GIM at TrecVid 2012

The Light Semantic Indexing Task

Teresa Alonso, Manuel Barrena, Pablo G. Rodríguez, Antonio Polo, Miryam Salas, Andrés Caro, Marisa Durán, Félix Rodríguez, Luís Arévalo, Antonio Corral, Mar Ávila, Jorge Martínez, Francisco Ruiz, Soledad Martín

University of Extremadura, Media Engineering Group
Escuela Politécnica, Av. Universidad s/n, 10003 Cáceres, Spain
barrena@unex.es
October 29, 2012

ABSTRACT

The Media Engineering Research Group (GIM) participated in the semantic indexing task at TRECVID 2012. In order to detect semantic content inside the videos, we used a simple approach based on visual features including color, motion and SURF combined with textual metadata from annotations. We finally submitted three runs based on the following approaches:

L_A_GIM_RUN1_1: training based in a subset of shots annotated by the collaborative annotation. We calculate centroids of every visual and textual feature for every distinct concept and compute distances from the shot that is being evaluated to the centroids. Probability of a shot representing one concept is equal to 100 minus wheighted distances: color->28%, motion->21%, SURF->21%, textual features->30%. Relationships between concepts add or subtract 10% to the final probability and a threshold value 25% is applied on it in order to discard shots.

L_A_GIM_RUN2_2: same as L_A_GIM_RUN1_1 but using another subset of shots for training and a threshold value 33%.

L_A_GIM_RUN3_3: similar to previous but using a different training set and a different weighting: color->52.5%, motion->17.5%, SURF->0%, textual features->30%, and a threshold value 0%.

This paper describes the experiments we carried out for this task, the problems we faced during its development and a final proposal to enhance the system evaluation process in this particular task by means of the use of a new benchmark.

I. Introduction

The GIM group at the University of Extremadura has been working for several years on content-based analysis of images, feature extraction and the subsequent use of them in similarity search, classification, relevance feedback and interpretation of semantic content among others. Features traditionally studied by GIM belong to sets of COLOR, SHAPE and TEXTURE.

Research developed by GIM between 2007 and 2010 on images has been materialized in successive versions of the application *Qatris iManager* [8], software designed and implemented by our group in collaboration with the company SICUBO™ [13]. *Qatris iManager* is a CBIR system

(Content-Based Image Retrieval) designed for storage, indexing, classification and retrieval of images based on their content.

Since 2009, GIM extends its research area from images to video, and we began with the implementation and development of the tool Qatris vManager [3, 9, 12]. Qatris vManager is an application to catalog video collections according to a set of taxonomies defined by the user. It allows the user to break the video in a set of segments and to store different kind of metadata for every segment and for the whole video. All the metadata information is encoded in XML format according to the standard MPEG-7 [11]. Qatris vManager is, therefore, an application for storage, cataloging and retrieval of video and video shots based on textual information, but it is ready to add visual information to the metadata. Such visual information will be used in a near future for a more selective and accurate recovering.

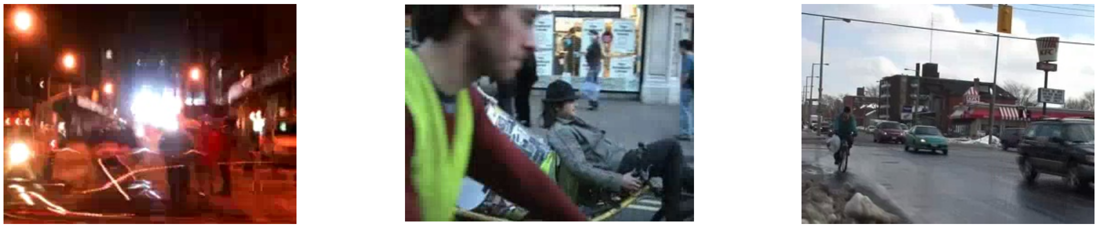
The interest of GIM in TRECVID [1] started at 2010, but it has not been until 2012 that we have decided to actively participate in it and test our previous work on video and metadata provided by the TRECVID platform. Among the various tasks available, the choice has been the "Light Semantic Indexing Task" because GIM group feels that the automatic assignment of labels representing visual concepts and the use of them for searching and classifying video shots is closely related to our research activity.

