# REGIMVID at TRECVID 2009: Semantic Access to Multimedia Data

*Nizar ELLEUCH, Issam FEKI, Anis BEN AMMAR, Hichem KARRAY, Adel M. ALIMI*
*REGIM : Research Group in Intelligent Machines, ENIS, Tunisia*
*[elleuch.nizar, feki_issam, anis.benammar, hichem.karray, adel.alimi]@ieee.org*
*http://www.regim.org*

## Description of Submitted Runs

### High-Level Feature Extraction

- Regim_1: local feature alone with SVM classification results and ARGs using detection scores of image concept as features.
- Regim_2: linear weighted learning image concept with SVM classification results for each concept using various feature representation choices.
- Regim_3: local feature alone with SVM classification results for each concept using various feature representation choices.

### Automatic Search

- Regim_AS: Type-A interactive run with 24 semantic concepts targeted in the 2009 HLF task.

## Abstract

In this paper we describe our TRECVID 2009 video retrieval experiments. The REGIMVID team participated in two tasks: High Level Feature Extraction and Automatic Search. Our TRECVID 2009 experiments focus on increasing the robustness of a small set of sensors and the relevance of the results using a probabilistic weighting of learning examples.

***Keywords:*** *Bag-of-words, support Vector Machines, ARGs, feature fusion, ISTRDO.*

## 1 Introduction

During the last decade, the automatic processing of audio-visual aids such as video indexing is a focus for several teams [1, 4, 11, 15]. Indeed, several techniques have been proposed and several systems have been developed. They are mainly interested in the use of new local descriptor [14] or the use of learning techniques in cascade with the integration bag-of-words [4, 15]. In this sense, REGIMVID team tries to improve its system for video indexing [11] by the use of new technique.

To surmount the semantic gap which results in the phase shift between the raw descriptions of different modalities (text, sound and image) and conceptual descriptions understandable by the user, the video indexing systems by engagement and the concept still develop new descriptors [14]. During this session TRECVID, REGIMVID proposes a new shape descriptor invariant to scale, translation, rotation, deformation and even to partial occlusions. It is to submit a form by selecting the discontinuity of its contour shape to locate an object or a particular shape in a short time.

In recent years, a new method called bag-of-features algorithm, which successfully employs text retrieval approach for High Level Feature Extraction, has shown promising results in various applications, like general object categorization [8,6].

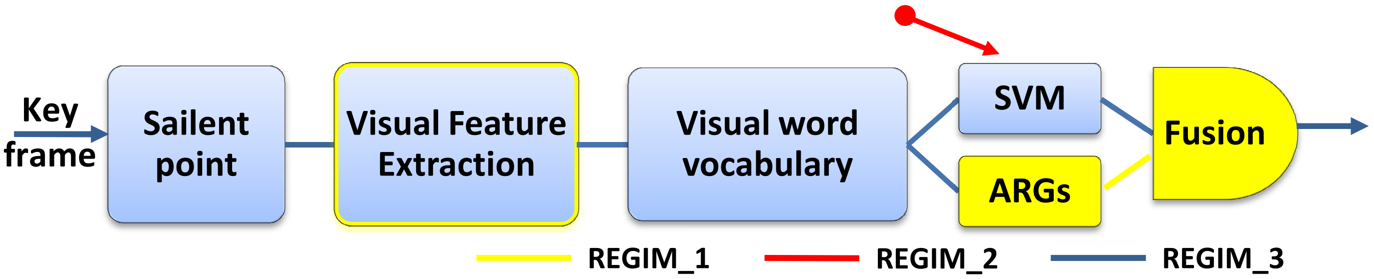

**Figure 1. Overview of the REGIMVID toolbox for video input**

In analogy with the text retrieval task using "key words", images are preprocessed to extract "salient" regions and some invariant descriptor is employed to represent the local visual texture information of each region. Then a sample group of descriptors are clustered to build a vocabulary tree of "visual words". After building the vocabulary tree, all extracted salient regions from each image are assigned to the closest "visual words" in vocabulary under a certain distance measure. Thus, each image can be represented as a bag of visual words which number depends on indexing system. For each visual word we traduce its frequency [15]. However, bag of visual words does not take into account the spatial distribution of the information in key frames. A mathematical model is, so, proposed for structuring the spatial distribution of keywords extracted from each image. It reports the semantic dependencies between the signatures of the neighborhood salient points while maintaining the distribution of special salient points.

