit knowledge of the target semantic concept and ground-truth labels such as the ontology hierarchy and the Bayesian networks (manually constructed in most cases). Second, the number of correlated concepts for a given concept is generally small, compared to the un-correlated ones; thus using all the concepts as in [45] will significantly affect the performance. Third, some detectors may provide inaccurate probability estimation, especially for the concepts with very few positive training samples. Therefore, such detectors will have a significantly detrimental impact on the learning relation.

3 Overview of the semantic video indexing system: SVI_REGIMVid

In this section, we present an overview of our semantic video indexing system, called SVI_REGIMVid. The Fig. 1 presents the main system components. SVI_REGIMVid is based on three levels of analysis. The first level (level1) focuses on low-level processing such as video shot boundary/key-frame detection, annotation tools, key-points detection and visual features extraction tools. The second level (level2) aims to build the semantic models for a supervised learning of the concepts/contexts. The third level (level3) enriched the semantic interpretation of concepts/contexts by exploiting the fuzzy knowledge.

3.1 Level1:low-level processing

3.1.1 Video shot boundary/key-frame detection

Shot boundary detection (SBD) is the process of automatically detecting the boundaries between shots in a video. From 2001 to 2007, automatic SBD was one of the main tasks for the TRECVid evaluation campaign [43]. There was a large variety of approaches proposed. We adopt the shot boundary detection system proposed by Fraunhofer HHI [39]. For each video shot, the middle image is selected as a key-frame.

3.1.2 Annotation tools

The semantic video indexing systems learn the concepts that will be detected from a large collection of positive and negative examples which are manually annotated. This need of knowledge has raised a new challenge for several competitions such as TRECVID and ImageCLEF Photo Annotation. In this context, various annotation tools have been proposed such as Efficient Video Annotation (EVA) system [52], VideoAnnEx [25], TRECVID 2007 Collaborative Annotation system [3], etc. These annotation tools try to specify with a binary label whether an image contains a given concept or not. However, these training images are not, in all cases, relevant to a target concept. Thus, it is inescapable to define new strategies to select a set of deterministic and relevant knowledge for each concept.

To this end, we have developed our own semi-automatic soft collaborative annotation system [22]. As shown in Fig. 2, it aims to aggregate the training data at three relevance levels or classes, namely “Highly Relevant” (TP), “Relevant” (P) and “Somewhat Relevant” (PP). In

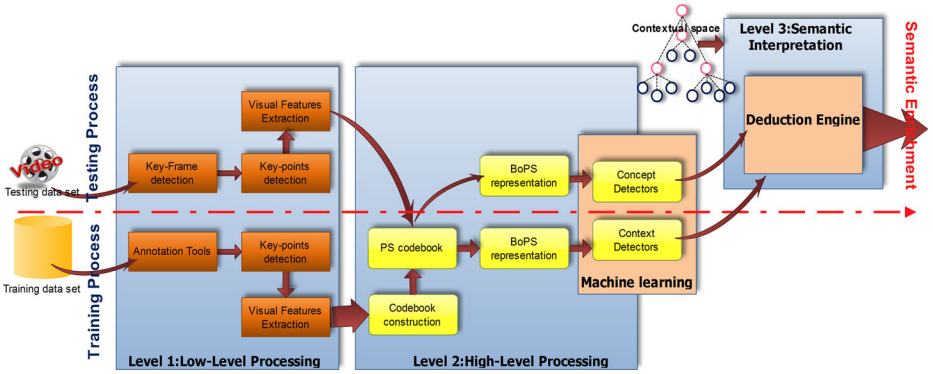

Fig. 1 Overview of the SVI_REGIMVid system

fact, we have proposed a Fuzzy annotation, a modeling of the user's judgments and an evaluation of the annotation quality. In addition, we have introduced a relevance feedback process by jointly exploiting annotated and unannotated data.

3.1.3 Key-points detection

A very large set of interest point detectors have been already proposed in the literature. This wide variety of detectors is mainly due to a lack of a definition of the concept of the interest points. However, most of these works basically assume that key points are equivalent to corners or more generally speaking to points of the image that are characterized by a significant gradient amount in more than one direction. Obviously, a differential framework results from these definitions.

Studies on visual attention, more related to the human vision, propose very different models. The basic information is still the variation in the stimuli but it is no longer taken into account in a differential way but mainly as an energy point of view [42]. According to [5], our approach joins with the last purpose [10]. The main idea is to exploit a detector based on luminance and variation of the edge orientation. As luminance variations occur at almost every scale in an image, we have to use multi-resolution and multi-orientation frameworks (8 orientations and 4 scales). In fact, a pixel is considered as an interest point if its luminance is significantly higher than those of its neighbors or it changes the orientation of the edge. When extracted at a low level in the pyramid, this pixel is related to a local luminance. When extracted at a higher level in the pyramid, this pixel is related to a region luminance. These pixels luminance are then combined in a top-down scheme in a pyramid yielding a final representation where both local and global information are taken into account.

3.1.4 Visual feature extraction

The aim of feature extraction is to derive compact descriptions and represent the interest pattern. Hence, we use a set of different visual descriptors at various granularities for each representative key-frame of video shots on a neighborhood of each detected key-point. In fact, we use several visual descriptors of different modalities (color, texture and shape) as Color Histogram, Co-occurrence Texture, Gabor, etc. [8, 19].

After extracting the visual features, we proceed to the early fusion step. We combine a selection of the feature vectors resulting from visual feature extraction. We adopt the method proposed in [48], using vector concatenation to unify the features representation. After feature normalization, which is performed for each feature vector before fusion through applying the L2 normalisation, we obtain an early fusion vector. This vector serves to extract codebook information for semantic concepts retrieval. Then, a set of Self-Organizing Maps (SOMs) is trained on these features to provide a common indexing structure across the different representations.

3.2 Level2:high-level processing

3.2.1 Codebook construction through Kohonen maps

One of the most important constraints of discrete visual codebook generation is the uniform distribution of visual words over the continuous high-dimensional feature space. Self-Organizing Maps (SOM) proved their performance in so doing. In fact, it has been successfully utilized for indexing and browsing by projecting the low-level input features to the two-

dimensional grid of the SOM map [21]. Thus, to generate a codebook of prototype vectors from the above features, we utilize the SOM-based clustering.

The SOM is basically an unsupervised and competitive learning algorithm. Relying on a grid of artificial neurons whose weights are adapted to match input vectors in a training set, SOM finds the optimal set of prototypes. It is basically made up of two layers: an input layer which corresponds to a set of input neuron vectors and a competitive output layer which corresponds to a set of M connected neurons $Ne=\{Ne_1,Ne_2,\ldots,Ne_M\}$. Every input N-dimensional vector $x=\{x_1,x_2,\ldots,x_N\}$ is connected to N neurons of the map through weights W_i.

After the learning process, we tried to discover the optimal number P of clusters. In fact, when the number of SOM units is large, similar units need to be grouped, i.e., clustered, to facilitate quantitative analysis of the map and the data. This is due to the topological ordering of the unit maps. Thus, after the learning process of the SOM map, we grouped the similar units using the partitive clustering applying the K-means introduced by J. Vesanto and E. Alhoniemi [51] as it allows very fast clustering with an acceptable accuracy.

At this stage, each image can be described by a set of visual elementary words by assigning each region descriptor of the image to the nearest cluster in the description space. We obtain a histogram model ignoring the spatial relationships linking all the visual elementary words. Thus, after extracting BoW, we interpret the dispersion of different words in images to further discriminate the concepts categorization. In the next section, we present how to build visual PS vocabulary.

