



REGIMVID at TRECVID 2010: Semantic Indexing


Nizar ELLEUCH, Mohamed ZARKA, Issam FEKI, Anis BEN AMMAR, Adel M. ALIMI
REGIM: Research Group in Intelligent Machines, ENIS, Tunisia
{elleuch.nizar, mohamed.ezzarka, feki_issam, anis.benammar, adel.alimi} @ieee.org
<http://www.regim.org>

Description of Submitted Runs

Semantic Indexing

 Regim_4: The indexing process is based on the visual modality analysis and relationships within LSCOM Ontology to improve the detection of large set of semantic concepts. The visual modality analysis is orientated towards an automatic categorization of video contents to create relevance relationships between low-level descriptions and semantic contents according to a user point of view.

 Regim_5: This indexing system is based on Multimodal fuzzy fusion using positive rules extracted from LSCOM Ontology. The fusion process employs both a deduction and abduction reasoning engines.

 Regim_6: This indexing system is based on Multimodal fuzzy fusion using positive and negative rules extracted from LSCOM Ontology. The fusion process employs also both a deduction and abduction reasoning engines.

Abstract

In this paper, we describe an overview of a software platform that has been developed within REGIMVid project for TRECVID 2010 video retrieval experiments. The REGIMVID team participated in Semantic Indexing task. In TRECVID 2010, we explore several novel techniques to perform the detection of semantic concepts, including multi classifiers with supervised learning process, discriminative feature representation based on local keypoints, visual concept fusion using LSCOM rules, and also multimodal concept fusion using LSCOM Ontology.

Keywords

Classification; Concepts; Data fusion model; Multimodal; Semantic; Signal.

1 INTRODUCTION

The images categorization and the objects detection in large collection of videos / images is a very challenging problem for many applications as video / image content-based retrieval, semantic indexing, history segmentation, video browsing and annotation. In fact, the vast amount of audiovisual data currently represents the main challenge of TREC Video Retrieval Evaluation (TRECVID) on the information processing particularly in representing and semantic indexing.

The goal of the TRECVID 2010 semantic indexing task is to promote research on methods for indexing of large number of semantic concepts (130) with a number of generic-specific relationships among them by using LSCOM Ontology from a new set of videos (IACC.1) which represents the truth of ground such as "web video". The IACC.1 collection is characterized by a high degree of diversity in creator, content, style, production qualities, original collection device/encoding, language, etc.

Furthermore, it is acknowledged that real progress in addressing this challenging task requires key advances in many complementary research areas such as scalable coding of each modality from audiovisual contents and its metadata, fuzzy fusion of each coding and database technology. Thus, the REGIMVid project integrates many of these issues. Therefore, a key effort has been integrated within the project, at this session, which is focuses on increasing the robustness of our semantic indexing scheme by employing a semantic representation of semantic concepts and analyzing the relationships between them.

Traditionally, the concept detection is most often carried out using classifiers or networks of classifiers [1, 2, 3] to build models. These classifiers typically use a supervised learning which consists in training a system from sets of positive and negative examples. So, the performance of semantic indexing systems depends a lot upon the implementation choices and details on one hand, and, on the other hand, it strongly depends upon the size and quality of the training examples.