After studying some papers [2, 4, 5, 6, 7, 14] by teams participating in previous TRECVID calls, we saw that most commonly followed strategies for automatic annotation that achieve good performance for concept detection turn around support vector machines, visual vocabularies in the bag-of-visual words model using local feature detectors as SIFT (Scale Invariant Feature Transform) or SURF, combination of local and global features in a feature vector, use of textual metadata excluding stop-words and validating the rest of the terms, and hybrid approaches mixing different methods. Differently from previous works, we decided however to check the validity of previous basic research on video classification by content [3, 12] with the benchmark provided by TRECVID, and although the final results from the runs submitted were not good, lessons learned from our participation in TRECVID will be really useful for the next future of GIM team. In fact we make a proposal to the TRECVID community concerning the use of different video benchmark collections.

II. Training process

To build the system IACC1-A and IACC1-B video collections were used. These sets present a wide range of topics filmed under very different conditions. They are really heterogeneous collections captured from the Internet with unpredictable quality and content. Most of them are videos in color but another part of the collection are black and white videos, their content ranges from cartoons, animations, home movies to newsreels. According to this, there is a clear difficulty in making training sets representing the whole collection.

To train the system one needs to choose a relevant sample of videos containing the concept to be detected and the way we chose to do it was by means of a subset of videos in IACC1-A annotated through the collaborative annotation activity, which provided annotated video shots for every concept. The point with this annotation process is that it is not controlled by an authority imposing a set of rules limiting the conditions in which a key frame truly represents a particular concept. Because of that, relevant key frames for some concepts may present "surprises" like figure 1 shows. Also, the number of relevant key frames for different concepts varies, so that some concepts have a very poor representation in number of relevant key frames.



(a) (b) (c)

Figure 1.- Reference key frames for concepts (a) Airplane_Flying, (b) Bicycling and (c) Bus

We also wanted to use textual metadata attached to the video in the training process, so that text from metadata fields like title, description, keywords and subject was collected for each video in the training set. We had to fight with similar problems for this kind of text features, because they contain too much imprecisions like spelling or grammatical mistakes, repeated words, words with an obscure meaning and with no apparent relation to the video content, and so on.

Taking into account all these problems, our team designed a strategy in order to try to minimize their effects and to obtain a training data set that meet a few basic requirements. The first step consisted of extracting the reference key frame (RKF) for each video shot provided by TRECVID collaborative annotation as relevant to each concept out of the 50 concepts of the *light run* subtask. Then we stored those RKFs in 50 separate folders, one per concept. Each folder was sent to two different members of the GIM team with the aim to filter key frames not clearly representative of the concept, by avoiding so the use of key frames like those in the figure 1. Once this data cleaning process was done, two training video sets were created and named UNION and INTERSECTION. The last one contained video shots corresponding to RKFs that were declared as visually relevant to the corresponding concept for *both* GIM members, while the former one was built with video shots corresponding to RKFs declared as relevant *by one* of the members. Based on this distribution, we train the system. Thus, the initial set of annotated shots was reduced to 55,645 supervised annotations in the UNION training set and to 29,360 supervised annotations in the INTERSECTION training set.

III Feature extraction

For the experiments we have chosen to work with two different groups of features: visual and textual ones.

Visual features

Vectors of COLOR, MOTION and SURF features have been extracted from each training shot.