3.2.2 Spatial information representation

Spatial information is very important in image retrieval [24, 55]. Thus, to incorporate such information, we use a grouping of the visual words to describe this model. We assumed that each visual word is spatially dependent on others to form a given pattern. As an image is considered a set of patterns, we group together the visual elementary words to yield visual pseudo-sentences which are more informative words. Thus, our approach is based on the assumption that the visual words with their locations are represented on the basis of exact coordinates. There are spatial relations between visual words which are metric (distance) or ordinal distance (direction).

Although the detected key points are invariant to rotation, scale and geometric partial distortions, the spatial relationships between visual words are not invariant. Thus, we should use a metric that ensures the invariance of the spatial relationship. In [40], P. Punitha proposed a novel data structure called relative spatial distance (RSD) matrix, which is used for the representation of symbolic images. RSD matrix is built using the relative Euclidean distances (RED) between the iconic objects in the symbolic image and is also invariant to all the 2-dimension transformations such as rotation, translation, scaling and composite transformations like flipping and reflection. Therefore, we adopt RED to count spatial relations between all visual words for each image.

Given an image I which contains M visual words. Each visual word V_i, in this image, is identified by its Cartesian coordinates (x_i,y_i). The pseudo-sentences algorithm can be summarized as follows:

1. Calculate the matrix of dimension M^2 which represents a RED between each visual word and others according to:

$$RED(V_i,V_j)=\begin{cases} 0 & \forall(i,j)\in\{1,\ldots,M\}^2, i\le j \\ \sqrt{\left[\left(x_i/\sqrt{x_i^2+y_i^2}\right)-\left(x_j/\sqrt{x_j^2+y_j^2}\right)\right]^2+\left[\left(y_i/\sqrt{x_i^2+y_i^2}\right)-\left(y_j/\sqrt{x_j^2+y_j^2}\right)\right]^2} \end{cases} \quad (1)$$

2. For each visual word, find the closest visual word V_{min}. This can be done by looking up relative Euclidean distance which satisfies equation below.

$$\|V_i - V_{i\_min}\|_2 = \underset{j=1}{\overset{M}{argmin}} \left(\|V_i - V_j\|_2 \right) \tag{2}$$

3. Find all very close pair of visual word satisfying this condition:

IF V_i *has a minimum distance to* V_j
AND V_j *has a minimum distance to* V_i
THEN (V_i, V_j) *is a very close pair of visual word*

The result is a set of visual words, each of which containing two words (at most 3 in case of a tie) with the smallest relative Euclidean distance, that could be subsequently merged.

4. Merge the two visual words V_i and V_j to one word V_{ij}^*.
5. Calculate the new Cartesian coordinates of the merged words V_{ij}^* or V_{ijt}^* (in case of a tie). This can be done by the below equation.

$$\begin{cases} V_{ij}^* = \left(\dfrac{x_i + x_j}{2}, \dfrac{y_i + y_j}{2} \right) \\ V_{ijt}^* = \left(\dfrac{x_i + x_j + x_t}{3}, \dfrac{y_i + y_j + y_t}{3} \right) \end{cases} \tag{3}$$

6. Repeat only once the steps 1–5, taking into consideration the merged and unmerged words.

The result of this combination is a new set of visual words, each of which containing at most 9 elementary words. With a set of 9 words, we can built 362880 (9!) PS. Thus, we have to analyze the arrangement of different elementary words to distinguish between them. In fact, for each set merged, we project its different visual elementary words on an axis Δ to identify the sequence of visual words: PS. The Δ equation is defined as follows:

$$\Delta : Y = a + b * X \tag{4}$$

Where:

$$a = \left(\sum y_i - B * \sum x_i \right) \Big/ n_1$$

$$n_1 : \textit{dimension of set merged } V^*$$

$$b = \left(n_1 \sum x_i * y_i - \sum x_i * \sum y_i \right) \Big/ n_1 * \sum {x_i}^2 - \left(\sum x_i \right)^2$$

Once the equation of the axis is determined, we use an orthogonal projection of different visual words to determine the syntax of this pseudo-sentence. The syntax of each pseudo-sentence is determined according to a fixed direction. We have chosen arbitrarily the direction from top to bottom and left to right.

Finally, our method allows generating a new codebook which represents pseudo-sentences that are based on the grouping of visual elementary words. Although we used only two stages

of spatial clustering, the size of the new codebook is very large compared to the elementary codebook. The size of the obtained codebook allows having more discriminative models, but also a need for the memory, storage and the computing time to train a much more important classifier. Therefore, we perform a refinement step to reduce the size of the obtained pseudo-sentences codebook.

The pseudo-sentence refinement process: The refinement process is likened to a problem of optimization of the pseudo-sentences construction. For example, let be a pseudo-sentence V^*_{12345} which consists of a chronological grouping of 5 visual words ($V_1V_2V_3V_4V_5$) with just one occurrence in the image collection to categorize, contrariwise another pseudo-sentence V^*_{34} which consists of a chronological grouping of 2 visual words (V_3V_4) with very large occurrence in the same image collection. Therefore, the pseudo-sentence V^*_{12345} only increases the size of the codebook. Thus, it would be very interesting to subdivide the pseudo-sentence. To achieve this goal, two steps are considered: the analysis of syntax and the occurrence of all constructed pseudo-sentences and the subdivision of low occurrence pseudo-sentences.

The syntax analyzing process The most natural way of representing a dictionary is through building a tree. The tree creation is based on the share of the common prefixes. Thus, to analyze the syntax of all the created pseudo-sentences, we use an M-ary tree.

In our context, the M-ary tree will have 8 levels and each node will have at most M children (size of elementary codebook). The construction of this M-ary tree is based on the following principles:

- For each node, we associate a visual elementary word, a boolean variable that indicates if the pseudo-sentence is complete and an integer variable to calculate the occurrence of each pseudo-sentence in the image collection to categorize.
- The first level of the M-ary tree will be represented by all the visual elementary words.
- To add a sentence in the tree, we use a recursive procedure. Indeed, we add, firstly in the M-ary tree, each pseudo-sentence that is formed by 2 and next by 3 visual elementary words. The addition principle procedure consists in adding word by word starting from the root and going down gradually from one level to another. The break condition is reached when we no longer have words to add in the tree. In this case, the boolean variable value of the current node is modified to true and also the integer variable is incremented by 1. Also, while going through the tree, node by node to insert a new pseudo-sentence, the value of the integer variable of each node is incremented by 1 if the boolean variable value is equal to true. Secondly, every pseudo-sentence that are formed by 4 or 5 or 6 or 7 or 8 or 9 visual elementary words are then added by applying the same addition procedure. Furthermore, for each of these pseudo-sentences, we will apply the following strategy, n – 2 times for each pseudo-sentence (where n is the number of visual elementary words in the pseudo-sentence). Whenever we remove the first visual elementary word encountered in the pseudo-sentence: we check whether the new pseudo-sentence, obtained after the removal, is already in the M-ary tree. In this case, we increment the integer variable by 1. Otherwise we move to the next iteration until we finally obtain two visual elementary words in the pseudo-sentence.

For each image in the dataset, the same procedure is performed simultaneously in order to categorize their content later on.

The subdivision process The purpose of the subdivision process is to reduce the obtained codebook size by analyzing the syntax of every pseudo-sentence having a visual elementary

words count strictly upper to 3. It involves removing some pseudo-sentences that have an occurrence less than or equal to a threshold *th* (we fixed th to 5) from the codebook and to replace them with other elementary pseudo-sentences. The choice of one or more elementary pseudo-sentences that will replace a generic pseudo-sentence is treated as a problem of maximization of the size of the elementary pseudo-sentences. Thus, we proceed as follows:

- We select all the pseudo-sentences which will be eliminated from the pseudo-sentences codebook to be replaced afterwards.
- We sort out all the selected pseudo-sentences in a descending order according to the number of visual elementary words that they contain.
- For each of these pesudo-sentences, we determine the set of its sub-pseudo-sentences that are defined in the dictionary. Thus, we replace each of them by the longest sub-pseudo-sentence or by a combination of two sub-pseudo-sentences.