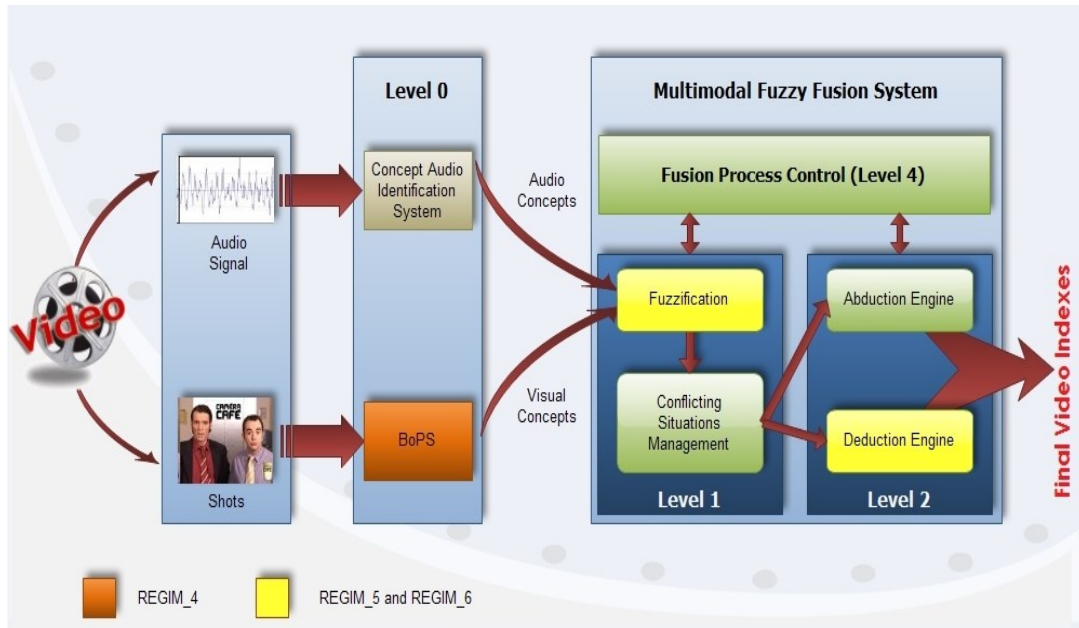


Fig 1. Overview of the REGIMVID toolbox for video input.

To achieve this goal, many requirements are taken into account for our platform of semantic indexing. They can be summarized as follows:

- Provide an extraction of subset of efficient local and global features.
- Employ a semantic structure which can represent and discriminate objectively between the concepts.
- Use an ontology to refine and generalize the detection of others concepts that are semantically dependent.
- Classify the training examples considering to their semantic relevance according to a user point of view.
- Use a multiple training of Intelligent Machines to build semantic models for each semantic concept.
- Use an ontology to generalize multimodal concept detection.

In this paper, we describe the version of our platform developed within the framework in TRECVID 2010. The paper is organized as follows; in section 2, we present a general overview of the toolbox. It includes a description of the architecture, visual concepts indexing, audio concepts indexing and multimodal scheme. Section 3 describes the obtained experimental results via the TRECVID benchmark. Finally, we describe our future plans for both the extension of the toolbox and its use in different scenarios.

2 REGIMVID: MULTIMODAL INDEXING SYSTEM

In this section, we present an overview of the multimodal indexing system structure and we describe its different sub-systems.

2.1 Architecture

The main objective of our system is to provide an automatic analysis of video contents by using frame description based on low-level features. Thus, the system extracts, first, the low-level features for each modality of the video shot and then, represents their contents for labeling them, later, by basing on score detection via classification process. After that, the predicted score are merged to obtain multimodal fusion [23]. The fusion architecture, based on the JDL/DFS model [4], is shown in figure 1. After the extraction of visual and audio semantic concept (Level 0), the Level 1 is called to search and to eliminate conflicting situations. Then, Level 2 is applied on Level 1 results aiming at finding further semantic concepts based on abduction and deduction engines. These engines analyses relationships between concepts. Finally, the Level 4 is used to control the whole fusion process.

Below is the description of the different Levels of the proposed fusion architecture.

2.2 Level 0 : unimodal semantic interpretation

2.2.1 Visual analysis

The visual video indexing focuses on a automatic concepts categorization based on a pseudo-sentences representation with a co-training scheme. Thus, for each key frame, we extract the salient point, for which the low-level visual features are extracted, like Color Histogram, wavelet, Gabor, etc. After combining together in an "early fusion" manner, a visual elementary word vocabulary is firstly constructed through clustering of these features by using Self-Organizing Maps (SOM). Based on these elementary words, we group them together to yield visual pseudo-sentences which are more informative words. Indeed, an image is considered a set of patterns, thus we assumed that each visual word is spatially dependent on others to form a given pattern.

Based on these visual pseudo-sentences, each key frame is described by one vector according to the occurrence of each visual pseudo-sentence. Then, we process a multiple classifiers with supervised learning process using SVM classifier for each model.

a. Pre-processing

In the context of the TRECVID campaigns, the MRIM (Multimedia Information Modeling and Retrieval) team has provided a web-based collaborative annotation tool to enhance annotation process. It is focuses on the description of images contents with a single semantic label. Unfortunately, when the images is annotated with a single label involve the same semantic meaning, however the contents are not often homogeneous. For example, Figure 1 shows the four images that contain the same label "sky", but their semantic contents are very different – blue sky, cloud or cloudless, sunset, night, and sky with more dominants others concepts (beach, building). Therefore, the positive examples annotated "somewhat relevant" are the most present in the training set. They are considered by the classifier the most important examples. Thus, the classifier built an inefficient model because it is influenced by the quality of the training samples. Consequently, the classification using this model generates errors.