We first present the REGIMVID toolbox. Next, we provide details of the visual feature extraction process. The experimental results for both search and high-level feature are then detailed. We conclude with some future directions, including the incorporation of textual and acoustic transcriptions.

## 2 REGIMVID Toolbox

In this section, we present an overview of the structure of the toolbox and briefly describe the novel visual feature extraction techniques currently supported.

### 2.1 Architecture

The framework of our baseline system is shown in Figure 1. For each key frame image, we extract the salient point. For each salient point, the low-level visual features are extracted, like SIFT, wavelet, Gabor, etc. These features are clustered by using Self-organizing maps (SOM). After combining together in an "early fusion" manner, that is, their feature vectors are concatenated as a new feature vector which presents a visual word vocabulary. Then we process a supervised learning process using SVM classifier. Also, we lay the examples of learning to improve system performance and display the relevant documents to the top of the results generated by the supervised classification technique.

Below is the description of the supervised learning process we perform.

### 2.2 Supervised learners

Supervised classification is mainly employed to bridge the gap between low-level image features and high-level semantic information presented in the images. Similar to our past participation [11], we use the support vector machine framework for supervised learning of semantic concepts. Here we use the LIBSVM implementation (for more detail see *http://www.csie.ntu.edu.tw*) with probabilistic output [12]. It is well known that the parameters of the support vector machine algorithm have a significant influence on concept detection performance [3, 7]. The parameters of the support vector machine we optimize are the kernel function K(.) and C.

Despite the performance results provided by SVM at the last session TRECVID, it neglects the spatial distribution of relevant areas. In fact, the spatial locations of key points in an image carry important information for classification process. So, REGIMVID system integrates the Attributed Relational Graph for the spatial information. These represent an extension of the graph by joining the real and multidimensional attributes of vertices and edges. Indeed, we equate the vertices at the points of interest extracted by the descriptor form edges semantic dependencies between signatures in the neighborhood of each point of interest. These dependencies are based on semantic analysis of structural and evolutionary semantic fields using algorithms for sorting large arrays. These are used to extract the blocks representing the emerging concepts and consistent from the sort of diagonal matrix of co-occurrence crossing emerging regions between them.

The fusion of result of SVM and ARGs is based on probabilistic scores.

## 2.3 Visual feature Extraction

We used a set of different visual descriptors at various granularities for each representative key frame of video shots. The relative performance of the specific features within a given feature modality is shown to be consistent across all concepts/topics. However, the relative importance of one feature modality vs. another may change from one concept to another. The following descriptors had the top overall performance of both search and concept modeling experiments:

- Color Histogram: Global color represented as 128- dimensional histogram in HSV color space.
- Color Moments: localized color extracted from 3x3 grid and represented by the first 3 moments for each grid region in Lab color space as normalized 255-dimensional vector.
- Edge Histogram: global edge magnitude with 8 edge direction bins and 8 edge magnitude bins based on a sobel filter (64-dimensional).
- Co-occurrence Texture: global texture represented as a normalized 96-dimensional vector of entropy, energy, contrast and homogeneity extracted from the image gray–scale co-occurrence matrix at 24 orientations.
- Gabor Texture: Gabor functions are Gaussians modulated by complex sinusoids. The Gabor filter masks can be considered as orientation and scale tunable and lines detectors. The statistics of these micro features in a given region can be used to characterize the underlying texture information. We take 4 scales and 6 orientations of Gabor textures and further use their mean and standard deviation to represent the whole key frame and result in 48 textures.
- GLCM: The GLCM (Gray- Level Co-occurrence Matrix) is a common technique in statistical image analysis that used to estimate image properties related to second-order statics. GLCM considers the relation between two neighboring pixels in one offset, as the second order texture, where the first pixel is called reference and the second one neighbor pixel. GLCM is the two dimensional matrix of joint probabilities $P_{d, \theta}$ (I, J) between pairs of pixels, separated by a distance **d** in a given direction **θ**. We used four statistical features (contrast, Correlation, Energy and Homogeneity) from gray-level Co-occurrence matrix for texture classification.
- Fourrier: The Fourrier- transforming said image to find a radial reference point, normalizing said Fourrier- transformed image with reference to said reference point, and then describing said texture descriptor by using said normalized values of said Fourier- transformed image. Here, the radial reference point is set by determining an arc in which one of energy,