After the subdivision process, the generated visual pseudo-sentences vocabulary can be effectively utilized as a codebook of visual concepts for image encoding and representation.

3.3 Concepts/contexts learning

The classification plays an important role to bridge the semantic gap. In our work, we use the LIBSVM implementation (for more details see http://www.csie.ntu.edu.tw).

In order to decrease the influence of the imbalance of different distributions of relevance classes, we propose to generate three repartitions of the training database. Indeed, the first considers the examples annotated "Highly Relevant" as positive examples and the other represents the negative ones. The second merges the two classes "Highly Relevant" and "Relevant" in a positive class and others are considered as negative examples. The third considers the examples of "Highly Relevant", "Relevant" and "Somewhat Relevant" as positive examples, and examples of "neutral" and "irrelevant" as negative examples.

Once the repartitions for all the training images are built, the classifiers are learnt for each repartition to build concept models. So, for each concept, three classifiers are simultaneously learnt. Generally, the models are built through a process of supervised learning. A supervised learning algorithm is run with a set of training data, containing both positive images and negative ones from the visual concept to learn and provides the model. This algorithm needs to find the most discriminative information to represent concepts later. We employ Support Vector Machines (SVMs) as our base learning algorithm for their effectiveness in many learning tasks. The primary goal of an SVM is to find an optimal separating hyper plane that gives a low generalization error while separating the positive and negative training samples.

SVMs return binary output y_i for each test image, $y_i \in \{-1, 1\}$. To fit the SVM binary outputs into probabilities, we use Platt's method that produces a probabilistic output using a sigmoid function:

$$P\left(y\middle|f\right) = \frac{1}{1+\exp(Af+B)} \tag{5}$$

Where A and *B* are estimated by training using the maximum likelihood estimation from the training data set and $f=f(x)$ is defined as follows:

$$f(x) = \sum_{i=0}^{N-1} \alpha_i y_i K(x, x_i) + b \tag{6}$$

Where N is the dimension of x, $K(.,.)$ is a radial basis kernel function, $K(x,x_i)=\exp(-\gamma\|x-x_i\|^2)$, $\sum_{i=0}^{N-1}\alpha_i y_i = 0$, α_i is the learned weight of the training sample x_i,, $\alpha_i>0$ and γ is a kernel parameter to be tuned for the best accuracy, $\gamma>0$.

Once the three classifiers are learnt with probabilistic SVM, we merge the three outputs by calculating the weighted average to obtain the final model using this equation:

$$C = \delta * C_{tp} + \beta * C_{tp+p} + \varepsilon * C_{tp+p+pp} \tag{7}$$

Where δ, β and ε are predefined by experts on the SVM training. These averages are then ordered to select the examples. Accordingly, each video shot Vs_k is ranked with a probabilistic measure $P(Vs_k|c_i)$ or $P(Vs_k|t_j)$.

3.4 Level3: semantic interpretation

The contextual information is an ambiguous term. It has been defined and interpreted in several ways in different domains. In the multimedia literature, visual context was introduced in [32, 34, 50], as an extra source of information for both object detection and scene classification.

Accordingly, the context may provide an important cue in order to enrich the semantic interpretation and further enhance the performance of semantic indexing and multimedia retrieval content systems. Thus, we model the contextual information in order to exploit and understand the high-level content efficiently. The context information modeling consists of three steps: semantic knowledge representation/interpretation, semantic concept/context categorization and refinement process. The semantic knowledge representation focuses on building the context space that represents the relationships (roles and rules) among every context and its semantic concepts. Such information is extracted, represented and stored via a proposed fuzzy abduction engine mated with an inference engine. More specifically, the inference engine provided by fuzzy description logics (fDLs) is used for context ontology construction that links each context space with its semantic concepts. The second step, semantic concepts/contexts categorization, focuses on the construction of a small, compact, vocabulary that effectively discriminates concepts and contexts. Based on this vocabulary, semantic concepts/contexts models are trained via an SVM classifier. The third step, refinement process, aims to enrich and enhance the semantic interpretation of our video indexing system. Based on fuzzy rules defined in our fuzzy ontology, a deduction engine is used to handle new richer results.

To specify the proposed ontology to describe the semantic knowledge, a less expressive fuzzy description logic is applied to facilitate fast computations. We detail how we constructed a contextual knowledge model for semantic interpretations below. In this field, our fuzzy ontology O^f is modeled as $O^f=\{T,C,R^f_{tc},R^f_{ct},R^f_{cc|t},Q\}$ where:

- $T=\{t_1,t_2,\ldots,t_n\}$ is a set of n contexts
- $C=\{c_1,c_2,\ldots,c_m\}$ is a set of m concepts
- $R^f_{t_ic_j}: TxC\rightarrow[0,1]$; $i\in\{0,\ldots,n\}$ *and* $j\in\{0,\ldots,m\}$; is a fuzzy rule that the context t_i performs for the concept c_j
- $R^f_{c_it_j}: CxT\rightarrow[0,1]$; $i\in\{0,\ldots,m\}$ *and* $j\in\{0,\ldots,n\}$; is a fuzzy rule that the concept c_i performs for the context t_j
- $R^f_{c_ic_{j|t_k}}: CxC\rightarrow[0,1]$; $i\in\{0,\ldots,m\}$, $j\in\{0,\ldots,m\}$ and $k\in\{0,\ldots,n\}$; is a fuzzy rule that the concept c_i performs for the concept c_j within the context t_k

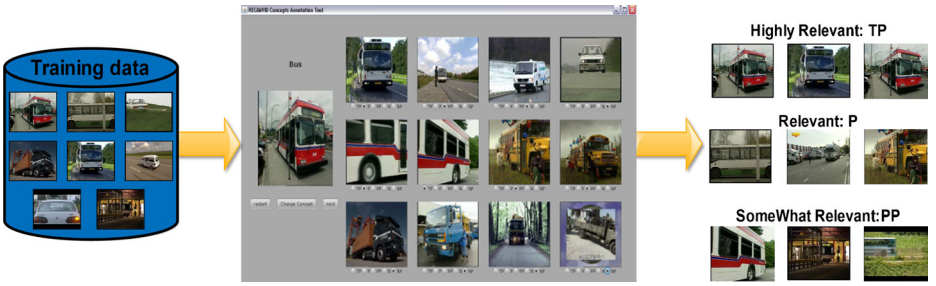

Fig. 2 REGIMVid Semi-Automatic Collabrative Annotation tools

- Q is a set of fuzzy qualifiers. In O^f, we define two qualifiers: "weak" and "strong" (Fig. 3).

We have also defined some roles between concepts and contexts {Generalization, IsRelatedTo, IsPartOf, Includes}. Their interpretation is rather simple and detailed in Table 1. The choice of these specific roles is motivated by the generic reasoning scenarios designed to improve concept detection. However, these roles can be further enriched depending on referred applications.

Based on probabilistic scores of concepts and contexts provided by the indexing process, a fuzzification step is performed. This step aims to handle the imprecision and inexactness of concepts and contexts detectors, on the one hand, and to generate the fuzzy inputs required by fuzzy rules on the other. Thus, we consider a concept c_i or a context t_j "Relevant" in a video shot Vs_k if $P(Vs_k|c_i)$ respectively $P(Vs_k|t_j)$ is greater than 0.7. However a concept c_i or a context t_j is qualified by "Not-Relevant" in a video shot Vs_k if $P(Vs_k|c_i)$ respectively $P(Vs_k|t_j)$ is between 0.3 and 0.7. These two intervals are empirically chosen according to expert observations and the probabilistic mesures provided by concept detectors.