Fig 2. Different image contents with label "sky".

Obviously, it must be more complex if many kinds of labels are mixed. That is the main reason that we propose to aggregate the training data at three relevance levels or classes, namely "highly relevant" (TP), "relevant" (P) and "somewhat relevant" (PP). In fact, this strategy allows a better capture the user's subjective perception about the relevance of the image and will reduce errors generated by classification process.

b. Detection of interest points

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 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 [5].

Our approach joins with the last purpose [6]. The main idea is to exploit a detector based on luminance and variation of the orientation of edge. As luminance variations occur at almost every scale in an image, we must 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.

c. Visual features extraction

The aim of feature extraction is to derive a compact descriptive, and representing the pattern of interest. 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. The relative performance of the specific features within a given feature is shown to be consistent across all semantic concepts. However, the relative importance of one feature modality vs. another may change from one concept to the other. In fact, we use several visual descriptors of different modalities (color, texture and shape) as Color Histogram, Co-occurrence Texture, Gabor, etc [7].

After extracting the visual features, we proceed to the early fusion step. For this, we combine a selection of the feature vectors resulting from visual feature extraction. We adopt the method proposed in [8], using vector concatenation to unify the features representation. After feature normalization, we obtain 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.

d. Elementary CodeBook construction through Kohonen maps

One of the most important constraints of discrete visual codebook generation is in the uniform distribution of visual words over the continuous high-dimensional feature space. Self-Organizing Maps (SOM) proved their performance in doing that [9, 10]. 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 [11]. Thus, to generate a codebook of prototype vectors from the above features, we utilize the SOM-based clustering.

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 by using of hierarchical agglomerative clustering and partitive clustering using K-means introduced by J. Vesanto and E. Alhoniemi [12] as it allows very fast clustering with an acceptable accuracy.

e. Spatial Information representation: Bag of Pseudo-Sentences (BoPS)

To represent the image content, the most common way is to use the bags of words (BoW). This BoW representation cannot describe objectively and discriminately the content of an image and also neglects the spatial distribution of relevant areas. So, some works [1] have been interested in spatial distribution of key-points to enhance the classification process and concepts categorization. For integrating such information, we group together the elementary words to yield visual pseudo-sentences which are more informative words. To generate these pseudo-sentences, we used only two stages of spatial clustering based on the Relative Euclidean Distance (RED) calculated between each visual elementary word in each image. Two visual elementary words are spatially dependent if their distance RED is minimal compared to all other distances with the other visual elementary words. Thus, they could be subsequently merged. The result of two merger stages is a set of visual pseudo-sentences, each of which containing at most 9 elementary words (or 7, 6, 5, 4, 3, 2 and 1) with the smallest RED. After this, for each visual pseudo-sentence, we project its different elementary visual words on an axis Δ to identify the sequence of visual words: pseudo-sentence. Δ is the linear regression of the Cartesian coordinates of elementary visual words forming each pseudo-sentence.

Finally, our method allows generating a new codebook which represents pseudo-sentences. 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 classifier much more important. Therefore, we perform a refinement step to reduce the size of the obtained pseudo-sentences codebook.

The refinement process is likened to a problem of optimization of the pseudo-sentences construction. To resolve this problem two steps are considered: the analysis of syntax and the occurrence of all constructed pseudo-sentences, and the subdivision of pseudo-sentences having a low occurrence (for more details, the paper "Semantic Categorization Concepts Based on Pseudo-Sentences Representation" will appear) [24]. Therefore, each image is represented by one vector presenting the occurrence of each pseudo-sentence defined in the visual pseudo-sentences codebook.

f. Sample learning filtering

The purpose of this step is to check the robustness of our approach in clustering of the same semantic contents and discrimination between various semantic contents. In addition, we eliminate certain representation in order to further enhancing the classification process.