- COLOR.- 44 feature values computed from a palette of 11 colors: white, black, gray, red, orange, yellow, green, cyan, blue, purple and magenta. For each of these 11 colors it is calculated the percentage or level of such a color and its deviation as regards its centroid in the RFK of any shot.
- SURF.- The RKF image is segmented and its keypoints are located. Those keypoints subsequently serve to recognize places and objects of all kinds. They are invariant to scaling and translation and partly to rotation and illumination changes. Keypoints inside the RKF of each shot have been detected and a descriptive vector containing 64 medium values calculated.
- MOTION.- 13 features extracted from all the frames that make up the shot, not just its RKF. Within each shot, successive comparisons are made between a frame and the next one so that the motion intensity and direction (0 ° - 360 °) is computed in 5 different areas or regions of the RKF image. These five areas are obtained by dividing the rectangle occupied by the frame in 4 subrectangles identical in size and it is considered a 5th subrectangle placed at its center. Figure 2 shows the location of these 5 areas in a video frame.

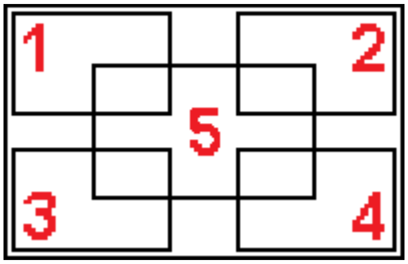


Figure 2.- Distribution of the 5 areas in a video frame

So, 10 out of 13 motion features are computed: 5 intensities and 5 directions. The remaining 3 features correspond to the most frequent motion intensity among those calculated for the five areas, the direction of that one and a feature called "camera movement" that indicates whether the camera that filmed the shot added movement to the scene or, conversely, it was filmed from a static and fixed point of view.

When the visual features of all shots in the training sample have been calculated and since these shots are arranged in 50 subfolders, one for each concept in the light run, it has been possible to obtain the centroid and standard deviation of each of these visual characteristics for each class. In total, 6050 centroids were calculated with their corresponding deviations: 50 x 44 (one per class times one per feature) for COLOR features, 50 x 64 for SURF features and 50 x 13 for MOTION ones.

Textual features.-

Based on the classification of all training shots in 50 different classes or folders, we have extracted metadata associated with videos representing a particular concept, that is, videos containing shots annotated with the concept. Remember that TRECVID provides a metadata file in XML format related to each video. Within all the information provided as textual metadata, we have just read that contained between <keywords> ... </ keywords> and <subject> ... </ subject> tags. Stop words have been removed from this textual information and some other words, slogans or possible substitutes have been added. Finally, we have counted the number of occurrences of each of the keywords and subjects in the videos associated with a particular class or concept because the cardinality of repetitions will be very important in determining the importance of each word.

In addition to visual and textual features, we have used the set of relationships between concepts that TRECVID provides. These relationships are limited to two types: i) the concept A implies the concept B, and ii) the concept A excludes the concept B. In our experiments we have just taken into account the relationships in which both the concept A and the concept B belong to the subset of 50 concepts of light run. Then these relations have been extended with those others deductible by transitivity.

IV Measuring the relevance

The Semantic Indexing Task for TRECVID 2012 consisted of given the test collection IACC1-C, master shot reference, and concept definitions, return for each concept a list of at most 2000 shot IDs from the test collection ranked according to their likeliness of containing the concept, being this probability or likeliness a percentage value.

We have carried out various tests based on video shots tagged by the collaborative annotation, resulting in different weighting of visual features, textual features and relationships. The weights that produced the best results on the training supervised subsets have been after applied in the runs submitted to TRECVID.

Thus, in order to calculate the relevance percentage of each test shot concerning a certain concept, we have followed the strategies described in the next paragraphs.

Visual features

We have obtained COLOR, SURF and MOTION features from each test video shot to evaluate and then its distance to each of the centroids computed on the training supervised set for each of the 50 concepts of the light run. These formulas describe how distances to centroids are calculated:

$$\forall\ concept\ a_j = a_1, a_2,\ ..a_{50} \mid a_j \text{€} A :$$

$$dist\_COLOR[a_j] = \sum_{i=1}^{44} \frac{|c_i - x_i|}{s_i}$$

$$dist\_SURF[a_j] = \sum_{i=1}^{64} \frac{|c_i - x_i|}{s_i}$$

$$dist\_MOTION[a_j] = \sum_{i=1}^{13} \frac{|c_i - x_i|}{s_i}$$

Where:

a_j represents the j-th concept among those 50 concepts in the Light Run.
c_i is the centroid value of the i-th visual feature, computed on the training set of shots.
x_i is the i-th visual feature value for the test shot that is being evaluated.
s_i is the standard deviation of the i-th visual feature, computed on the training set of shots.