Based on these fuzzy inputs, the deduction engine explores all the defined rules in order to infer the most appropriate ones and thus generates an optimal score for the target rule output. In this field, two cases arise: when a fuzzy rule is qualified as "Strong" or "Weak".

In the first case, the deduction engine proceeds as follow: Let R_k^f be a fuzzy rule defined as : R_k^f:c_i is Strong RelatedTo c_j within $t_{k'}$ and let $P(Vs_i|c_i)$ and $P(Vs_i|t_{k'})$ be, respectively, a

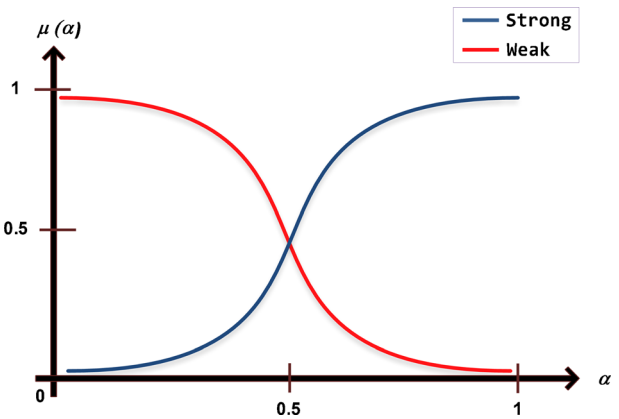

Fig. 3 Two Fuzzy βeta function to represent O^f qualifiers

Table 1 Semantic relationships between concepts and contexts

Name	Symbol	Meaning	TxT	CxC	CxT	TxC	Definition
Generalization	$t_i:t_j$	The concept t_i is the generalization of the concept t_j	X				LSCOM
IsRelatedTo	$c_i\|t_k \rightarrow c_j$	The concept c_i is related to the concept c_j within t_k		X			Learning
IsPartOf	$\{c_i\} \in t_j$	A set of concept c_i is a part of the context t_j			X		Learning
Includes	$t_i \supset c_j$	The context t_i includes the concept c_j				X	Expert

score detection of concept c_i and context $t_{k'}$ in the same video shot Vs_i. The optimal score, or the deduced relevance degree, of the fuzzy rule R_k^f outputs, denoted as $\alpha_k^{'}(c_j)$, is computed as follow:

$$\alpha_k^{'}(c_j) = \mu_k\left(\max\left\{P\left(Vs_k\middle|c_i\right), P\left(Vs_k\middle|t_{k'}\right)\right\}\right) * \mu_{Strong}(\alpha_k) \tag{8}$$

Where μ_k and α_k are, respectively, the βeta membership function and the confidence degree of the k^{th} fuzzy rule according to the role "IsRelatedTo".

In the second case, the deduction engine applies the following equation.

$$\alpha_k^{'}(c_j) = \mu_k\left(\min\left\{P\left(Vs_k\middle|c_i\right),\ P\left(Vs_k\middle|t_{k'}\right)\right\}\right) * \mu_{Weak}(\alpha_k) \tag{9}$$

The same approach is built by the deduction engine for the other rules according the role "IsPartOf", "Includes" and "Generalisation".

4 Experiments

The main goal of our experimentation is, first to check the effectiveness of the semantic concepts categorization using pseudo-sentences (PS), and, second, test the added value of the context space for semantic concept detection. Hence, we first present the experimental setup and the obtained results of the SVI_REGIMVid within TREC Video Retrieval Evaluation (TRECVID) 2010 at the Semantic Indexing task. Second, we examine the performance for representation of the image contents obtained in the space of the semantic concepts categorization induced by pseudo-sentences, multiscale bag-of-words and graphs. Third, we compare the effectiveness of our fuzzy ontology O^f for the enhancement of the semantic concept detection to existing techniques. Finally, the scalability of the proposed approach is discussed.

We use the inferred average precision (infAP), the precision (P) and the recall (R) as a performance metric and the best scores which are reported in Tables 3, 4 and 5 are given in bold.

4.1 TRECVID 2010 evaluation

In order to check the SVI_REGIMVid scalability on the detection of a large set of concepts, the experimentations are mainly conducted with different scale and difficulties on TRECVID 2010 benchmark data set.

4.1.1 Data sets

At TRECVID 2010 Semantic Indexing task there are two data sets provided by the National Institute of Standards and Technology (NIST): a test and a development data set. The development data set IACC.1.tv10.training contains 3200 Internet Archive videos (50GB, 200 h) while the test data set IACC.1.A contains approximately 8000 Internet Archive videos (50GB, 200 h). IACC.1.A is annotated with 130 semantic concepts.

4.1.2 Experimentation setup

We have participated with two instances of SVI_REGIMVid system: SVI_REGIMVid_1 and SVI_REGIMVid_2 [10]. The first system integrates only the first and the second level of analysis while the semantic interpretation process is incorporated within SVI_REGIMVid_2. We note that the results given by TRECVID 2010 are limited to the detection of 30 among 130 semantic concepts which were initially suggested. The set of 30 semantic concepts is reported in Table 2.

The experimental findings of SVI_REGIMVid_1 and SVI_REGIMVid_2 at TRECVID 2010 are shown in Fig. 4. SVI_REGIMVid_2 achieves the best mean inferred average precision (infAP=0.089) among SVI_REGIMVid_1 where the infAP is 0.085. Also, we observe from the results that the SVI_REGIMVid achieved a good performance to detect the high level features in video data. In fact, most of the infAP scores are on the top of the median curve. In addition, for some concepts such as Bicycling, Singing, Telephones and Car_Racing SVI_REGIMVid fulfilled the best results. Moreover, our novelty stems from the fact that is now clearly noticeable that the use of the semantic interpretation process allows improving semantic concept detection. The infAP score of 3 among 30 concepts (Animal, Boat_Ship, and Vehicle) increased since a small number of rules (268 among 2 876 rules defined in O^f) was inferred at the semantic refinement for the 30 concepts.

Table 2 TRECVID 2010: high-level features

ID	ID_TREC	High level features	ID	ID_TREC	High level features
1	004	Airplane_Flying	16	052	Female_Human_Face
2	006	Animal	17	053	Flowers
3	007	Asian_People	18	058	Ground_Vehicles
4	013	Bicycling	19	059	Hand
5	015	Boat_Ship	20	081	Mountain
6	019	Bus	21	084	Nighttime
7	022	Car_Racing	22	086	Old_People
8	027	Cheering	23	100	Running
9	028	Cityscape	24	105	Singing
10	029	Classroom	25	107	Sitting_Down
11	038	Dancing	26	115	Swimming
12	039	Drak_skinned_People	27	117	Telephones
13	041	Demonstration_or_Protest	28	120	Throwing
14	044	Doorway	29	126	Vehicle
15	049	Explosion_Fire	30	127	Walking

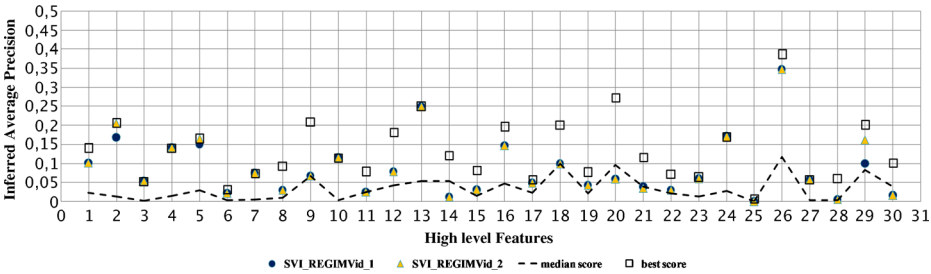

Fig. 4 TRECVID 2010: SVI_REGIMVid evaluations

4.2 Impact of pseudo-sentences on semantic content categorization

To highlight the effectiveness of using the context space for semantic concept detection, we have run more experiments. We extended our interest to deal with other concepts and data sets mainly the Caltech data set.