entropy and a periodical component of said Fourier- transformed image apart at the same distance from the origin in said frequency domain is most distributed, and then setting a radius of said founded arc as said radial reference point.

- Sift: The SIFT descriptor [5] is consistently among the best performing interest region descriptors. SIFT describes the local shape of the interest region using edge histograms. To make the descriptor invariant, While retaining some positional information, the interest region is divided into a 4x4 grid and every sector has its own edge direction histogram( 8bins ).The grid is aligned with the dominant direction of the edges in the interest region to make the descriptor rotation invariant.
- Combined Sift and Gabor.
- Wavelet Transform for texture Descriptor: wavelets are hybrids that are waves within a region of the image, but otherwise particles. Another important distinction is between particles that have place tokens and those that do not. Although all particles have places in the image it does not follow that these places will be represented by tokens in feature space. It is entirely feasible to describe some images as a set of particles, of unknown position. Something like this happens in any description of texture. We performed 3levels of a Daubechies wavelet [13 decomposition for each frame and calculate the energy level for each scale, which resulted in 10 bins features data.
- Hough Transform: As descriptor of shape we employ a histogram based on the calculation of Hough transform [2]. This histogram gives information better than those by the edge histogram. We thus obtain a combination of behavior of the pixels in the image along the straight lines
- ISTRDO: he is based mainly on measurements of Haralick [10] (contrast, energy, entropy, ...) applied on a selection of some points from the contour of a given shape. First, for each image, we define pyramid 4 scales and 8 orientations. For these 32 images, we detect the contour using the method of Canny. Subsequently, for the entire set of contour points, we calculate the orientation of the contour through the gradient, in order to detect the discontinuity. It is based on the change of direction. Then, all discontinuous points of the contour are described through the Haralick measures for a 3 * 3 neighborhood.

## 2.4 TRECVID 2009 evaluation

We investigated the contribution of each component discussed in Sections 2.1–2.3, emphasizing in particular the role of linear weighted image concept, the use of bag-of-words, the influence of spatial information, and the effectiveness of kernel-based learning parameters.

The experimental results of our 3 runs are shown in figure 2, and REGIM_1 achieves the best mean inferred average precision (infAP = 0.127) among all runs (infAP_REGIM_2 is 0.123 and infAP_REGIM_3 is 0.111).

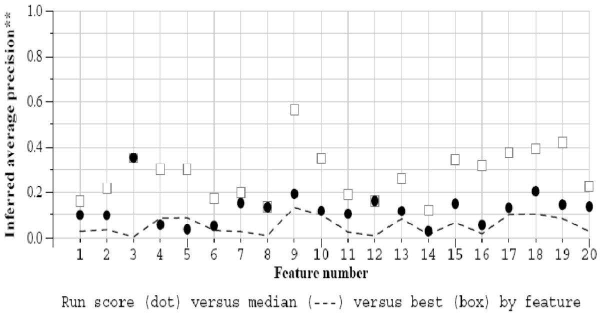

**Figure 2. TRECVID 2009: Evaluation of REGIM_1**

Table 1 shows the precision at number of shot of each runs in our system. It demonstrates the value of linear weighted learning image concept.