To achieve this goal, we use also Self-Organizing Maps (SOM) which is composed of 390 units (130 semantic concepts and each concept is composed by 3 classes). Then, the SOM map is trained on the representation, based on visual pseudo-sentences, of all training database images with respect to all semantic concepts. After the training process, we analyze the set of vectors assigned to each unit of the map. Generally, each set must be formed by the same semantic entity. However, there are some failures. There are two types of failures. The first is when a feature vector belongs to the same semantic class, but at another relevance levels. In this case, we change the semantic tag of the image. The second is when a feature vector does not belong to its semantic class. In this case, we remove the image from the collection.

g. Concept learning

The classification plays an important role for bridging the semantic gap. In our work, we use the LIBSVM implementation (for more detail 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 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 consider the examples of "highly relevant", "relevant" and "irrelevant" 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 learned for each repartition to build concepts models. So, for each concept, three classifiers are paralleled 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 images from the visual concept to learn and provides the model. This algorithm needs to find later the most discriminative information to represent concepts. 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 for each test sample. To map the SVM binary outputs into probabilities, we use Platt's method that produces probabilistic output using a sigmoid function:

$$P_{A,B}(f) = \frac{1}{1 + \exp(Af + B)}, \text{ where } f = f(x) \quad (1)$$

Where A and B are estimated by training using the maximum likelihood estimation from the training data set. 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 formula:

$$C = \alpha * C_{ip} + \beta * C_{ip+p} + \gamma * C_{ip+p+pp} \quad (2)$$

Where α , β and γ are predefined by annotator. These averages are then ordered to select the examples.

2.2.2 Audio analysis

The audio indexing approach consists of three steps. After pre-processing segmentation and removal of silence segments, the separation modules built acoustic speech, music and environmental sounds classes. The feature extraction block makes a characteristic vector of the audio signal that is sent to the learning and classification modules for the recognition of different audio concepts [22].

a. Pre-processing modules

The segmentation modules in our system aims to separate heterogeneous input audio into speech, music and environmental sound regions. Initial segmentation is achieved in the time domain based on silence detection. The audio signal is sampled at 22050 Hz and 16 bits is ample. The audio stream is segmented into clips that are 3 seconds long with 1 second overlapping with the previous ones. Each clip is then divided into frames that are 512 samples long and are shifted by 256 samples from the previous frames. For each frame we extract ZCR (Zero Crossing Rate) feature value [19], set that will be used to determine a silence segments. These segments are automatically eliminated because they are not part of our indexing. A merge module of no silence segments remaining runs for the preparation to a new segmentation, oriented to the detection of speech and music classes of the audio stream obtained.

b. Acoustic sources separation modules

A two-step scheme is proposed to classify audio clips into one of three audio classes: speech, music, and environment sound. First, the input audio segments are separated into speech and non-speech segments by two features: Low short time energy ratio (LSTER) and Spectrum flux (SF). LSTER is defined as the ratio of the number of frames who's STE is less than 0.5 times of average short time energy in a one second window. LSTER is an effective feature, especially for discriminate speech and other no speech signals [19]. In general, there are more silence frames in speech, so the LSTER measure will be much higher for speech than that for no speech segments. Spectrum Flux (SF) is defined as the average variation value of spectrum between the adjacent two frames in one second window [19]. In our experiments, we found that, in general, the SF values of speech are higher than those of all other no speech segments. Second, no speech segments are further classified into music and environmental sound, by a Ban Periodicity feature (BP). Band periodicity is defined as the periodicity of each sub-band. It can be derived by sub-band correlation analysis. It is observed that the music band periodicities are in general much higher than those of environment sound [20].

This two-step scheme is suitable for different application, and it can achieve high classification accuracy. Fig 3 shows the two steps of acoustic sources separation module. Then, the segmentation of an audio stream can be got, by using these classification results.

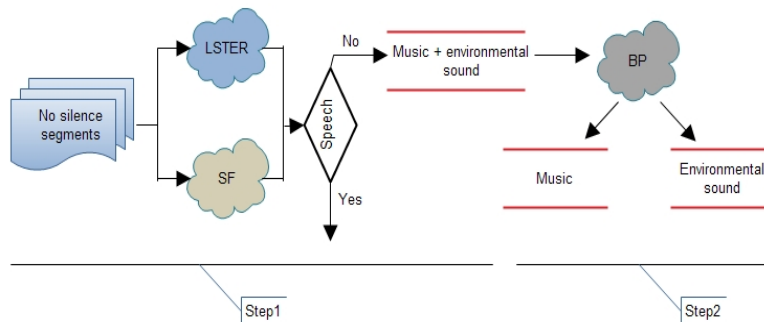


Fig 3. Acoustic sources separation module.