In Caltech-101, we have used 10 different object classes, notably Airplanes (800), Motorbikes (798), Car_side (123), Sunflower (085), Leopards (200), Ketch (114), Helicopter (88), Schooner (063), Water_lilly (037) and Cougar_body (047). For each class, a third of these images are randomly picked for learning process and the reaming two third for the testing.

We ran experiments for SVI_REGIMVid_1 at TRECVID and Caltech. The obtained results are reported in Tables 3 and 4. Considering the achieved findings, we can make the following comments:

- Our PS approach clearly outperforms the SVM classifiers in terms of inferred average precision classification performance. In fact, for most concepts, the PS process was able to improve the semantic categorization (eg. crowd, car_side concepts) better than multiscale bag-of-words and graphs. For some other concepts such as water_lilly and sunflower the improvement is even clearer and more significant. This improvement is due to the PS codebook model that yields a distribution over codewords that represents the whole image, making this model well-suited for describing the context.
- The experimental results show that, the PS codebook model can be well categorized between the semantic concepts and more particularly between the concepts that are visually similar. In fact, our contribution achieves the best inferred average precision for (Water_lilly, Sunflower), (Cougar_body, Leopard) and (Suburban, Building). This improvement is due to the PS codebook model investigated using the background context. In fact, our method of key-points detection tries to extract significant patches from the object and the background.
- For Caltech 101 subset, the feature graph perform almost as well as BoW, while with TRECVID 2010 subsets BoW are at least 17 % better, because the number of concept categories increases and category size decreases. In addition, the experimental results show that the feature graph performance is closely dependent on the image resolution. For instance, in Caltech 101 the feature graph achieves

Table 3 Concept retrieval performance (infAP in %) of different methods on TRECVID 2010 data sets

Method			
Semantic concepts	SVI_REGIMVid_1	Lazebnik [20]	R.Behmo [4]
Airplane_Flying	**0.102**	0.031	0.023
Animal	**0.169**	0.008	0.012
Asian_People	**0.054**	0.002	0.004
Bicycles	**0.142**	0.070	0.001
Birds	**0.151**	0.042	0.032
Building	**0.022**	0.002	0.003
Car	**0.075**	0.018	0.005
Charts	**0.030**	0.003	0.018
Cheering	0.068	**0.129**	0.064
Cityscape	**0.116**	0.001	0.002
Crowd	**0.025**	0.012	0.013
Dancing	**0.079**	0.030	0.037
Daytime_Outdoor	**0.250**	0.068	0.021
Dogs	0.012	0.045	**0.046**
Entertainment	**0.031**	0.019	0.023
Female_Person	**0.147**	0.077	0.049
Female-Human-Face-Closeup	**0.049**	0.041	0.014
Greeting	0.099	**0.117**	0.066
Ground_Vehicles	**0.043**	0.025	0.012
Motorcycle	0.059	**0.200**	0.114
News_Studio	0.039	**0.055**	0.051
Office	**0.030**	0.028	0.021
Roadway_Junction	**0.060**	0.026	0.011
Shopping_Mall	**0.171**	0.016	0.052
Single_Person	0.000	0.000	**0.005**
Suburban	**0.348**	0.032	0.243
Teenagers	**0.058**	0.003	0.004
Tent	0.006	**0.016**	0.002
Vegetation	0.100	**0.103**	0.055
Vehicle	0.016	0.032	**0.033**

the best inferred average precision for Airplane (infAP=0.35), contrary to TRECVID 2010 dataset, where the inferred average precision for the same concept is 0.023.

4.3 Impact of contextual information on semantic content indexing

In order to emphasize the efficiency of the use of the context space for semantic concept detection, our attention was focused on other sets of high-level features (6 contexts and 11 concepts) defined below in Fig. 5.

Table 4 Concept retrieval performance (infAP in %) of different methods on Caltech 101 data set

Semantic concepts	Method		
	SVI_REGIMVid_1	Lazebnik [20]	R. Behmo [4]
Airplane	0.310	0.151	**0.35**
Motorbike	0.152	0.131	**0.234**
Car_side	**0.182**	0.171	0.174
Sunflower	**0.27**	0.2	0.117
Leopard	**0.229**	0.162	0.078
Kitchen	0.136	**0.17**	0.095
Helicopter	0.272	0.093	**0.275**
Schooner	0.182	**0.2**	0.098
Water_lilly	**0.28**	0.183	0.136
Cougar_body	**0.234**	0.131	0.098

As shown in Fig. 5, the produced hierarchy is a M-ary tree such-structure, where leaf nodes (pink ellipses) are the contexts which are recovered from LSCOM based on the role "Generalization", while their concepts (blue octagons) are integrated based on Experts observations through the role "Includes". The M-ary hierarchy has a crucial role on the O^f evolving.

The obtained results are reported in Table 5 and three illustrative examples are presented in Fig. 6. Relying on the obtained results, we can make the following comments:

- The effectiveness of video indexing systems is clearly improved when a knowledge-based approach is integrated. In fact, when the LSCOM ontology is incorporated, the precision improvement of semantic concept detection is around 11 %. However, we have obtained 21 % via our ontology O^f. This variation is mainly due to the hierarchical roles of each one. The LSCOM ontology, based on "Generalization" roles, provides enrichment only for the concepts of a higher level. However, the O^f ontology expounds other roles such as "IsPartOf", "Includes" and "IsRelatedTo". These allow us to highlight the relationship between a context and its concepts and concept-concept within a target context space. In fact, as shown in the first example of Fig. 6, the role "Includes" has reduced the detection score of the concept "Mountain" however the detection score of the concept "Tree" has

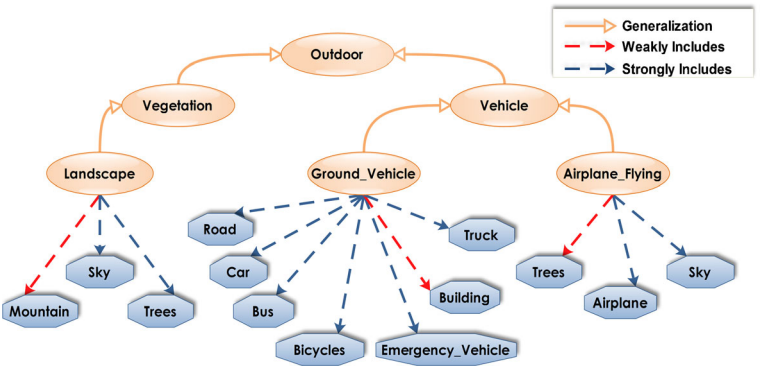

Fig. 5 Partial view of concepts/contexts relationships

Table 5 Concept retrieval performance (Inferred Average Precision infAP, Precision P and Recall R) for different Concept detection methodologies applied on TRECVID 2010 data set