**Tableau 1. Precision at n shot**

| N shot | Precision REGIM_1 | Precision REGIM_2 | Precision REGIM_3 |
|---|---|---|---|
| 5 | 0.8400 | 0.8600 | 0.6400 |
| 10 | 0.7300 | 0.7400 | 0.6300 |
| 15 | 0.7400 | 0.7334 | 0.6466 |
| 20 | 0.6750 | 0.6750 | 0.6150 |
| 30 | 0.6834 | 0.6766 | 0.6000 |
| 100 | 0.5420 | 0.5330 | 0.4990 |
| 200 | 0.3880 | 0.3854 | 0.3814 |
| 500 | 0.1932 | 0.1922 | 0.1922 |
| 1000 | 0.1116 | 0.1106 | 0.1106 |
| 2000 | 0.0632 | 0.0630 | 0.0630 |

# 3 Search

The REGIM team's has implemented three sub systems for automatic search. The first one is a retrieval engine based on textual transcription detection within video sequence. The second one focuses on visual feature detection. The last one is based on automatic speech recognition for audio feature detection (show figure 3).

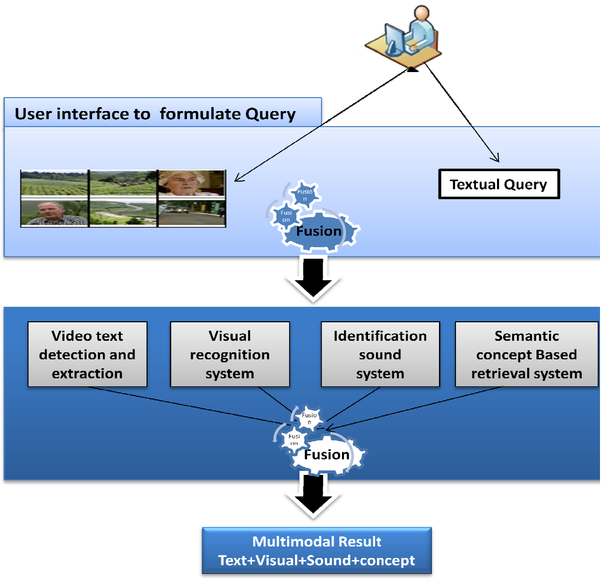
User interface to formulate Query
Textual Query
Fusion
Video text detection and extraction
Visual recognition system
Identification sound system
Semantic concept Based retrieval system
Fusion
Multimodal Result
Text+Visual+Sound+concept

**Figure 3. Overview of REGIM Interactive Search System**

## 3.1 Video text detection and extraction

Text within an image is a particular interest for indexing and retrieval of video because of its capability to describe the contents of an image, and its relationship to the semantic information; it also enables applications such as keyword based search in multimedia databases. In this work we use a wavelet- based method to detect an extract the text from the video frames [16], this method works without the dependency on the availability of an OCR. Extraction of text information involves detection, localization, enhancement and recognition of the textual content in the video frames; a method involves a frame by frame processing on the entire video for locating textual blocks.

## 3.2 Visual recognition system

Our indexing system is based on visual feature extraction visual low-level (color, texture and shape) from the key frames and translate them into semantic information by using bags of words. The latter serve as good examples for supervised learning of each concept. Our system uses the SVM as a learning technique which gives better performance. For the configuration of various parameters of the SVM, we proceed by a gradual change of function learning until to be the best.

## 3.3 Identification sound system

If the methods of extraction of video content have just focused on visual, audio has also proven to contain useful information. For example, it is possible to identify a sports scene by the noise of a crowd, or landing of aircraft by the hum of a jet engine [9]. The audio track does not match the images, even for speech and music. It is also much more difficult even for humans to distinguish the sounds as image. As expected, the results of average precision for audio-based research were lower than visual. Yet some can be attributed to the fact that the shots were labeled using only images, and a review of the audio blows retrieved provided some encouraging results. A segmentation of the audio and a separation of the source could possibly improve the results. These methods can be explored in future work TRECVID.

## 3.4 Evaluation Results

This section describes submitted runs within TRECVID09 search task. Only 10 test topics evaluation were performed. However, only the second component of REGIMVID search engine was used. In fact, the baseline of our video retrieval engine is based on visual transcription. The integration of textual and speech transcription in retrieval process is in progress (multi modal fuzzy fusion).

The graph below presents the obtained results.