In extracting audio features in our classification scheme, whatever the sample rate of input signal could be, we all down sample it into 16KHz sample rate and then segment it into sub segments by one-second window. This one-second audio clip is taken as the basic classification unit in method. It is further divided into forty 20ms frames. Each feature is extracted based on these frames in one-second audio clip. In the end of this module, speech and music classes are considered in final audio data indexing. Thereafter, environment sound segments are the input of recognition concepts module of our system.

c. *Learning and identification concepts modules*

Audio concepts identification is based on two sound stages, namely low-level feature extraction based on MFCC descriptor and concepts classification based on supervised SVMs classifiers.

- Concepts introduce and MFCC extraction

The first stage consists on labeling user sets positively and negatively the audio concepts for indexing. His choice is based on an expanded study of the sound taxonomy which can give semantic meaning to the final user of a video search system.

So the audio samples of each concept are introduced for a spectral description. To obtain a good description for environmental sound segments, we extract MFCC features. Mel-Frequency Spectral Coefficients (MFCCs) are a set of perceptual parameters calculated from the STFT [21]. They provide a compact representation of the spectral envelope, such that most of the signal energy is concentrated in the first coefficient. After introducing training signal audio of concepts, the feature extraction module gets in output a vector for each sound frames.

- Analyses of audio SVM classifiers scheme

Our audio classification scheme is based on SVM classifier. Each SVM corresponds to a unique semantic class (concept) and shall contain in the input layer an individual MFCC feature. Since there can have a massive amount of negative examples, we shall be selecting a limited sub-set out of them, which are in the closest proximity of the positive examples and thus have the highest potential for erroneous classification.

2.3 Level 1: object refinement

The level 1 deals with mixed unimodal semantic interpretations. These last are structured through this data format: every concept has a list of indexed video content sorted by their descending pertinent ranks. These ranks are normalized then analyzed.

2.3.1 Concept ranks fuzzification

The purpose of this step is to calculate the fuzzy membership degree of a concept to a video content. This is given by a normalization function (see equation 1).

Let r be the rank of a concept for a video content, and R is the highest rank of the same concept for other video contents. We seek for a transformed rank called r_N as follow:

$$r_N = \left(\frac{\varepsilon - 1}{R - 1} \right) * (R - r) + 1 \quad (3)$$

Where ε is a small positive integer.

2.3.2 Conflicting situations handling

Sometimes, certain conflicting situations can be found in aggregated semantic interpretations. Since we are using fuzzy logic, two contradictory semantic interpretations can coexist for the same video content. This situation is permitted until the equation 2 is verified.

Let C_1 and C_2 be two contradictory concepts. And let $\mu(C_1)$ and $\mu(C_2)$ be respectively the membership degree of the first and the second concept.

$$\mu(C_1) \approx 1 - \mu(C_2) \quad (4)$$

If the equation is not verified, the concept, extracted from the less confident modality, is deleted. This trust degree is established for each modality by the fusion control process (level 4).

In other situations, we find that the concept relevance analysis in a video content varies from one modality to another. This conflict over the relevance degree for the same concept is solved using of equation 3. (Same equation used in [16]).

Let C a concept. And let $\mu_1(C)$ and $\mu_2(C)$ the relevance degrees computed respectively from the first and the second modality.

$$\mu(C) = \alpha * \mu_1(C) + \beta * \mu_2(C) \quad (5)$$

Where α and β are trust degrees respectively for the first and second modality fixed by the fusion process control (level 4), and $\alpha + \beta = 1$.

We can see that level 1 is important since it eliminates any extraneous information. However, several actual indexing systems do not well account for this component (as in [16], [17] and [18]). The set of fuzzyfied and filtered concepts are finally passed to the level 2.

2.4 Level 2: Situation refinement

The purpose of this level is to look for new concepts by analyzing available interpretations. To do this, we propose the use of two different intelligent techniques.

2.4.1 Deduction engine

By using a Mamdani fuzzy system [13], a deduction engine infers new concepts using fuzzy rules extracted from the LSCOM ontology [14]. This ontology is based on generalization relationships. Fig 4 illustrates how we extract fuzzy rules from the LSCOM ontology. The Fig 5 shows an overview of deduction engine using extracted fuzzy rules.