Both the values of visual features and their distances from the centroids are normalized yielding a probability value between 0 and 100. Therefore, the percentage of probability that the shot represents a concept j-th is calculated as follows:

Prob_COLOR[a_j] = 100 – dist_COLOR[a_j]
Prob_SURF[a_j] = 100 – dist_SURF[a_j]
Prob_MOTION[a_j] = 100 – dist_MOTION[a_j]

Textual features

We have calculated the number of times each keyword and subject is repeated within the annotated training shots for each of the 50 concepts. The probability based on dictionary (Prob_DICC [a_j]) is calculated as follows. Keywords and subjects associated to the test video whose shots are being evaluated are compared with those read from the training set, that is, with the dictionaries previously created for each concept a_j and

- If the keyword or subject matches any of those in the j-th dictionary and it was repeated more than 5 times in the training set, then it is added 50% to the probability that this shot represents the concept a_j.
- If the keyword or subject matches any of those in the j-th dictionary and it was repeated more than 1 time in the training set, then it is added 35% to the probability that this shot represents the concept a_j .
- If the keyword or subject matches any of those in the j-th dictionary and it was found just 1 time in the training set, then it is added 20% to the probability that this shot represents the concept a_j .

If the final value of Prob_DICC [a_j] exceeds 100% then it is truncated to 100%.

V Run descriptions

We have submitted to TRECVID 2012 three different runs for evaluation. We describe how these runs calculate the probability that a given test shot represents a concept.

L_A_GIM_RUN1_1

- The training set INTERSECCIÓN is used in previous calculation of centroids, standard deviations and diccionaries.
- Then we computed distances to centroids of each group of visual features and probabilities depending on dictionaries.
- After that, those probabilities were weighted according to previous weights resulting in an initial estimation of the probability of one shot representing one concept. The exact weighting was:

$$\text{Prob\_initial}[a_j] = (\text{Prob\_COLOR}[a_j]*0{,}4 + \text{Prob\_SURF}[a_j]*0{,}3 + \text{Prob\_MOTION}[a_j]*0{,}3) \ * \ 0{,}7 \ + \ \text{Prob\_DICC}[a_j]*0{,}3$$

- Finally, knowledge about relationships betwen concepts were summed up as follows:

$$\text{Prob\_final}[a_j] = \text{Prob\_initial } [a_j]$$

$$\forall\, concept\ b_k = b_1, b_2, \,..b_{50}\ |b_k\ implies\ a_j$$

$$\text{Prob\_final}[a_j] = \text{Prob\_final}[a_j] + \text{Prob\_initial } [b_k]*0{,}1$$

$$\forall\, concept\ b_k = b_1, b_2, \,..b_{50}\ |b_k\ excludes\ a_j$$

$$\text{Prob\_final}[a_j] = \text{Prob\_final}[a_j] - \text{Prob\_initial } [b_k]*0{,}1$$

- A threshold has been defined as a minimum 25% value of probability required so as, concepts with a final probability lower than that were considered not to be represented in the shot being evaluated. Thus, the shot is discarded to represent the concept.
- Retrieved shots are ranked depending on their greater value of likeliness (Prob_final[a_j]) and the system returns at most 2000 shot IDs for each concept a_j.

L_A_GIM_RUN2_2

- The training set UNION is used in previous calculation of centroids, standard deviations and diccionaries.
- The likeliness of a shot representing a concept is calculated exactly the same as for L_A_GIM_RUN1_1 but applying a threshold value 33% instead of 25%.