Concept detection methodologies										
High level features	SVI_REGIMVid_1			SVI_REGIMVid_2 with LSCOM			SVI_REGIMVid_2 with O^f		CBCF with SVM	
	InfAP	P	R	infAP	P	R	P	R	P	R
Outdoor	–	0.52	0.59	–	0.88	0.77	**0.9**	**0.82**	0.42	0.71
Vegetation	0.1	0.74	0.68	0.1	0.74	0.68	**0.93**	**0.87**	0.56	0.7
Landscape	–	0.6	0.79	–	0.6	0.79	**0.7**	**0.82**	0.52	0.8
Sky	–	0.66	0.9	–	0.66	0.9	**0.85**	**0.95**	0.66	0.9
Trees	–	0.62	0.72	–	0.62	0.72	**0.73**	**0.82**	0.62	0.72
Mountain	–	0.68	0.8	–	0.68	0.8	**0.83**	**0.85**	0.68	0.8
Vehicle	0.016	0.22	0.5	**0.103**	0.7	0.77	**0.72**	**0.78**	0.32	0.6
Ground_Vehicle	0.043	0.3	0.66	**0.18**	0.6	0.73	**0.69**	**0.75**	0.4	0.69
Road	–	0.43	0.6	–	0.43	0.6	**0.88**	**0.9**	0.43	0.6
Car	0.075	0.42	0.64	**0.17**	0.58	0.73	**0.79**	**0.83**	0.42	0.64
Bus	–	0.52	0.73	–	0.52	0.73	0.52	0.73	0.52	0.73
Bicycles	0.142	0.67	0.92	**0.185**	0.82	**0.97**	**0.83**	**0.97**	0.67	0.92
Emergency Vehicle	–	0.9	0.83	–	0.9	0.83	0.9	0.83	0.9	0.83
Building	0.022	0.18	0.22	**0.1**	0.5	0.43	**0.55**	**0.45**	0.18	0.22
Truck	–	0.35	0.37	–	0.35	0.37	0.35	0.37	0.35	0.37
Airplane flying	0.102	0.8	0.78	0.102	0.8	0.78	**0.83**	**0.79**	0.82	**0.79**
Airplane	–	0.5	0.6	–	0.5	0.6	**0.71**	**0.69**	0.5	0.6
Total:	0.071	0.53	0.66	**0.134**	0.64	0.71	**0.74**	**0.77**	0.52	0.68

improved due to "Landscape Strong Includes Tree" and "Landscape Weak Includes Mountain". In addition, the role "IsRelatedTo" has performed the detection of "Sky". Moreover, we can note the effectiveness of the "IsPartOf" role to enhance the detection of contexts within the third example.

- The proposed approach improves not only the precision of contexts detection, but also concepts detection. In fact, our ontology O^f performs the best precision scores for 16 (6 context and 10 concepts) out of 17 high level features. This result is rather obvious: the proposed ontology O^f tries to represent the context space; with 4 roles ("Generalization", "IsPartOf", "Includes" and "IsRelatedTo"); using an Abduction Engine [9]. It automatically generates fuzzy rules and optimizes them. These fuzzy rules, that represent the ground truth, further improve the effectiveness of video indexing systems.
- The context-based concept fusion framework enhances the high level feature detection. In fact, the recall is improved for 6 (Outdoor, Vegetation, Landscape, Vehicle, Ground_Vehicle, Airplane-Flying) out of 17 high level feature. We can see that the enrichment has only targeted the context. Despite this recall improvement (about 2 %), precision has declined. The origin of this degradation is its requirement of an explicit knowledge about contextual space that is provided manually based on intuition and human knowledge. In addition when a detector provides inaccurate probability (e.g. Sky, truck, bus, road), the effectiveness is low (the precision improvement of Vehicle is 10 %). In these fields, O^f reached a rate of about 50 %.

		
	Groundtruth :	
Bicycles – Tree – Landscape – Sky - Vegetation - Outdoor	**Bus – Road - Ground_Vehicles – Vehicles - Outdoor**	**Airplane – Sky – Airplane-Flying – Vehicle - Outdoor**
	Concept detectors output :	
Mountain : 0.534 – Landscape : 0.821 – Tree : 0.712 – Vegetation :0.9	Bus : 0.633 – Road : 0.715 - Ground_Vehicles : 0.82 – Landscape : 0.22	Airplane : 0.812 – Sky :0.971 – Bird :0.12
	Includes Role	
Mountain : 0.124 – Landscape : 0.821 – **Tree : 0.789** – Vegetation :0.9	Bus : 0.621 – Road : 0.715 - Ground_Vehicles : 0.82 – ~~**Landscape**~~	Airplane : 0.812 – Sky : 0.971 – ~~**Bird**~~
	IsRelatedTo Role	
Mountain : 0.124 – Landscape : 0.821 – Tree :0.789 – **Sky : 0.802** - Vegetation :0.9	**Bus : 0.69** – Road : 0.715 - Ground_Vehicles : 0.82	Airplane : 0.812 – Sky : 0.971
	IsPartOf Role	
Mountain : 0.124 – Landscape : 0.821 – Tree : 0.789 – Sky : 0.802 - Vegetation :0.9	Bus : 0.69 – Road : 0.715 - Ground_Vehicles : 0.82	Airplane : 0.812 – Sky : 0.971 – **Airplane-Flying : 0.9531**
	Generalization Role	
Mountain : 0.124 – Landscape : 0.821 – Tree : 0.789 – Sky : 0.802 – Vegetation : 0.9 – **Outdoor : 0.9**	Bus : 0.69 – Road : 0.715 - Ground_Vehicles : 0.82 **– Vehicles : 0.82 – Outdoor : 0.82**	Airplane : 0.812 – Sky : 0.971 – Airplane-Flying : 0.9531 – **Vehicle : 0.9531 - Outdoor : 0.9531**

Fig. 6 Illustrative examples of semantic concepts detection based on SVI_REGIMVid

4.4 SVI_REGIMVid scalability

In this section, we address the problem of SVI_REGIMVid scalability. We focus on the scalability of the indexing process through a large set of concepts detection, on the one hand, and the scalability of the refinement process through a large number of relationships, on the other hand.

The semantic concepts/contexts detection process is performed on all video shots individually and independently. The average execution times per key-frames for this process within TRECVID 2010 benchmark (146788 key-frame) took about 22.235 s on using a WINDOWS PC with 980 MHZ INTEL(R) CORE(TM) 2 CPU and 2 GB RAM. This large computing time versus the semantic Pathfinder [46] (about 1 s), invites us to parallelize the concept detection process or to use a GPU computing. However, within the refinement process, the fuzzy role are inferred simultaneously. To evaluate the scalability of the refinement process, we plot the execution times versus the number of fuzzy rules inferred for all key-frames of the TRECVID

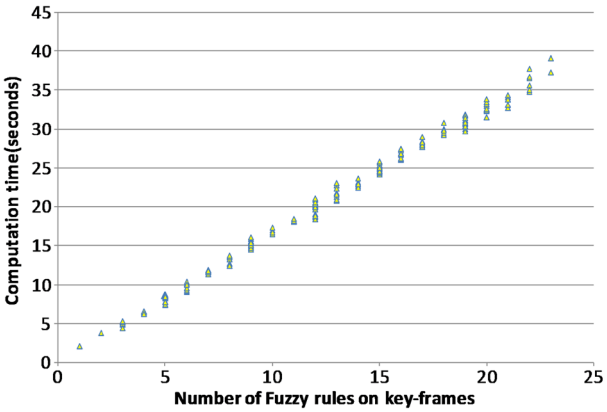

Fig. 7 Computation time versus the number of fuzzy rules on TRECVID 2010 data sets

2010 benchmarks in Fig. 7. We note that, the number of the fuzzy rules defined in our ontology O^f for the 130 high-level features is 2 876 rules. However the maximal number of the inferred rules for each key-frame is 23.

As shown in the Fig. 7, the computational time is approximately linear to the number of the inferred fuzzy rules in all visual content. Thus, we can confirm that the refinement process is scalable.

5 Conclusion and future works

In this paper we presented the semantic video indexing system of REGIMVid group for semantic access to multimedia archives, called SVI_REGIMVid. The SVI_REGIMVid is a generic approach for video indexing. It exploits the spatial information in visual content categorization and the contextual knowledge in concept detection enrichment to bridge the semantic gap. Its effectiveness in terms of precision and recall was proved on diverse concepts and data sets.

Future works will consist in improving the indexing process scalability, integrating the spatial coherence between the context space with the exploration of cross-concept correlation and inter-shot dependency.