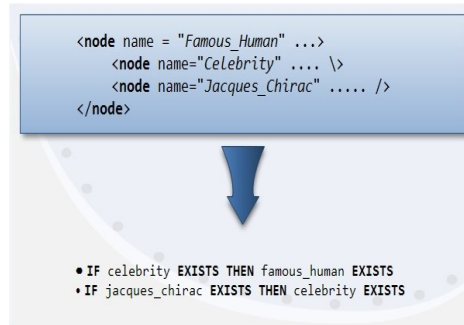


Fig 4. Extracting Fuzzy Rules from LSCOM Ontology.

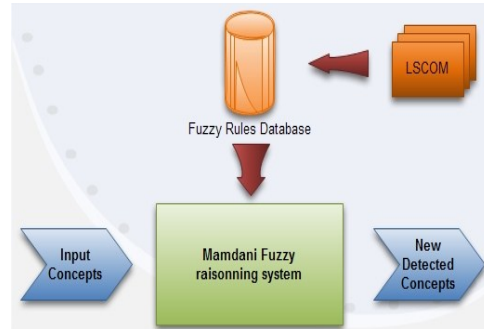


Fig 5. Deduction engine for situation refinement.

2.4.2 Abduction engine

The abduction engine automatically searches for possible relationships between concepts by analyzing training data (shows figure 6). These rules are used then in extracting of further concepts in real situations. This engine uses the Beta fuzzy systems (BFS) based on a multiagent genetic algorithm [15].

Using a video database for the learning process, the genetic learning system analyzes input concepts and output ones. The goal is to find possible fuzzy relationships between these concepts. Then, in the test process, these detected fuzzy relationships are used as fuzzy rules to deduct new concepts from ones extracted from a test video database.

2.5 Level 4 :Fusion Process Control

This level aims at controlling the fusion process by manipulating trust degree of each modality (sound and images) and for each reasoning engine (abduction and deduction). These trust degrees are determined following a supervised learning.

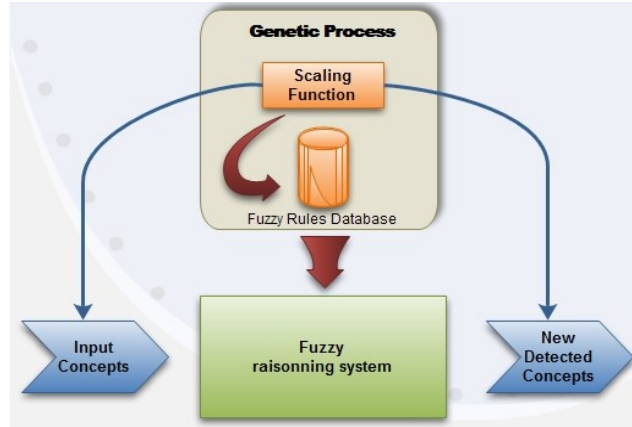


Fig 6. Abduction engine for situation refinement.

We can conclude that the multimodal fuzzy fusion that we propose is different from the majority of video indexing system by its ability to deliver a rich and coherent semantic interpretation. This is mainly due to:

- The use of fuzzy logic and fuzzy reasoning.
- A maximum compatibility with the JDL / DFS data fusion model.
- Operating a maximum of information extracted from a video content (multi-modality).

However, the quality of this fusion system is still dependent on the quality of initial semantic interpretations delivered by unimodal analyzers.

3 Experiments

We investigated the contribution of each component discussed in Sections 2.1–2.3. We emphasize in particular the visual indexing process based on semantic concepts categorization by using bag of pseudo-sentences (BoPS), and the effectiveness of our fusion system.

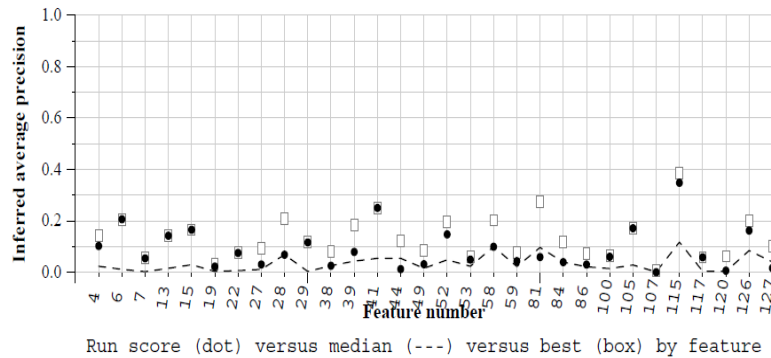
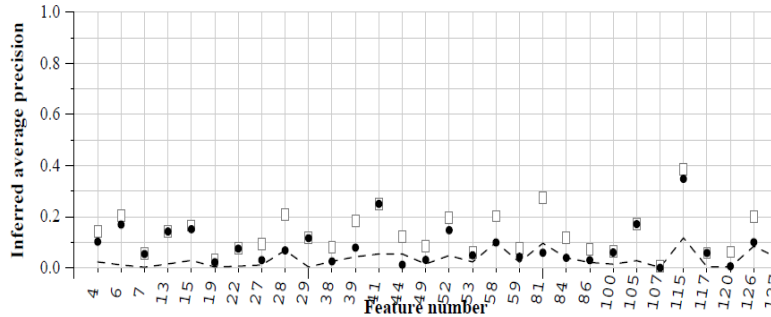