L_A_GIM_RUN3_3

- The training set UNION is used in previous calculation of centroids and standard deviations, and all the videos tagged by the collaborative annotation of data set IACC1-A are used to create the dictionaries.
- The weighting of visual and textual features is:

Prob_initial[a_j] = (Prob_COLOR[a_j]*0,75 +
Prob_MOTION[aj]*0,25) * 0,7 + Prob_DICC[aj]*0,3

As it can be seen, local descriptors SURF are not taken into account in this run.

- Then, relationships between concepts are applied:

Prob_final[a_j] = Prob_initial [a_j]

$\forall$ concept $b_k = b_1, b_2, \ ..b_{50}\ |b_k$ implies a_j

Prob_final[a_j] = Prob_final[a_j] + Prob_initial [b_k]*0,1

$\forall$ concept $b_k = b_1, b_2, \ ..b_{50}\ |b_k$ excludes a_j

Prob_final[a_j] = Prob_final[a_j] - Prob_initial [b_k]*0,1

- There is no threshold defined, so all shots with a likeliness value greater than 0% are considered candidates to represent a given concept, although they will be only kept at most 2000 shot IDs per concept.
- As before, selected shots are ranked depending on their greater value of likeliness (Prob_final[a_j]) and the system returns at most 2000 shot IDs for each concept a_j.

VI Results and conclusions

During tests on supervised training sets of shots, we obtained mean percentages of success among the 50 concepts of 24.59% precision and 54.62 % recall.

Recall percentages are favored by the fact of having few shots for training and therefore the maximum number of potential positives for each concept was low. As up to 2000 shots were selected as possible solutions, and this cardinality is much higher than the number of training samples for some of the 50 concepts, this causes a deceptively high value in the recall measure.

However, when the same criterion has been followed in order to annotate automatically test data set IACC1-C, the average success rate has not exceeded 4.1%, a figure far below that expected and that has led us to some interesting conclusions.

The wide diversity present in the video data set IACC1-C makes that visual features identify very few common patterns among shots representing one concept.

For time saving reasons, during the calculation of the COLOR and SURF features, only RKFs have been evaluated. As one shot can represent multiple concepts, it often happens that the RKF selected for that shot does not include the visual representation of all the concepts related to the shot, but only some of them. Thus, the RKF is insufficient in some cases to determine whether the shot represents some concepts.

In the absence of a checking or supervision of textual metadata associated with each video, we found that many words are not correctly identified by the system because they contain spelling or grammatical mistakes. The automatic annotation system based on faulty information yields low results and some times it is even counterproductive when applied.