Acknowledgments The authors would like to acknowledge the financial support of this work by grants from General Direction of Scientific Research (DGRST), Tunisia, under the ARUB program. Also, the authors are grateful to NIST and the TRECVID coordinators for the benchmark organization's effort.

References

1. Amir A, Berg M, fu Chang S, Iyengar G, yung Lin C, (paul Natsev A, Neti C, Nock H, Naphade M, Hsu W, Smith JR, Tseng B, Wu Y, Zhang D, Watson ITJ (2003) Ibm research trecvid-2003 video retrieval system. In: In NIST TRECVID-2003

2. Amores J, Sebe N, Radeva P (2007) Context-based object-class recognition and retrieval by generalized correlograms. IEEE Trans Pattern Anal Mach Intell 29:1818–1833
3. Ayache S, Quenot G (2008) Video Corpus annotation using active learning. In: European Conference on Information Retrieval (ECIR), pp. 187–198. Glasgow, Scotland
4. Behmo R, Paragios N, Prinet V (2008) Graph commute times for image representation. In: CVPR
5. Bres S, Jolion JM (1999) Detection of interest points for image indexation. In: VISUAL, pp. 427–434
6. Crandall D, Felzenszwalb P, Huttenlocher D (2005) Spatial priors for part-based recognition using statistical models. In: Proceedings of the 2005 I.E. Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1 - Volume 01, CVPR'05, pp. 10–17. IEEE Computer Society, Washington, DC, USA
7. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L (2009) ImageNet: a large-scale hierarchical image database
8. Elleuch N, Ben Animar A, Alimi AM (2010) A generic system for semantic video indexing by visual concept. In: I/V Communications and Mobile Network (ISVC), 2010 5th International Symposium on, pp. 1–4. IEEE
9. Elleuch N, Zarka M, Ammar AB, Alimi AM (2011) A fuzzy ontology: based framework for reasoning in visual video content analysis and indexing. In: Proceedings of the Eleventh International Workshop on Multimedia Data Mining, p. 1. ACM
10. Elleuch N, Zarka M, Feki I, Ammar AB, Alimi AM (2010) Regimvid at trecvid 2010: Semantic indexing. In: TRECVID 2010. Citeseer
11. Gemert JC, Geusebroek JM, Veenman CJ, Smeulders AW (2008) Kernel codebooks for scene categorization. In: Proceedings of the 10th European Conference on Computer Vision: Part III, ECCV'08, pp. 696–709. Springer-Verlag, Berlin, Heidelberg
12. Grauman K, Darrell T (2005) Efficient image matching with distributions of local invariant features. In: Proceedings of the 2005 I.E. Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 2 - Volume 02, CVPR'05, pp. 627–634. IEEE Computer Society, Washington, DC, USA
13. Hauptmann A, Baron RV, y. Chen M, Christel M, Duygulu P, Huang C, Jin R, h. Lin W, Ng T, Moraveji N, Snoek CGM, Tzanetakis G, Yang J, Yan R, Wactlar HD (2003) Informedia at trecvid 2003: analyzing and searching broadcast news video. In: Proc. Of TRECVID
14. Hauptmann AG, Yan R, Qi Y, Jin R, Christel MG, Derthick M, yu Chen M, Baron RV, Lin WH, Ng TD (2002) Video classification and retrieval with the informedia digital video library system. In: TREC
15. Helmer S, Helmer (2004) CS object recognition with many local features. In: In Workshop on Generative Model Based Vision (GMBV)
16. Jiang W, Chang SF, Loui AC (2006) Active context-based concept fusion with partial user labels. In: IEEE International Conference on Image Processing (ICIP 06). Atlanta, GA, USA
17. Jiang YG, Wang J, Chang SF, Ngo CW (2009) Domain adaptive semantic diffusion for large scale context-based video annotation. In: International Conference on Computer Vision (ICCV). Kyoto, Japan
18. Jurie F, Triggs B (2005) Creating efficient codebooks for visual recognition. In: Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1 – Volume 01, ICCV'05, pp. 604–610. IEEE Computer Society, Washington, DC, USA
19. Karray H, Wali A, Elleuch N, Ammar AB, Ellouze M, Feki I, Alimi AM (2008) Regim at trecvid2008: high-level features extraction and video search. In: TRECVID
20. Kennedy LS, Chang SF (2007) A reranking approach for context-based concept fusion in video indexing and retrieval. In: Proceedings of the 6th ACM international conference on Image and video retrieval, CIVR'07, pp. 333–340. ACM, New York, NY, USA
21. Kohonen T, Schroeder MR, Huang TS (eds) (2001) Self-organizing maps, 3rd edn. Springer-Verlag New York, Inc., Secaucus
22. Ksibi A, Elleuch N, Ammar AB, Alimi AM (2011) Semi-automatic soft collaborative annotation for semantic video indexing. In: EUROCON, pp. 1–6
23. Laptev I, Marszalek M, Schmid C, Rozenfeld B (2008) Learning realistic human actions from movies. In: CVPR
24. Lazebnik S, Schmid C, Ponce J (2006) Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: CVPR (2), pp. 2169–2178
25. Lin CY, Tseng BL, Smith JR (2003) Video collaborative annotation forum: establishing ground-truth labels on large multimedia datasets. In: Proceedings of the TRECVID 2003 Workshop
26. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vis 60(2):91–110
27. Marszałek M, Schmid C (2006) Spatial weighting for bag-of-features. In: IEEE Conference on Computer Vision & Pattern Recognition, CVPR 2006, June, 2006, vol. 2, pp. 2118–2125. IEEE, New York, NY, Etats-Unis
28. Matas J, Chum O, Urban M, Pajdla T (2002) Robust wide baseline stereo from maximally stable extremal regions. In: BMVC, pp. 1–10
29. Mikolajczyk K, Schmid C (2004) Scale & affine invariant interest point detectors. Int J Comput Vis 60(1): 63–86