Fig 7. TRECVID 2010: Evaluation of REGIM_5.

The experimental results of our 3 runs are shown in figure 6 and 7, and REGIM_5 achieves the best mean inferred average precision (infAP = 0.089) among all runs (infAP_REGIM_4 is 0.085 and infAP_REGIM_6 is 0.089).

We note that results given by TRECVID2010 are limited to the detection of 30 among 130 concepts which are initially suggested. In total, we recorded a detection improvement of 14 among 130 concepts for the multimodal fuzzy fusion system vs. visual indexing system.



Run score (dot) versus median (---) versus best (box) by feature

Fig 8. TRECVID 2010: Evaluation of REGIM_4.

Table 1 shows concept detection improvement, given by our multimodal fusion system vs. the unimodal visual analysis system, in terms of indexed shots number.

Table 1. CONCEPT DETECTION IMPROVEMENT.

ID	Concept	Nb Frames REGIM_4	Nb Frames REGIM_5
6	Animal	627	737
12	Bicycles	55	249
15	Boat Ship	177	246
21	Car	565	599
50	Face	1800	1925
51	Female Person	1501	1874
67	Indoor	336	972
75	Male Person	1883	2407
87	Outdoor	383	4636
90	Person	1998	9672
91	Plant	323	527
93	Politicians	391	418
108	Sky	845	845
111	Sports	1111	1277
125	Vegetation	1909	1909
126	Vehicle	728	1165

We note in the table 1 that for some concepts, there has been no improvement (eg the vegetation or sky concept concepts). This is due to the fact that the system could not find any rules allowing it to detect new video contents containing these concepts.

In addition, and for other concepts, the fusion system was able to improve the semantic interpretation by increasing the number of video contents indexed by these concepts (eg face and car concepts). However, in certain situations (eg person or outdoor concepts), the improvement is clearly significant. This is due to fact that the unimodal visual analysis focuses on the detection of sub concepts (eg Male person or Femal person concepts) without realizing that they are sub-concepts of the same generalized concepts (eg Person).

Table 2 shows the precision at number of shot of each runs in our system. It demonstrates the effectiveness of the multimodal fuzzy fusion system indexing.

Table 2. PRECISION AT n SHOT.

N shot	Precision REGIM_4	Precision REGIM_5	Precision REGIM_6
10	0.630	0.630	0.630
100	0.536	0.528	0.527
1000	0.181	0.193	0.194
2000	0.094	0.102	0.102

4 Conclusion

REGIM Research Group participated in this TREC Video Retrieval Semantic Indexing task. In this paper, we have presented preliminary experiments and obtained results. The main direction for the REGIMVID tool enhancement is the multi modal video indexing. Actually, the different video modalities indexing (visual and audio) are collectively performed. In future, we plan to incorporate motion information to detect concepts involving activities more effectively and the textual modality in our multimodal fusion system. Furthermore, the REGIMVid Toolbox functionalities will be enhanced by complementary tools as personalization and visualization. These last subsystems are under development.