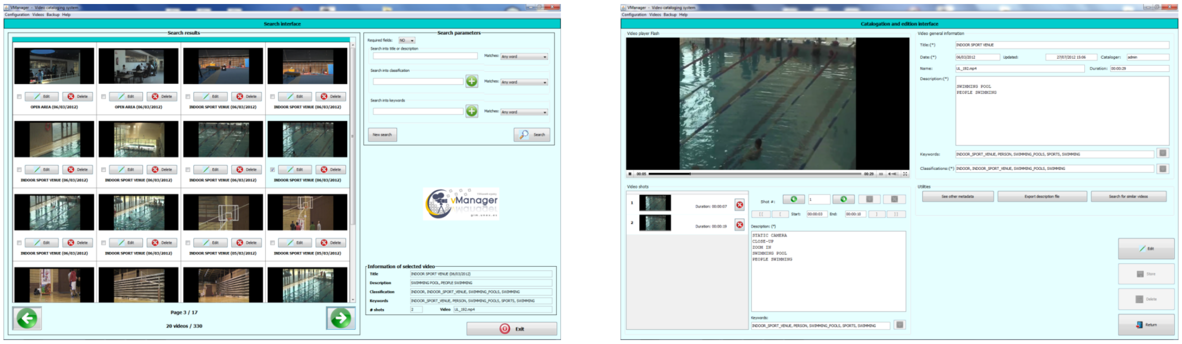
VII The final proposal

Our group was very interested in knowing what the results would be if the same criteria were applied on a supervised and restricted data set. With this goal, we have created a video database named UNIVERSITY LIFE that consists of 330 videos segmented into shots and annotated according to a specific ontology and using a supervised vocabulary. In this sense we would like to introduce the results of a new initiative we have started in our laboratory called University Life Video Database which is a video collection based on a well-known thematic around the life in a university campus. The annotation of that video database is based on LSCOM (Large Scale Concept Ontology for Multimedia) [10] and it is currently organized according to an ontology called ULO (University Life Ontology) .

Currently, the University Life Video Database has been annotated and organized according to ULO that consists of 26 different categories. Videos have been segmented into shots and both videos and shots have been added textual information as title, description and up to 80 keywords or semantic concepts among others. All these metadata are stored as XML files named the same as the corresponding video and encoded following the standard MPEG-7 scheme.

By following the same training process as we did on IACC1-A video collection to train the runs we submitted to TRECVID 2012, we have got better results with the University Life database, having 57,61% precision and 76,98% recall. Our purpose is to make available both the database and the ontology in order they can be extended by the research community to serve as a testbed for content-based video retrieval methods, techniques and systems. It is available for downloading from the web site gim.unex.es under Creative Commons license.

The creation of the UL taxonomy, the selection of a restricted set of semantic concepts or keywords as well as the segmentation and tagging of videos of University Life database have been carried out by means of our application *Qatris vManager* that makes it easy for the user to manage all those elements. Figure 3 shows two screen captures of *Qatris vManager*.



OPEN AREA (06/03/2012)
OPEN AREA (06/03/2012)
INDOOR SPORT VENUE (06/03/2012)
INDOOR SPORT VENUE (06/03/2012)
INDOOR SPORT VENUE (06/03/2012)
INDOOR SPORT VENUE (06/03/2012)
INDOOR SPORT VENUE (06/03/2012)
INDOOR SPORT VENUE (06/03/2012)
INDOOR SPORT VENUE (06/03/2012)
INDOOR SPORT VENUE (06/03/2012)
INDOOR SPORT VENUE (05/03/2012)
INDOOR SPORT VENUE (05/03/2012)

Page 3 / 17
20 videos / 330

Required fields: NO
Matches: Any word
Matches: Any word
Matches: Any word

Information of selected video
Title: INDOOR SPORT VENUE (06/03/2012)
Description: SWIMMING POOL, PEOPLE SWIMMING
Classification: INDOOR, INDOOR_SPORT_VENUE, SWIMMING_POOLS, SWIMMING
Keywords: INDOOR_SPORT_VENUE, PERSON, SWIMMING_POOLS, SPORTS, SWIMMING
# shots: 2 Video: UL_192.mp4

Title:(*): INDOOR SPORT VENUE
Date:(*): 06/03/2012 Updated: 27/07/2012 15:06 Cataloger: admin
Name: UL_192.mp4 Duration: 00:00:29
Description:(*):
SWIMMING POOL
PEOPLE SWIMMING
Keywords: INDOOR_SPORT_VENUE, PERSON, SWIMMING_POOLS, SPORTS, SWIMMING
Classifications:(*) INDOOR, INDOOR_SPORT_VENUE, SWIMMING_POOLS, SWIMMING

00:05 00:29

1 Duration: 00:00:07
2 Duration: 00:00:19

Shot #: 1
Start: 00:00:03 End: 00:00:10
Description: (*)
STATIC CAMERA
CLOSE-UP
ZOOM IN
SWIMMING POOL
PEOPLE SWIMMING
Keywords: INDOOR_SPORT_VENUE, PERSON, SWIMMING_POOLS, SPORTS, SWIMMING

Figure 3.- Screen captures of *Qatris vManager* application.

References

[1] Alan F. Smeaton and Paul Over and Wessel Kraaij. (2006). ”***Evaluation campaigns and TRECVid***”. MIR'06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval.ISBN 1-59593-495-2, pp.321-330, Santa Barbara, California, USA.