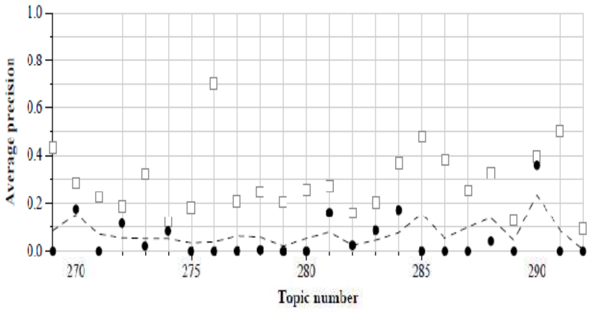

**Figure 4. TRECVID 2009 Search Results**

## 3.5 Research conclusion

We introduced the system in text, visual and audio used for the automatic search. The operating time was sometimes high, because the time taken to train the SVM models which included local characteristics. As this was our second year participating in the evaluation TRECVID much time was devoted to development work.

## 4 Conclusion

REGIM Research Group participated in this TREC Video Retrieval High-level features Extraction and Search tasks for the second time. In this paper, we have presented preliminary results and experiments for both tasks. The main direction for the REGIMVID tool enhancement is the multi modal video indexing. Actually, the different video modalities indexing (visual, textual and speech) are separately performed.

## References

[1] Apostol Natsev¤, Wei Jiangy, Michele Merlery and al, "IBM Research TRECVID-2008 Video Retrieval System", *NIST TRECVID Workshop, 2008.*

[2] Boujemaa N. Ferecatu M. and Gouet V, "Approximate search vs. precise search by visual content in cultural heritage image databases". *In Proc. Of the 4-th International Workshop on Multimedia Information Retrieval (MIR 2002) in conjunction with ACMMultimedia, 2002.*

[3] C. G. M. Snoek, 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 Trans. PAMI, 28(10):1678−1689, 2006.

[4] C.G.M. Snoek, K.E.A. van de Sande and al, "The MediaMill TRECVID 2008 Semantic Video Search Engine", *NIST TRECVID Workshop, 2008.*

[5] David G. Lowe, "Distinctive image features from scale-invariant keypoints", *International Journal of Computer Vision, 60, 2 (2004), pp. 91-110.*

[6] D. Nister and H. Stewenius, 2006, "Scalable Recognition with a Vocabulary Tree", *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*

[7] D. Wang, X. Liu, L. Luo, J. Li, and B. Zhang. "Video Diver: generic video indexing with diverse features. ", *In Proc. ACM SIGMM MIR Workshop, pages 61−70, Augsburg, Germany,2007.*

[8] E. Nowak, F. Jurie, and B. Triggs, 2006, "Sampling Strategies for Bag-of-Features Image Classification," *in Computer Vision ECCV 2006, pp. 490-503.*

[9] Guodong Guo and Stan Z. Li, "Content-based audio classification and retrieval by support vector machines", *IEEE Transactions on Neural Networks, 2003.*

[10] R.M. Haralick, K. Shanmugam, and Dinstein I. "Textural features for image classification.", *IEEE Trans. on Systems*

*Man and Cybernetics, SMC-3(6):610–621, November 1973.*

[11] H. Karray, A. wali, N. Elleuch and al, "REGIM an TRECVID2008: High Level Features Extraction and Video Search", *NIST TRECVID Workshop, 2008.*

[12] H.-T. Lin, C.-J. Lin, and R. C. Weng. "A note on Platts probabilistic outputs for support vector machines". ML, 68(3):267−276, 2007.

[13] I. Daubechies. CBMS-NSF series in app. Math., *chapter SIAM. 1991.*

[14] Koen E. A. van de Sande, Theo Gevers, Cees G. M. Snoek, "Evaluating Color Descriptors for Object and Scene Recognition", *IEEE Transactions on Pattern Analysis and Machine Intelligence (in press), 2010.*

[15] Shih-Fu Chang, Junfeng He, Yu-Gang Jiang and al, "High-Level Feature Extraction and Interactive Video Search ", *NIST TRECVID Workshop, 2008.*

[16] Wali A. Karray H. and Alimi M.A. Sirpvct: "System of indexing and the search for video plans by the content text", *In proc. Treatment and analyses information: methods and Application, TAIMA 07, pages 291, 297, Tunisia Hammamet, May 2007.*