Acknowledgment

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] C.W. Ngo, Y.G. Jiang, X.Y. Wei, W. Zhao, Y. Liu, J. Wang, S. Zhu, S.F. Chang. "High Level Feature Extraction, Automatic Video Search, and Content-Based Copy Detection", Proc of the International Conference TREC 2009.
- [2] C.G.M. Snoek and all. "The MediaMill TRECVID 2009 Semantic Video Search Engine", Proc of the International Conference TREC 2009.
- [3] A. Natsev and all. "IBM Research TRECVID-2009 Video Retrieval System". Proc of the International Conference TREC 2009.
- [4] E. L. Waltz and J. Llinas, "Multisensor Data Fusion". Norwood, MA, USA: Artech House, Inc., 1990, foreword By-White, Franklin E..
- [5] C. Schmid and all. "Evaluation of interest point detectors". Internat. J. Comput. Vision 37 (2),151–172, 2000.
- [6] N. Elleuch, A. Ben Ammar, A. M. Alimi. "A Generic System for Semantic Video Indexing by visual concept" , 5th International Symposium on Image/Video Communications and Mobile Networks. 30 September 2010, RABAT, MOROCCO.
- [7] H. Karray and all. "REGIM at TRECVID2008: High-level Features Extraction and Video Search", Proc of the International Conference TREC 2008.
- [8] C.G.M. Snoek and all. "Early versus late fusion in semantic video analysis". In proceedings of ACM Multimedia, 2005.
- [9] M. Ellouze, N. Boujemaa, A. M. Alimi. "Scene Pathfinder: Unsupervised Clustering Techniques for Movie Scenes Extraction", In Proc of the International Conference of Multimedia Tools and Applications, No 2, pp., 325-346, 2010.
- [10] M. Ellouze, , H. Karray, ,W.B. Soltana, A. M. Alimi, 2007. "Utilisation de la carte de Kohonen pour la détection des plans présentateur d'un journal télévisé", In Proceedings of international conference TAIMA 2007, cinquième édition des ateliers de travail sur le traitement et l'analyse de l'information .pp. 271-276.
- [11] Kohonen T., 1990. "The Self-Organizing Map". In Proceedings of the IEEE, pp., 1464-1480.
- [12] J. Vesanto and E. Alhoniemi, "Clustering of the Self-Organizing Map" , In proc of IEEE Transactions on Neurral Networks, Vol. 11, No. 3, pp., 586-600, MAY 2000
- [13] E. Mamdani and S. Assilian, "an experment in linguistic synthesis with a fuzzy logic controller," *Int. J. Man-Machine Studies*, vol. 7, pp. 1–13, 1975
- [14] L. Kennedy and A. Hauptmann, "Lscom lexicon definitions and annotations (version 1.0)," Columbia University, Tech. Rep., 2006
- [15] I. Kallel and A. Alimi, "Magad-bfs: A learning method for beta fuzzy systems based on a multi-agent genetic algorithm," *Soft Computing - A Fusion of Foundations, Methodologies and Applications*, vol. 10, pp. 757–772, 2006
- [16] S. Vrochidis, A. Moutzidou, P. King, A. Dimou, V. Mezaris, and I. Kompatsiaris, "Verge: A video interactive retrieval engine," pp.1–6, 2010
- [17] C. G. M. Snoek, S. Member, M. Worring, J.-m. Geusebroek, D. C. Koelma, F. J. Seinstra, and A. W. M. Smeulders, "The semantic pathfinder: Using an authoring metaphor for generic multimedia indexing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, pp. 1678 – 1689, 2006.
- [18] S. Ayache, G. Quenot, and J. Gensel, "Image and video indexing using networks of operators," *J. Image Video Process.*, vol. 2007, no. 4, pp.1–13, 2007.
- [19] E. Scheirer and M. Slaney, Construction and Evaluation of a Robust Multifeature Music/Speech Discriminator. Proc. ICASSP 97, vol. II, pp 1331-1334. IEEE, April 1997.
- [20] K. El-Maleh, M. Klein, G. Petrucci and P. Kabal Speech/music discrimination for multimedia application. ICASSP, 2000.
- [21] K. Umaphathy, S. Krishnan and R. Rao. "Audio signal feature extraction and classification using local discriminate bases". IEEE Transactions on Audio, Speech and Language Processing, vol. 15(4), pp. 1236–1246, 2007.
- [22] F. Issam, A. Ben Ammar, M. A. Alimi, "AUDIO CONCEPTS IDENTIFICATION FRAMEWORK BASED ON BINARY CLASSIFIERS ENCAPSULATION", ICASSP 2011, *in press*.
- [23] M. Zarka, A. Ben Ammar, A. M. Alimi, "Multimodal Fuzzy Fusion System for Semantic Video Indexing", CIMSIVP 2011, *in press*.
- [24] N. Elleuch, A. Ben Ammar, A. M. Alimi., "Semantic Concepts Categorization Based on Pseudo-Sentences Representation", CIMSIVP 2011, *in press*.