30. Mikolajczyk K, Tuytelaars T, Schmid C, Zisserman A, Matas J, Schaffalitzky F, Kadir T, Gool LV (2005) A comparison of affine region detectors. Int J Comput Vis 65(1–2):43–72
31. Miller GA (1995) Wordnet: a lexical database for english. Commun ACM 38(11):39–41
32. Murphy P, Torralba A, Freeman W (2003) Using the forest to see the trees: a graphical model relating features, objects and scenes. In: Adv. Neur. Inform. Process. Syst. 16 (NIPS). MIT Press, Cambridge
33. Mylonas P, Athanasiadis T, Wallace M, Avrithis YS, Kollias SD (2008) Semantic representation of multimedia content: knowledge representation and semantic indexing. Multimedia Tools Appl 39(3):293–327
34. Mylonas Ph, Avrithis Y (2005) Context modelling for multimedia analysis. In: Proc. 5th Int. Interdisciplinary Conf. Modeling and Using Context (CONTEXT'05), Paris, France, Jul. 2005
35. Mylonas P, Spyrou E, Avrithis Y, Kollias S (2009) Using visual context and region semantics for high-level concept detection. Trans Multi 11(2):229–243
36. Naphade M, Smith JR, Tesic J, Chang SF, Hsu W, Kennedy L, Hauptmann A, Curtis J (2006) Large-scale concept ontology for multimedia. IEEE Multimedia 13(3):86–91
37. Natsev A, Tesic J, Xie L, Yan R, Smith JR (2007) Ibm multimedia search and retrieval system. In: CIVR, p. 645
38. Ngo CW, Zhu S, Tan HK, Zhao W, Wei XY (2010) Vireo at trecvid 2010: semantic indexing, known-item search, and content-based copy detection. In: TRECVID
39. Petersohn C (2004) Fraunhofer hhi at trecvid 2004: shot boundary detection system. In: Proc. TRECVID Workshop
40. Punitha P, Guru DS, Vikram TN "Proceedings of the international conference on cognition and recognition relative spatial distance matrix: a novel and invariant data structure for representation and retrieval of exact match symbolic images"
41. Rubner Y, Tomasi C, Guibas LJ (2000) The earth mover's distance as a metric for image retrieval. Int J Comput Vis 40(2):99–121
42. Schmid C, Mohr R, Bauckhage C (2000) Evaluation of interest point detectors. Int J Comput Vis 37(2):151–172
43. Smeaton AF, Over P, Kraaij W (2006) Evaluation campaigns and trecvid. In: MIR'06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval, pp. 321–330. ACM Press, New York, NY, USA
44. Smeulders AWM, Worring M, Santini S, Gupta A, Jain R (2000) Content-based image retrieval at the end of the early years. IEEE Trans Pattern Anal Mach Intell 22(12):1349–1380
45. Smith J, Naphade M, Natsev A (2003) Multimedia semantic indexing using model vectors. In: Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on, vol. 2, pp. II–445–8 vol.2
46. Snoek CGM, Worring M, Geusebroek JM, Koelma DC, Seinstra FJ, Smeulders AWM (2006) The semantic pathfinder: using an authoring metaphor for generic multimedia indexing. IEEE Trans Pattern Anal Mach Intell 28(10):1678–1689
47. Snoek C, Worring M, Hauptmann AG (2006) Learning rich semantics from news video archives by style analysis. TOMCCAP 2(2):91–108
48. Snoek CGM, Worring M, Smeulders AWM (2005) Early versus late fusion in semantic video analysis. In: Proceedings of the 13th annual ACM international conference on Multimedia, MULTIMEDIA'05, pp. 399–402. ACM, New York, NY, USA
49. Spyrou E, Mylonas P, Avrithis Y (2008) A visual context ontology for multimedia high-level concept detection. In: Kofod-Petersen A, Cassens J, Leake D, Zacarias M (eds) HCP-2008 Proceedings, PArt II, MRC 2008—Fifth InternationalWorkshop on Modelling and Reasoning in Context, pp. 25–36. TELECOM Bretagne
50. Torralba A (2005) Contextual influences on saliency. In: Neurobiology of attention. Academic, London
51. Vesanto J, Alhoniemi E (2000) Clustering of the self-organizing map
52. Volkmer T, Smith JR, Natsev AP (2005) A web-based system for collaborative annotation of large image and video collections: an evaluation and user study. In: Proceedings of the 13th annual ACM international conference on Multimedia, MULTIMEDIA'05, pp. 892–901. ACM, New York, NY, USA
53. Wang L, Song D, Elyan E (2012) Improving bag-of-visual-words model with spatial-temporal correlation for video retrieval. In: Proceedings of the 21st ACM International Conference on Information and Knowledge Management, CIKM'12, pp. 1303–1312. ACM, New York, NY, USA
54. Wei XY, Jiang YG, Ngo CW (2009) Exploring inter-concept relationship with context space for semantic video indexing. In: Proceedings of the ACM International Conference on Image and Video Retrieval, CIVR'09, pp. 15:1–15:8. ACM, New York, NY, USA
55. Wu L, Hu Y, Li M, Yu N, Hua XS (2009) Scale-invariant visual language modeling for object categorization. Trans Multi 11(2):286–294

56. Zampoglou M, Papadimitriou T, Diamantaras KI (2010) From low-level features to semantic classes: spatial and temporal descriptors for video indexing. Signal Proc Syst 61(1):75–83
57. Zha ZJ, Mei T, Hua XS, Qi GJ, Wang Z (2007) Refining video annotation by exploiting pairwise concurrent relation. In: Proceedings of the 15th international conference on Multimedia, MULTIMEDIA'07, pp. 345–348. ACM, New York, NY, USA

Nizar Elleuch received the Engineering degree in computer science (2004) from the National School of Engineers of Sfax in the University of Sfax (Tunisia) and the Master Degree in automation and industrial computing (2006) from the National School of Engineers of Sfax in the University of Sfax (Tunisia).

He is currently a PhD student at the National School of Engineering of Sfax. He is also an assistant in the Higher Institute of Biotechnology of Monastir and a research member in the Research Groups of Intelligent Machine (REGIM-Lab.). His research interests include multimedia indexing and knowledge management.

Anis Ben Ammar is assistant professor in Higher Institute of Business Administration of Sfax, Tunisia. He is Member of the IEEE computer society. He obtained a master degree and then a PhD both in Computer 1999 and 2003 in Paul Sabatier-FRANCE University.

Anis Ben Ammar is a member of REGIM Laboratoy (REsearch Groups on Intelligent Machines).

His research interests include applications of intelligent Multimedia processing including patter recognition, audiovisual analysis, indexing and retrieval.

He focuses his research on information retrieval system especially for social WEB.

Adel M. Alimi (S'91, M'96, SM'00) He graduated in Electrical Engineering in 1990. He obtained a PhD and then an HDR both in Electrical & Computer Engineering in 1995 and 2000 respectively. He is full Professor in Electrical Engineering at the University of Sfax since 2006.

Prof. Alimi is founder and director of the REGIM-Lab. on intelligent Machines. He published more than 300 papers in international indexed journals and conferences, and 20 chapters in edited scientific books. His research interests include applications of intelligent methods (neural networks, fuzzy logic, evolutionary algorithms) to pattern recognition, robotic systems, vision systems, and industrial processes. He focuses his research on intelligent pattern recognition, learning, analysis and intelligent control of large scale complex systems. He was the advisor of 24 Ph.D. thesis. He is the holder of 15 Tunisian patents. He managed funds for 16 international scientific projects.

Prof. Alimi served as associate editor and member of the editorial board of many international scientific journals (e.g. "IEEE Trans. Fuzzy Systems", "Pattern Recognition Letters", "NeuroComputing", "Neural Processing Letters", "International Journal of Image and Graphics", "Neural Computing and Applications", "International Journal of Robotics and Automation", "International Journal of Systems Science", etc.). He was guest editor of several special issues of international journals (e.g. Fuzzy Sets & Systems, Soft Computing, Journal of Decision Systems, Integrated Computer Aided Engineering, Systems Analysis Modelling and Simulations). He organized many International Conferences ISI'12, NGNS'11, ROBOCOMP'11&10, LOGISTIQUA'11, ACIDCA-ICMI'05, SCS'04ACIDCA'2000.

Prof. Alimi has been awarded with the IEEE Outstanding Branch Counselor Award for the IEEE ENIS Student Branch in 2011, with the Tunisian Presidency Award for Scientific Research and Technology in 2010, with the IEEE Certificate Appreciation for contributions as Chair of the Tunisia Computational Intelligence Society Chapter in 2010 and 2009, with the IEEE Certificate of Appreciation for contributions as Chair of the Tunisia Aerospace and Electronic Systems Society Chapter in 2009, with the IEEE Certificate of Appreciation for contributions as Chair of the Tunisia Systems, Man, and Cybernetics Society Chapter in 2009, with the IEEE Outstanding Award for the establishment project of the Tunisia Section in 2008, with the International Neural Network Society (INNS) Certificate of Recognition for contribution on Neural Networks in 2008, with the Tunisian National Order of Merit, at the title of the Education and Science Sector in 2006, with the IEEE Certificate of Appreciation and Recognition of contribution towards establishing IEEE Tunisia Section in 2001 and 2000.

He is the Founder and Chair of many IEEE Chapters in Tunisia section. He is IEEE CIS ECTC Education TF Chair (since 2011), IEEE Sfax Subsection Chair (since 2011), IEEE Systems, Man, and Cybernetics Society Tunisia Chapter Chair (since 2011), IEEE Computer Society Tunisia Chapter Chair (since 2010), IEEE ENIS Student Branch Counselor (since 2010), He served also as Expert evaluator for the European Agency for Research. since 2009.