[2] Anastasia Moumtzidou[1], Panagiotis Sidiropoulos[1], Stefanos Vrochidis[1,2], Nikolaos Gkalelis[1], Spiros Nikolopoulos[1,2], Vasileios Mezaris[1], Ioannis Kompatsiaris[1], Ioannis Patras[2]. (2011). “***ITI-CERTH participation to TRECVID 2011***“. [1]) Informatics and Telematics Institute/Centre for Research and Technology Hellas, Thermi-Thessaloniki, Greece. [2]) Queen Mary, University of London, UK.

[3] Andrés caro, Pablo G. Rodríguez, Rubén Morcillo and Manuel Barrena. (2011). “***vManager, developing a complete CBVR system***”. 5th Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA’11), Las Palmas (Spain). Lectures Notes in Computer Science (LNCS), Vol.6669, pp.604-611, Springer-Verlag.

[4] Chao Chen[1], Dianting Liu[1], Qiusha Zhu[1], Tao Meng[1], Mei-Ling Shyu[1], Yimin Yang[2], HsinYu Ha[2], Fausto Fleites[2], Shu-Ching Chen[2]. Winnie Chen[3], Tiffany Chen[3]. (2011). “***Florida International University and University of Miami TRECVID 2011***“. [1]) Department of Electrical and Computer Engineering. University of Miami, USA. [2]) School of Computing and Information Sciences Florida International University, Miami, USA. [3]) Miami Palmetto Senior High School, USA.

[5] Chao Zhu, Boyang Gao, Charles-Edmond Bichot, Emmanuel Dellandré Liming Chen, Ningning Liu, and Yu Zhang. (2011). *ECL-LIRIS at TRECVID 2011: Semantic Indexing.* Universitéde Lyon, CNRS, Ecole centrale de Lyon, France.

[6] Chong-Wah Ngo† , Shi-Ai Zhu , Wei Zhang , Chun-Chet Tan, Ting Yao, Lei Pang, Hung-Khoon Tan‡. (2011). “***VIREO@TRECVID 2011: Instance Search, Semantic Indexing, Multimedia Event Detection and Known-Item Search***”. Video Retrieval Group (VIREO), City University of Hong Kong. ‡Faculty of Information and Communication Technology, University Tunku Abdul Rahman.

[7] David Scott, Jinlin Guo, Colum Foley, Frank Hopfgartner, Cathal Gurrin and Alan F. Smeaton. (2011). “***TRECVid 2011 Experiments at Dublin City University***“. School of Computing Dublin City University, Ireland.

[8] ***GIM: Qatris iManager***. Retrieved from http://gim.unex.es/qatrisim on October 2012.

[9] J.M. Lanza, Miryam Salas and Manuel Barrena. (2010). "***vManager: una herramienta para la gestión de videos***". XV Jornadas de Ingeniería del Software y Bases de Datos (JISBD’10), Valencia, Spain.

[10] ***LSCOM (Large Scale Concept Ontology for Multimedia)***. Retrieved from http://www.lscom.org/ on October 2012.

[11] ***MPEG-7 Overview***. Retrieved from http://mpeg.chiariglione.org/ standards/mpeg-7/mpeg-7.htm on October 2012.

[12] Rubén Morcillo, Pablo G. Rodríguez, Andrés Caro and Manuel Barrena. (2010). "***vManager: un sistema CBVR basado en color local***". XV Jornadas de Ingeniería del Software y Bases de Datos (JISBD’10), pp. 163-174, Valencia, Spain.

[13] ***SICUBO: home page***. Retrieved from http://www.docugest.es/ on October 2012.

[14] Usman Niaz, Miriam Redi, Claudiu Tanase, Bernard Merialdo, Giovanna Farinella, Qian Li. (2011). “***EURECOM at TrecVid 2011: The Light Semantic Indexing Task***“. Multimedia Department, EURECOM Sophia Antipolis, France.